（预 览 版）

大模型应用：从提示工程到 AI 智能体

赵宇 陈星延 陈中普 陈珍珠 潘宁宁 著



西南财经大学

2025 年 5 月



赵宇，西南财经大学教授，博导，四川省学术和技术带头人后备人选，金融智能与金融工程四川省重点实验室副主任，通用人工智能与数字经济创新团队负责人，西南财经大学学术百人，美国罗切斯特大学联合培养博士，中国人工智能学会自然语言理解专委会委员，四川省计算机学会自然语言处理专委会副主任委员。长期致力于人工智能大模型研究，主要研究方向包括自然语言处理、大模型、AI 智能体、具身智能、数字经济等，打造“大模型三部曲”系列新教材，包括《自然语言处理：大模型理论与实践》、《大模型应用：从提示工程到 AI 智能体》和《AI 智能体产品：技术、设计与案例》，率队自研女娲智能体平台，创立“四川省高校大学生“AI 智能体 +”创新设计大赛”品牌，迄今发表高水平论文 40 余篇（其中中国计算机学会（CCF）A 类/B 类推荐论文 16 篇），包括 IEEE Trans. 系列（TKDE、TNNLS、TMM）以及人工智能顶会 ACL、KDD、WWW 等。主持 2 项国家自然科学基金（1 项面上，1 项青年），主持 2 项四川省自然科学基金（1 项重点，1 项面上），主研 2 项国家社科重大项目，主持 3 项省部级教改项目。撰写专著 1 部，获得专利 1 项。



陈星延，博士，北京邮电大学自动化学院副教授。主要研究方向为网路化计算和大语言模型。曾主持/参与国家重点研发计划、国家自然科学基金面上/青年项目、四川省科技厅重点/青年项目、中央高校基本科研年度项目等项目。

序一

AI 语法课：从提问到创造

“我们的智慧使我们成为人类，人工智能是这种特质的延伸。”

——杨立昆

随着人工智能技术的迅速发展，我们正在见证一场改变人类与机器互动方式的革命。智能体的“诞生”不仅在许多领域开创了新的应用场景，也推动了我们思维方式的转变。作为其中关键的技术之一，提示工程（Prompt Engineering）正是这场变革的起点之一。

2018 年，Google 发布了 BERT（Bidirectional Encoder Representations from Transformers），这项技术极大提升了自然语言处理（NLP）的性能，并奠定了 Transformer 架构在 AI 领域的基础。这一突破为后来更强大的语言模型（如 OpenAI 的 GPT 系列）铺平了道路。特别是 2020 年，OpenAI 发布了 GPT-3，一个拥有 1750 亿个参数的语言模型，这标志着语言模型的能力达到了前所未有的高度。GPT-3 不仅可以完成多种自然语言任务，还能生成高质量的文本，展现出类人般的语言理解和生成能力。

正是 GPT-3 的发布，引发了“提示工程”这一全新领域的兴起。提示工程，顾名思义，就是如何通过设计合适的输入提示，充分挖掘大语言模型的潜力。在早期，研究人员和开发者发现，通过精心设计提示，能够让模型在许多任务上表现得更好，从而开启了与 AI 对话的全新方式。研究者们逐渐意识到，提示不仅仅是简单的输入文本，而是与机器沟通的关键，影响着模型的理解和生成过程。

在这一过程中，Google Research 的 Jason Wei 及其团队提出了“思维链提示（Chain-of-Thought Prompting）”的概念，该方法通过引导模型进行逐步推理，显著提升了模型在复杂任务中的表现。思维链提示的方法，让 AI 能够像人类一样逐步推理，解答一些需要逻辑推理的复杂问题。该方法被证明在

解决数学和推理任务上非常有效，进一步推动了语言模型在更广泛应用中的潜力。

与提示工程相辅相成的是 AI 智能体的发展。智能体（AI Agents）与传统的人工智能工具不同，它们不仅能执行任务，还能够通过算法优化行为策略。这一思想最早可以追溯到 DeepMind 的 AlphaGo，它在 2016 年成功击败围棋世界冠军李世石，标志着 AI 在复杂决策问题上的巨大进步。AlphaGo 的成功不仅是围棋技术的突破，也展示了 AI 如何在环境中学习和调整策略，成为现代智能体技术的里程碑。

随着技术的进步，越来越多的智能体开始应用于实际场景中。现代智能助手，如微软的 Cortana、Google 的 Assistant、苹果的 Siri，已经从最初的简单语音助手，逐渐发展成能够结合上下文理解用户需求的系统。尽管它们的底层模型仍需依赖定期更新与训练优化，但已能通过预设策略灵活响应用户需求，成为日常生活中不可或缺的工具。

提示工程和智能体的结合，开启了 AI 技术新的应用前景。如今，我们不仅可以通过调整提示来提升语言模型的性能，还可以通过不断优化智能体的学习和决策机制，使它们在各种复杂环境中展现出高效的问题解决能力。无论是在金融、医疗，还是在教育和创意产业，AI 智能体正开始渗透到我们工作中。

例如，微软与 OpenAI 合作推出了基于 GPT 模型的 Azure AI 服务，使得企业能够将强大的语言处理能力应用于客户服务、内容创作等多个领域。这些应用无一例外地依赖于提示工程技术的支持，使得模型能够根据不同场景灵活应对各种任务。如今，随着技术的不断进步，提示工程和智能体的发展将继续影响着我们的工作和生活方式。对于任何一位致力于 AI 领域的研究人员或开发者来说，理解如何与这些系统有效互动、设计高效的提示，将是未来在这一领域取得成功的关键。

本书带你走进提示工程的核心领域，揭示这一新兴学科如何从基础概念到应用实践，帮助你掌握与智能体有效合作的技能。无论你是从事人工智能的研究，还是希望利用 AI 提升工作效率，这本书将为你提供有价值的思路和方法。阅读它，开启你与智能体共同创作未来的旅程。

未来不是等来的，而是写出来的。而你写下的第一句话，可能就是一个提示。



全国政协委员、湘江实验室副主任
西南财经大学大数据研究院院长
2025 年 4 月 7 日

前 言

近年来，随着生成式 AI 的崛起，尤其是以 DeepSeek、ChatGPT、千问、豆包、Kimi 为代表的大模型的普及，提示工程（Prompt Engineering）与 AI 智能体（AI Agent）作为一种利用大模型能力的技术手段，受到了广泛关注。近期，国家出台了《教育强国建设规划纲要（2024-2035 年）》等相关政策文件，大力推动人工智能通识教育，其中大模型通识教育尤为关键。然而，目前介绍大模型应用技术知识体系的教材较为缺乏。基于此考虑，催生了我编著本教材的想法。

本教材的核心目标是系统介绍大模型应用知识，帮助读者掌握提示工程和 AI 智能体技术，以发挥大模型在各类任务中的最大效能。全书内容分为三个部分：第一部分介绍大模型初阶应用提示工程的相关知识，包括初识大模型、大模型提示词、大模型辅助工作等；第二部分深入探讨大模型高阶应用的关键技术，涵盖从检索增强、认知框架、使用工具，帮助读者更深入地理解大模型的能力，同时为第三部分 AI 智能体奠定了技术基础；第三部分介绍大模型高阶应用 AI 智能体，并通过平台介绍和案例展示，帮助读者掌握在实际应用中通过构建 AI 智能体来解决问题的能力。

本书旨在为具有一般工程和逻辑思维能力的学生提供一个深入浅出的大模型应用技术学习资料。在编写本书的过程中，我深感大模型技术给人工智能领域带来的深刻变革。提示工程与 AI 智能体作为一种新兴的技术手段，正在成为大模型应用的关键之一。我希望本书能够为读者提供一个关于大模型应用的完整知识体系，帮助大家在大模型应用的浪潮中脱颖而出，并为未来的研究和开发提供指导。

感谢黄萱菁教授、王小捷教授、郭军教授对本书编写给予宝贵建议。感谢我们通用人工智能团队以及金融智能与金融工程四川省重点实验室的同仁们，为本书的完成提供了不可或缺的支持。本书初稿完成以后，宗成庆教授、文继荣教授、张岳教授和马占宇教授浏览了全书的内容。李彪博士、陈珍珠博士、邢容博士、陈中普博士和施龙博士校对了第一章的全部内容。吴钢博士、陈珍珠

博士、邢容博士、施龙博士和杜华明博士校对并修改了第二章的全部内容。潘宁宁博士、黄士罗博士、杜华明博士校对了第三章的全部内容。王俊博士、潘宁宁博士、黄士罗博士、杜华明博士与谢志龙博士校对了第四章的全部内容。庄福振博士、陈珍珠博士、邢容博士、陈星延博士和张蕊博士校对了第五章的全部内容。王昊奋研究员浏览了第五章全部内容。刘康博士、陈珍珠博士、邢容博士、潘宁宁博士和黄士罗博士校对了第六章的全部内容。袁彩霞博士浏览了第六章全部内容。郑海超博士、张蕊博士、陈中普博士和黄士罗博士校对了第七章的全部内容。陈运文博士、徐亮博士、张赟博士、陈中普博士、潘宁宁博士、施龙博士和温良剑博士校对了第八章的全部内容。张赟博士、陈中普博士校对了第九章的部分内容。张赟博士、杜华明博士校对了第十章的部分内容。冯飞高级架构师、肖雪松董事长对该书也提出了宝贵的意见。另外，特别感谢程雨霏、郭源欣、周家宝、沈伟、陈思源、王涛、陈斯康、孙宇、冯瑞琪、汝鑫、张杰、曾子昂、孙文熙、靳凯俊等同学在文稿编辑、图表绘制和审稿校对等方面所付出的努力，这是本书完成的基础。

本书引用了一些优秀参考文献中的图表和案例等，征求了相关作者的意见并得到了积极支持，在此表示由衷的感谢！

我衷心感谢全国政协委员寇纲教授、华为诺亚方舟实验室刘群教授在百忙之中为本书撰写序言，使我倍感荣幸。

此外，由衷感谢机械工业出版社的辛勤付出，感谢你们在本书编写和出版过程中所展现的专业精神和不懈努力，让本书最终得以面世。

本书的编写得到了国家自然科学基金项目的资助。

谨向所有在本书编写过程中给予帮助和支持的个人与机构致以诚挚的谢意！

由于编者水平有限，书中难免有疏漏和不足之处，恳请读者批评指正！如果您发现书中存在有任何错误或遇到任何问题，可以提交至本书网站[1]，您的反馈将是我们不断完善和提升的动力。

赵 宇

2025 年 5 月

[1]本书是“大模型三部曲”系列教材的第二本，第一本和第三本分别是《自然语言处理：大模型理论与实践》、《AI 智能体产品：技术、设计与案例》，更多资源及讨论见官网：https://nlp-book.swufenlp.group

目 录

序一 **i**

前 言 **iii**

第一章 绪论 **1**

1.1 人工智能概述 1
1.2 自然语言处理 5
1.3 语言模型与大语言模型 8
1.3.1 语言模型 8
1.3.2 大语言模型 10
1.4 大模型应用 11
1.4.1 大模型应用技术 11
1.4.2 大模型应用案例 13
1.5 本书内容安排 15
1.6 讨论 16
1.7 习题 16

第一部分 大模型初阶应用：提示工程 **17**

第二章 初识大模型 **19**

2.1 概述 19
2.2 大模型使用示例 20
2.2.1 思维导图生成 20
2.2.2 代码生成 21
2.2.3 图像绘制 22
2.2.4 音乐生成 23

2.3 常见大模型 24
2.3.1 DeepSeek 24
2.3.2 ChatGPT 26
2.3.3 通义千问 28
2.3.4 豆包 29
2.3.5 其它大模型 30
2.4 大模型分类 31
2.4.1 对话型与推理型大模型 31
2.4.2 基础与轻量化大模型 31
2.4.3 单模态与多模态大模型 32
2.5 大模型本地部署 33
2.5.1 Ollama 部署工具 34
2.5.2 Ollama 命令行交互 35
2.5.3 Ollama 可视化界面交互 37
2.6 大模型客户端 41
2.6.1 大模型配置方法 42
2.6.2 内置 Web 访问服务 47
2.7 讨论 48
2.8 习题 48

第三章 大模型提示词 51
3.1 概述 51
3.2 提示词组成要素 51
3.2.1 任务 52
3.2.2 角色 55
3.2.3 上下文 60
3.2.4 输出控制 64
3.2.5 推理型大模型提示词 67
3.3 提示词优化 71
3.3.1 优化策略 72
3.3.2 一键优化 75
3.4 提示词案例 77
3.4.1 写作助理 78
3.4.2 法律咨询 81
3.4.3 头脑风暴 85

3.4.4 学习辅助 89
3.5 讨论 92
3.6 习题 92

第四章 大模型辅助工作 95
4.1 概述 95
4.2 大模型联网检索 95
4.2.1 联网检索流程 96
4.2.2 联网检索分类 97
4.2.3 联网检索优势 99
4.3 大模型辅助办公 101
4.3.1 Word 101
4.3.2 PDF 111
4.3.3 Excel 115
4.3.4 PPT 123
4.4 大模型辅助创作 132
4.4.1 图片生成 132
4.4.2 视频生成 142
4.4.3 音乐生成 149
4.5 大模型辅助学习 153
4.6 大模型辅助编程 158
4.6.1 安装与注册 159
4.6.2 功能介绍 160
4.7 综合案例 169
4.8 讨论 174
4.9 习题 174

第二部分 大模型高阶应用：关键技术 177

第五章 大模型检索增强生成 179
5.1 概述 179
5.2 经典检索增强生成 180
5.2.1 文本嵌入表示 181
5.2.2 文档分块存储 183
5.2.3 高维向量索引 184

5.3 广义检索增强生成 185
5.3.1 基于联网检索的 RAG 185
5.3.2 基于关系数据库的 RAG 188
5.3.3 基于知识图谱的 RAG 190
5.4 讨论 193
5.5 习题 193

第六章 大模型认知框架 195
6.1 概述 195
6.2 情景学习 196
6.2.1 基本概念 196
6.2.2 框架设计 196
6.2.3 应用案例 198
6.3 思维链 203
6.3.1 基本概念 203
6.3.2 执行框架 203
6.3.3 应用案例 204
6.3.4 拓展方法 208
6.4 自我一致性 211
6.4.1 基本概念 211
6.4.2 框架设计 211
6.4.3 应用案例 212
6.5 计划与执行 214
6.5.1 基本概念 214
6.5.2 框架设计 214
6.5.3 应用案例 215
6.6 ReAct 框架 216
6.6.1 基本概念 216
6.6.2 框架设计 217
6.6.3 应用案例 218
6.7 反思 221
6.7.1 基本概念 221
6.7.2 框架设计 222
6.7.3 算法流程 223
6.7.4 应用案例 224

6.8 其它认知框架 228
6.8.1 蚁群智能 228
6.8.2 认知螺旋理论 230
6.9 讨论 233
6.10 习题 233

第七章 大模型使用工具 235
7.1 概述 235
7.2 工具调用能力 236
7.2.1 判断是否使用工具 236
7.2.2 选择合适工具 237
7.3 工具调用方式 239
7.3.1 配置工具 API 239
7.3.2 工具通信协议 240
7.3.3 GUI 模拟操作 249
7.4 常用工具 251
7.4.1 检索工具 251
7.4.2 学习教育 252
7.4.3 社交娱乐 253
7.4.4 图像音视频处理 254
7.4.5 日常生活 255
7.4.6 金融商业 256
7.4.7 其它工具 257
7.5 讨论 258
7.6 习题 258

第三部分 大模型高阶应用：AI 智能体 261

第八章 AI 智能体概述 263
8.1 概述 263
8.2 AI 智能体构成 263
8.2.1 四大核心组件 263
8.2.2 规划 265
8.2.3 记忆 268
8.2.4 工具与执行 271

8.3 多智能体系统 274
8.3.1 多智能体框架 275
8.3.2 多智能体通信协议 278
8.4 AI 智能体产品设计 281
8.4.1 产品设计原则 281
8.4.2 原型设计工具 285
8.5 讨论 287
8.6 习题 288

第九章 AI 智能体平台与应用 291
9.1 概述 291
9.2 智能体开发平台 292
9.2.1 女娲智能体平台 292
9.2.2 其它智能体平台 308
9.3 通用 AI 智能体 309
9.3.1 Manus 310
9.3.2 Flowith 311
9.3.3 其它通用智能体 312
9.4 讨论 313
9.5 习题 313

第十章 AI 智能体行业案例 315
10.1 概述 315
10.2 教育行业案例 315
10.2.1 关键设计方法 316
10.2.2 智能体实现 317
10.3 金融行业案例 319
10.3.1 关键设计方法 320
10.3.2 智能体实现 321
10.4 法律行业案例 321
10.4.1 关键设计方法 322
10.4.2 智能体实现 323
10.5 医疗行业案例 324
10.5.1 关键设计方法 325
10.5.2 智能体实现 326

10.6 讨论 328
10.7 习题 328

附录 A 缩略语表 **329**

附录 B 翻译对照表 **331**

参考文献 **332**

第一章　绪论

为更好地梳理并介绍大模型应用技术，帮助读者构建从提示工程到 AI 智能体的完整知识体系，本章将首先简要阐述其背景知识，包括人工智能、自然语言处理、语言模型与大模型等内容，最后概述了大模型应用相关技术和案例。

1.1　人工智能概述

人工智能 (Artificial Intelligence, AI) 是指通过计算机程序模拟和实现人类智能的技术领域。它赋予机器感知、推理、学习、规划、理解语言和解决问题等能力，从而完成传统上依赖人类智能的任务。

人工智能可分为弱人工智能 (Narrow AI)[1] 与强人工智能 (Strong AI)[2]。弱人工智能指的是设计用于执行特定任务的智能系统，如语音识别、图像分类、推荐系统等，这些系统在特定领域中表现出色，但无法完成预设任务之外的工作，难以如人类一样完成不同种类的任务。强人工智能也称为通用人工智能 (Artificial General Intelligence, AGI)，它是指像人类一样具备适应环境、自主学习和推理泛化等能力的智能系统，能够完成各式各样的现实任务。

1. 历史与发展

人工智能的历史可以追溯到 20 世纪中期，其发展历经多个阶段，每个阶段无疑都为今天的 AI 技术奠定了基础。

初期探索与概念提出（1940s-1950s）：人工智能的思想最早可以追溯到 20 世纪 40 年代的计算机科学与逻辑学研究。1943 年，逻辑学家沃尔特·皮茨（Walter Pitts）和神经科学家沃伦·麦卡洛克（Warren S. McCulloch）提出神经网络的概念，指出通过简单的计算单元可以模拟神经元的工作方式。1950 年，艾伦·图灵（Alan M. Turing）在其著名论文《计算机与智能》[3] 中提出了 “图灵测试”，作为衡量机器是否具备智能的标准。

早期发展与符号主义（1956-1970s）：1956 年，在达特茅斯会议上，约翰·麦卡锡（John McCarthy）等科学家首次提出了 “人工智能” 这一术语，并确立将人工智能作为一个学术领域。此后，该领域的研究重点集中在符号主义方法上，即通过推理、逻辑和规则构建人工智能系统。早期的 AI 系统，如 “逻辑理论家” 和 “通用问题解决者”，是基于这种符号推理的代表性模型，能够在特定规则和知识库的基础上进行问题求解。

知识工程与专家系统（1970s-1980s）：1970 年代末至 1980 年代，AI 研究进入一个新的阶段，称为知识工程。研究人员开始开发专家系统，这些系统能够模拟专家在特定领域中的决策过程。这一时期，AI 应用得到了广泛的关注，尤其是在医疗、金融和制造业等领域。专家系统依赖于大量的领域知识，并通过规则引擎进行推理。然而，尽管人工智能在该时期取得了一定成就，但是，由于技术的限制和过高的期望，AI 研究分别在 1970 年代和 1980 年代末期经历了两次 “冬天”。在此期间，AI 进展缓慢，资金短缺，许多 AI 项目未能实现预期目标，特别是随着专家系统的局限性逐渐显露，人们对 AI 的信心也产生了动摇。

机器学习与深度学习的崛起（1990s-2000s）：随着计算能力的提高和数据的丰富，AI 迎来了新的发展机遇。机器学习 (Machine Learning)特别是决策树 (Decision Tree, DT)、支持向量机 (Support Vector Machine, SVM)等算法取得了重要进展，这些算法突破了传统统计模型对数据分布的先验假设限制，为数据驱动型人工智能范式奠定基础。2006 年，杰弗里·辛顿（Geoffrey Hinton）等学者明确提出了深度学习 (Deep Learning)的概念[4]，利用多层神经网络来处理复杂的模式识别任务，特别是在语音、图像和自然语言处理等领域取得了突破性进展。

深度学习与大数据时代（2010s-至今）：随着深度学习的兴起，AI 进入了黄金发展期。2012 年，卷积神经网络 (Convolutional Neural Network, CNN)在 ImageNet 图像识别竞赛中取得了压倒性的胜利[5]，标志着深度学习的崛起。从此，深度学习成为 AI 研究的主流技术。AI 在语音识别（如苹果的 Siri、谷歌助手）、图像识别（如自动驾驶汽车）、自然语言处理（如 OpenAI 的 GPT 系列、大型语言模型）等多个领域取得了前所未有的成就。

大模型与 AI 智能体时代（2022-至今）：2022 年底，ChatGPT 的横空出世引爆全球对大语言模型（LLM）的关注，其对话能力与通用性使人类第一次摸到了通用人工智能的门槛。2023 年，GPT-4、Claude 等多模态模型的迭代进一步增强了复杂推理与跨模态处理能力，同时开源模型（如 Llama 2）的涌现降低了技术门槛，推动生态多元化。2025 年 1 月 20 日，DeepSeek-R1 推理

大模型发布，其推理能力进入了第一梯队（媲美 OpenAI o1），但训练（560 万美元）和推理成本低、速度快、全部开源。DeepSeek 打破了硅谷传统的“堆算力、拼资本”的大模型发展路径。随着大模型的推理能力的增强，由大模型驱动的 AI 智能体（Agent）随之兴起，以 AutoGPT、LangChain 为代表的技术使 AI 能自主规划任务链，渗透至办公、科研、医疗等领域。

2. 技术领域

AI 的核心技术涵盖多个领域和方法，旨在通过模拟和增强人类智能，赋予计算机或机器自主学习、推理、感知和决策的能力，主要技术包括机器学习、深度学习、强化学习等。机器学习是人工智能的基础技术之一，使计算机能够从数据中自动学习、识别模式并优化性能。深度学习是机器学习的一个分支，通过多层神经网络模拟人脑结构，擅长处理复杂的高维数据，广泛应用于图像识别、语音处理和自然语言理解。强化学习通过智能体与环境的交互，并基于奖励信号优化行为策略，适用于自动控制、游戏对抗等场景。

人工智能研究主要包括自然语言处理、计算机视觉、语音识别和机器人学四大部分，其分别赋予计算机理解语言、感知图像、处理声音和执行物理任务的能力，共同构建了人机交互的全方位智能体系。自然语言处理使计算机具备理解、生成和处理人类语言的能力，是实现人机对话、机器翻译和文本分析的关键技术。计算机视觉赋予计算机感知图像与视频中信息的能力，核心任务包括图像识别、目标检测和图像生成。语音识别技术可将语音转换为文本，语音合成技术则使计算机能够生成自然语言语音，实现人机语音交互。机器人学涉及机器人的感知、决策与控制，结合 AI 技术后，现代机器人可在自动驾驶、智能制造等领域执行复杂的物理任务。

人工智能的核心技术相互协同，共同推动了 AI 应用的广泛应用与持续发展。随着大数据的发展和计算能力的不断进步，AI 技术的潜力将不断被释放，将会对更多行业和领域产生深远影响。

3. 应用场景

人工智能技术已经渗透到社会的各个行业，涵盖了医疗、金融、交通、零售、制造、教育等多个领域，并逐渐改变着各行各业的运作模式。随着技术的不断发展，AI 将在更多领域发挥重要作用，推动各行各业的智能化和自动化，改善生产效率和生活质量。以下是人工智能的一些主要应用领域：

在医疗领域，人工智能正被广泛应用于疾病诊断、个性化治疗和药物研发等方面。通过机器学习和深度学习算法，AI 能够分析大量医疗数据，辅助医

生进行诊断工作。例如，AI 系统能够通过分析 X 光、CT 和 MRI 等医学影像来检测肿瘤或其他异常，其准确率甚至可以超过部分人类专家。在个性化治疗方面，AI 通过分析患者的基因数据和病史，帮助医生制定更适合的治疗方案。此外，AI 在药物研发中也展现出巨大潜力，能够显著加速药物筛选与新药发现的过程，提高研发效率。

在金融领域，人工智能的应用包括风险评估、投资决策、智能交易、客户服务等。AI 通过分析历史数据，帮助银行和金融机构评估贷款人的信用风险，预测股票市场的走势，提供个性化的投资建议。在投资领域，量化交易算法利用机器学习进行市场分析，自动执行交易决策，从而提高交易效率并减少人为错误。智能客服和聊天机器人也广泛应用于客户服务，能够快速处理客户的查询、解答问题，并提供个性化的理财建议。

自动驾驶是人工智能在交通领域的一个重要应用。通过计算机视觉、传感器融合、深度学习和强化学习等技术，AI 使得自动驾驶汽车能够感知周围环境、规划路径并做出驾驶决策。自动驾驶技术不仅可以减少交通事故，还能提升道路运输的效率，优化交通流量。此外，AI 还被用于交通管理系统中，通过智能信号灯、交通预测和优化调度来提高城市交通的效率，减少交通拥堵。

人工智能在零售和电子商务中的应用十分广泛，主要体现在个性化推荐、库存管理、客户服务和智能支付等方面。AI 通过分析用户的购物习惯、浏览记录和历史购买数据，为消费者提供个性化的商品推荐，提高转化率和客户满意度。在库存管理方面，AI 可以预测商品的需求量，从而帮助商家进行精准的库存控制，避免过度库存或缺货现象的发生。智能客服机器人也可以解答顾客的问题，并处理简单的订单查询或退换货请求。

在工业领域，人工智能技术促进了智能制造的发展。AI 技术被广泛用于生产过程中的自动化控制、质量检测、设备维护和优化生产流程。通过深度学习和计算机视觉，AI 可以实现实时的产品质量监测，及时发现生产中的缺陷，保证产品质量的一致性。此外，AI 还被用于预测设备的故障和维护周期，通过机器学习算法分析设备的数据，提前发现潜在故障，减少停机时间。

在教育领域，人工智能可以实现个性化学习，AI 系统能实时捕获并分析学生的多维度学习数据，包括知识掌握程度、认知负荷、学习风格偏好和情感状态等，并根据学生的认知特征、历史表现和学习目标，动态调整内容难度、呈现方式和学习节奏。AI 还可以通过智能辅导和答疑系统，帮助学生解答难题，提高学习效率。此外，AI 也能自动批改作业和考试，减轻教师的工作负担。

在娱乐和媒体行业，AI 技术被广泛应用于内容推荐、智能创作和个性化体验中。例如，视频流媒体平台（如国内抖音平台以及国外的 Netflix、YouTube

等）利用 AI 算法分析用户的观看历史，为用户推荐个性化的电影和电视剧。AI 还被用于内容创作，如自动生成新闻报道、影视剧本，甚至生成音乐、绘画等艺术作品。此外，AI 在游戏开发中也发挥着重要作用，尤其是在游戏中的 NPC（非玩家角色）行为模拟、自动难度调整等方面。

1.2 自然语言处理

自然语言处理 (Natural Language Processing, NLP) 是人工智能领域的重要研究方向之一。若要真正掌握 NLP 的核心理念，首先需弄清它所要解决的关键任务。正如人类在学习新语言时需要兼顾听、说、读、写等技能，计算机若想理解并应用人类语言，同样离不开多方面的技术协作。在这一过程中，分词、词性标注与句法分析等基础任务构成自然语言处理的学科基础，而机器翻译、问答系统、文本摘要等高级功能则体现了 NLP 技术在实际应用场景中的落地实现。

1. 基础技术

就像学习一门外语需要从认识单词开始，到理解语法，再到掌握意思，最后才能阅读整篇文章一样，NLP 也需要类似的步骤。NLP 的传统基础技术就像一个层层递进的学习体系。从最基础的词语识别，到句子结构分析，再到理解语义，最后达到理解整篇文章的水平。

在自然语言处理的词法分析层面，自动分词是对汉语、日语等词边界不明确的语言必不可少的预处理步骤。与英文等使用空格分隔词语的语言（如 “natural language processing”）不同，中文文本缺少明显的词边界（如 “自然语言处理”），需要将连续的字符序列正确划分成独立词语。分词过程面临三个主要挑战：分词规范问题（如 “北京大学计算机专业” 可分为 “北京/大学/计算机/专业” 或 “北京大学/计算机/专业”）、歧义切分问题（如 “研究生命题” 可分为 “研究生/命题” 或 “研究/生命/题”）、以及未登录词识别问题（如新词 “元宇宙”、专业术语 “RNA 干扰”、网络用语 “打工人” 等）。为解决这些问题，研究者提出了多种技术方法与策略。针对分词规范问题，采用领域特定词典与统计共现信息相结合的方式，识别” 北京大学” 等专有名词的整体性；对于歧义切分问题，引入句法分析与上下文语义理解机制，通过计算多重分词可能性的概率分布来确定最优切分；而解决未登录词识别问题则依靠基于字的互信息和熵的新词发现算法，结合预训练语言模型的迁移学习能力，动态扩充词库。

紧随分词之后，命名实体识别 (Named Entity Recognition, NER) 也是自

然语言处理领域的传统基础技术。它就像是一个智能标记员，负责找出文本中的特殊名词。比如在 “张三昨天去北京大学参加会议” 这句话中，“张三”（人名）、“北京大学”（机构名）和 “昨天”（时间）就是需要识别的命名实体。

词性标注就像是给句子中的每个词语贴上身份标签。比如在 “小明正在认真地学习” 这句话中，“小明” 是名词（表示人），“正在” 是助词（表示时态），“认真” 是形容词（表示状态），“地” 是助词（表示方式），“学习” 是动词（表示动作）。这种标注对于计算机理解人类语言非常重要，就像人类学习语文时需要分清楚词语的词性一样。

在自然语言处理的高级阶段，句法分析、语义分析和篇章分析是理解文本的关键技术。句法分析是通过短语结构分析与依存句法分析，对语言结构进行更高层次的解析。语义分析是迈向更深层次理解文本含义的关键步骤，它通过多个层次逐步深化对文本的理解。在词汇层面，词义消歧构成了语义分析的基本环节，其核心任务是解决多义词在特定上下文中的正确含义选择问题。篇章分析则扩展了处理的范围，不再局限于单句，而是关心多句子构成的文本或语篇。篇章分析的重要任务包括连贯性分析与指代消解。连贯性分析涉及局部与全局层面的连贯，如局部连贯体现为相邻句子的逻辑、语义、实体和话题关联性，全局连贯体现为整个文本主题、信息结构和文本组织形式的一致性。

总之，从分词、命名实体识别、词性标注，到句法分析、语义分析，再到篇章分析，这些传统基础技术层层递进地为自然语言的计算处理提供必需的基础。它们在自然语言理解方面各司其职，其中分词与词性标注在词法层面奠定基础；命名实体识别为信息抽取和知识图谱构建提供精准实体边界与类型信息；句法分析为句子结构建模，明确组块和依存关系；语义分析聚焦于词义与句子含义的理解，为机器提供更深层的文本意图认知；篇章分析则将目光扩展到整篇文档的连贯、主题延续和指代现象处理上，为高级 NLP 应用（如信息检索、自动文摘、问答系统、机器翻译、情感分析等）提供了基础支持。

2. 实际应用

自然语言处理的基础技术为众多应用场景提供了有力支撑。从机器翻译、文本分类与情感分析，到信息抽取与自动文摘、信息检索与问答系统，再到与人类交流近似的对话系统，这些应用已广泛融入人们的学习、工作和生活中。

机器翻译（Machine Translation）旨在通过自动化手段将源语言文本高质量地转换为目标语言文本。最开始，科学家们尝试教会计算机一些基本的翻译规则，就像给计算机一本翻译教科书。但这种方法太过死板，常常会闹出笑话。比如 “这个银行很大（This bank is big）” 被翻译成 “这个河岸很大”，因为计

算机不懂得根据上下文选择正确含义。现代的机器翻译系统就像一个精通多国语言的翻译专家，它不仅能理解单词的含义，还能根据整句话的上下文选择最恰当的表达。

文本分类（Text Classification）是将文本分配到预设类别的任务，在垃圾邮件识别、新闻分门别类、法律文档归类、舆情分析及广告投放定位等领域均有广泛应用。深度学习与预训练模型（如 BERT、GPT）赋予系统对文本语义与上下文的深度理解，使模型可自动捕捉语料中的特征与模式，大幅提升分类性能。作为文本分类的重要分支，情感分析（Sentiment Analysis）旨在判断文本的情感倾向（如正面、负面或中性）。该任务被广泛用于舆情监控、产品口碑分析、用户反馈优化与市场调研。在社交媒体、在线评论平台和企业客服系统中，情感分析为商业决策和公共政策制定提供重要参考。

信息抽取（Information Extraction）通过识别与提取非结构化文本中的实体、关系和事件，将繁杂的自然语言材料转化为结构化数据，支撑知识图谱构建和深层语义分析。例如，命名实体识别、关系抽取和事件抽取等技术可从新闻报道、科研论文中捕捉关键信息，为后续的信息检索与决策分析奠定坚实基础。自动文摘（Summarization）技术则通过精选与压缩文本中的关键信息，生成简洁明了的摘要。在海量信息时代，自动文摘帮助用户快速把握文档要点，提升信息获取效率。无论是新闻聚合、学术文献检索，还是商业报告与社交媒体内容的聚合与呈现，自动文摘为用户带来更清晰、高效的信息获取体验。

信息检索（Information Retrieval）技术支持从庞大文档集（如网页、学术论文库）中快速找到与用户查询最相关的内容，是搜索引擎与在线数据库的核心。倒排索引、查询解析及相似度计算等基础技术结合深度学习模型，使检索结果更加精准、实用。在此基础上，问答系统（Question Answering）进一步将检索范围缩小到精确答案的定位。基于检索的问答系统先从文本库中找到相关段落，再用阅读理解模型抽取或生成最终答案。生成式问答或基于知识库的问答系统可根据已有信息进行灵活的自然语言生成与推理，为在线客服、智能助手和教育辅导提供直接、简明的回答。评价指标如平均倒数排名（Mean Reciprocal Rank）[1]可用于衡量系统对测试问题集的回答质量。

对话系统（Dialogue Systems）模拟人与人对话的过程，满足交互式信息获取与任务完成需求。闲聊型对话系统注重语气自然与类人化反应，可作为虚拟助手提供娱乐或陪伴；而任务型对话系统则面向机票预订、餐厅订位、银行业务咨询等具体场景，强调对话策略和上下文追踪，确保用户目标高效达成。对话管理在此尤为关键，其负责分析对话历史、跟踪状态、识别用户意图，并

[1]维基百科 “平均倒数排名” 释义：`https://en.wikipedia.org/wiki/Mean_reciprocal_rank`

对系统响应进行规划与生成。随着语音识别和语音合成技术的成熟，对话系统已广泛应用于智能家居、客服机器人和车载系统，为用户带来更自然便捷的交互体验。

1.3 语言模型与大语言模型

本节将简要介绍语言模型和大语言模型的相关知识和技术[2]。

1.3.1 语言模型

语言模型（Language Model, LM）是自然语言处理中的一个基础概念，其基本目标是通过大量文本数据学习语言的结构和模式，进而能够处理各种语言任务。其核心任务是估计语言中词序列的概率分布，即给定一个词序列，模型预测下一个词出现的概率。这种预测能力与人类的语言认知类似，比如当听到“北京是中国的”时，人们很容易猜到下一个词是“首都”。语言模型通常以词或子词为单位，对语言的统计规律进行建模，主要分为统计语言模型（Statistical Language Model, SLM）、神经语言模型（Neural Language Model, NLM）和预训练语言模型（Pre-training Language Model, PLM），下面给出简要介绍。

1. 统计语言模型

统计语言模型基于概率统计理论方法，主要用于描述词序列的概率分布。它通过分析大规模语料库中词汇的分布规律和共现关系，构建能够预测特定上下文中词语出现概率的模型系统。在统计语言模型中，N-gram 模型是最经典的代表，其通过假设每个词只依赖于前面 $N-1$ 个词来简化模型的计算复杂度。其中，最常见的是 Bigram（2-gram）和 Trigram（3-gram）模型，它们分别通过当前词和前一个词或前两个词来预测下一个词。这种模型具有明显的优缺点。优点在于其简单、易于实现，并且能够在大规模数据集上训练出有效的模型。然而，其缺点也十分明显。首先，N-gram 模型的上下文窗口有限，无法捕捉长距离的依赖关系；其次，随着 N 的增大，模型的参数空间急剧膨胀，导致计算开销和存储需求增加。更重要的是，统计语言模型主要依赖于大量的标注数据和统计信息，无法有效理解自然语言的深层次语义。

[2]如果读者想详细学习语言模型和大语言模型详细知识，请参考教材《自然语言处理：大模型理论与实践》。

2. 神经语言模型

神经语言模型 (Neural Probabilistic Language Model) 是将深度神经网络应用于语言模型领域的早期工作之一。该模型利用一个神经网络来学习词语之间的条件概率关系，从而实现文本生成和预测的能力。与统计语言模型相比，神经语言模型能够突破传统 N-gram 模型只能看到有限窗口的局限，不仅可以捕捉句子中相距较远的词语之间的关联，还能够自动学习词语的分布式表示，理解词语在不同语境下的细微差别。在技术实现层面，神经语言模型主要采用三种架构。最基础的前馈神经网络（Feed-forward Neural Networks, FNN）通过固定大小的输入窗口处理文本。引入循环连接的循环神经网络（Recurrent Neural Networks, RNN）能够处理任意长度的序列输入。而长短期记忆网络（Long Short-Term Memory, LSTM）则通过设计的门控机制解决了 RNN 在处理长序列时的梯度消失问题。以句子 “经过长期的艰苦努力，他终于取得了成功” 为例，统计语言模型通常仅关注最后几个词来预测下一个词，而神经语言模型能够理解整个语境，捕捉到” 长期艰苦努力” 与” 取得成功” 之间的语义关联，从而做出更准确的预测。尽管神经语言模型在训练过程中需要更多的计算资源，但相较于统计语言模型，其在性能上取得了显著提升。

3. 预训练语言模型

随着大语言模型的发展，研究者们发现一个有趣的现象。如果让模型预先学习海量的通用文本数据，就像人类在成长过程中通过大量阅读和交流来积累语言知识一样，模型在特定任务上的表现会显著提升。这一发现催生了预训练语言模型的诞生。预训练语言模型在神经语言模型的基础上引入了 “预训练-微调” 的范式。首先，模型在海量无标注文本上进行预训练，就像 “博览群书” 一样，学习语言的基本规律和知识；然后，再针对特定任务进行微调，就像人类将通用知识应用到专业领域一样。

“预训练-微调” 范式继承了神经语言模型捕捉长距离依赖和理解上下文的优势，具有三个重要突破：1）它可以直接利用互联网上的海量文本进行训练，无需人工标注数据；2）预训练获得的语言知识具有通用性，使模型能够同时应对文本理解、问答和摘要等多种任务；3）通过预训练积累的知识使模型具备了一定的情景学习能力，只需少量示例即可快速适应特定任务。语言模型展现出广泛的应用价值，在文本生成、机器翻译、语音识别和情感分析等领域发挥着重要作用。

1.3.2 大语言模型

正如前面所讨论的预训练语言模型，研究者们逐渐认识到，模型的性能与其规模和训练数据量密切相关。这种认识促使研究方向从传统的预训练语言模型转向了更深层次的探索，即如何将更多的知识和能力整合到单个模型中？由此催生了一种全新的语言模型范式。

大语言模型简称“大模型”（Large Language Model，LLM），通常是基于 Transformer 架构[6]构建的预训练语言模型。与传统的预训练语言模型相比，大语言模型最显著的特点是采用了数量级更大的模型参数和训练数据。这种规模的提升不仅显著提高了性能，还产生了“涌现能力”（Emergent Abilities）[3]。例如，大语言模型展现出了强大的上下文学习能力，能够根据少量示例快速适应训练数据集中未出现的任务，还能进行复杂的推理和创造性思维。大模型出现标志着语言模型迈向了全新的发展阶段。

大语言模型的工作原理主要包括自回归（Autoregressive）和自编码（Autoencoder）两种范式。自回归模型像一个出色的写作者，能够根据已有内容预测和生成后续文本，擅长连贯性写作和对话生成，如 GPT 系列。自编码模型则像一个细心的阅读者，通过关注上下文的方式来理解文本含义，在阅读理解和文本分析等任务上表现出色，如 BERT 系列。

大语言模型的关键技术包括 Transformer 神经网络架构、大规模预训练和微调等，下面从这两个维度简要展开介绍。

1. Transformer 架构

Transformer 架构可以看成是构筑大语言模型“大脑”的基础，它彻底改变了机器处理语言的方式。想象一下阅读一本书，传统的 RNN 和 LSTM 就像是一个人必须从头到尾按顺序阅读，而 Transformer 就像是能够同时关注整页内容的专家读者，可以自由地在文本中建立联系。这种突破性的设计基于自注意力机制（Self-Attention Mechanism）[6]，使得模型能够同时处理文本中任意位置的词语关系，大大提升了并行计算效率和长距离信息处理能力。这种自注意力机制就像是给每个词配备了一个“智能探照灯”，能够自动调整关注度，聚焦到上下文中最相关的部分。例如，在理解“今天阳光明媚，但是气温很低”这句话时，模型会特别关注“但是”前后的关联信息。通过多层自注意力机制和前馈神经网络的基础结构组合，Transformer 构建了一个强大的语言理解系统，能够准确把握文本中细微的语义关系和复杂的表达模式。

[3]涌现能力：模型表现出在训练时并未明确设计的能力。

2. 大规模预训练和微调

大语言模型的训练过程采用 “预训练-微调” 范式。随着深度学习技术的发展，预训练模型的参数规模呈现出惊人的增长趋势。参数数量从百度 ERNIE 3.0 Titan 模型的 260 亿，跃升至智源 GLM 模型的 1300 亿，DeepSeek 系列进一步通过混合专家（Mixture of Experts, MoE）架构下的稀疏激活技术，将参数规模推升至 6710 亿。此类模型的训练需依赖分布式并行架构、混合精度训练等前沿技术，仍需数千 GPU 持续运行数月，消耗数百太瓦时的计算资源。研究人员发现了提升模型参数规模可以带来多方面的性能突破。首先，大规模模型展现出更强的知识储备能力，能够从训练数据中捕获和记忆更多的语言知识和世界常识。其次，随着参数规模增长，模型的理解能力和生成能力都得到显著提升，不仅能更准确地理解复杂的语言表达，还能生成更连贯、更符合语境的回应。此外，大语言模型在经历 “微调” 训练后，表现出强大的迁移学习能力，往往只需少量示例就能适应新任务，这种 “少样本学习”（Few-shot Learning）能力使其应用范围得到极大扩展。不仅在传统的自然语言处理任务表现出色，还可以扩展到图像识别、语音处理、视频理解等多模态任务中，展现出强大的通用人工智能潜力。

1.4 大模型应用

本节将从大模型应用技术和应用案例展开讨论。在大模型应用技术方面，将概述代表性的提示工程和 AI 智能体关键技术，这些技术是实现大模型高效应用的重要基础。在大模型应用案例方面，将探讨大模型在教育、金融、法律、医疗等垂直领域的创新实践。

1.4.1 大模型应用技术

1. 提示工程

提示工程 (Prompt Engineering)是一门通过设计和优化语言输入，来指导生成式人工智能 (Generative Artificial Intelligence, GAI)模型（如 DeepSeek[7]、ChatGPT[8]等大模型）产生预期输出的技术。近年来，基于 Transformer 架构[9]的大语言模型推动了人工智能应用的发展。随着工程应用需求的不断提升，业界开始了探索高质量提示工程的研究，以促进 GAI模型在各行各业发挥其性能。

提示工程被视为连接人类与大模型的桥梁，它能引导模型生成内容（如撰

写文章）和完成任务（如执行代码）。但是，大模型仍存在幻觉、欺骗等问题。例如，英国的人工智能安全研究机构阿波罗研究机构 (Apollo Research)最新的研究[10]发现，目前多个前沿大模型（包括 Claude 3.5、Gemini 1.5 Pro 等）已具备向使用者隐藏其真实目的的策划欺骗能力，且模型的欺骗行为具有持续性。为了提高模型输出内容可靠性，需要使用者能够设计出更高效、可靠的提示词。因此，使用者需要深入理解大模型对不同语言输入的响应方式，以及如何利用大模型优化提示词。正如人类通过与环境交互来学习一样，大模型的使用者也需要通过与模型的互动来了解其应答机制和潜在能力，以便快速高效地获取所需答案。

提示工程是一门需要不断实践和探索的学科。使用者们需要在理解模型工作原理的基础上，持续优化和创新提示词的设计，以充分发挥大模型的潜力。这项技术不仅包括基础的提示词编写，还涵盖了如内容检索增强、认知框架设计和大模型工具使用等关键技术。

检索增强生成是提示工程在知识获取方面的延伸。本质上，它是一种高级的知识检索提示技术。首先检索相关知识，然后将检索到的信息与用户问题结合，形成增强后的提示，从而帮助模型生成更准确的回答。这种方法有效解决了大模型知识时效性差和专业领域知识欠缺的问题。

认知框架是提示工程在思维方式层面的创新，它通过不同类型的提示方式来增强模型的认知能力，主要包括情景学习、思维链、自我一致性和计划与执行等。情景学习属于示例提示方法，通过示例来引导模型学习并产生迁移解决问题的能力；思维链采用分步提示方法，将任务拆解为多个步骤依次推理并解决；自我一致性采用多路径提示方法，从不同角度验证并整合输出结果的正确性；计划与执行是循环提示方法，将 “思考—行动—观察” 融为一个动态循环；反思框架利用二次提示（一次生成、一次检查）方法，对初步输出进行审查和修正。通过这些高级提示方法，群策群力地赋予大模型远超出简单问答水平的高级行为。

在提示工程实践中，充分利用大模型的工具调用能力是提升效果的关键策略。通过精心设计的提示，用户可以引导模型适时调用外部应用程序编程接口 (Application Programming Interface, API)、数据库或计算引擎等工具，从而弥补模型知识局限。有效的提示需明确指定可用工具类型、调用时机和结果处理方式，使模型能在回答过程中主动请求所需资源。例如，针对实时数据查询任务，提示可包含查询函数的使用说明；对于复杂计算，则需在提示中说明计算工具的调用方法。这种结合工具使用的提示工程技巧，显著扩展了模型的应用范围，使其在专业领域和动态环境中展现更强的实用价值。

这些技术虽然表现形式各异，但本质上都建立在提示技术的基础之上。例如，检索增强生成是知识检索与提示词相结合的技术方法；认知框架中的各种方法是在提示词方法技术上演进而来的复杂提示逻辑框架。通过认知框架，大模型展现出接近人类认知水平的智能表现。这种技术体系的建立，使得提示工程不再局限于简单的提示词编写，而是发展成为一个包含知识获取、认知增强和工具应用的完整技术领域。通过综合运用这些技术，能够更好地发挥大模型的潜力，从而实现更高级的人工智能应用。

2.AI 智能体

AI 智能体是将大模型与外部资源和工具深度整合的集合体，使其具备“认知—决策—执行”全流程自治的能力。简言之，AI 智能体指的是 AI 大模型与工具集的结合体，就像人类使用工具从而变得强大一样，大模型通过使用工具组成智能体，可以极大地增强处理各项事务的能力。

具体而言，AI 智能体通过在大模型提示系统中集成自主决策模块，结合工具调用和多步工作流机制实现复杂任务执行。其核心在于动态策略优化能力，系统可自主触发知识检索、数据库查询或计算引擎调用等操作，并基于实时反馈进行策略迭代。这种架构使大模型从被动响应提示的范式，转型为具备环境适应能力的主动系统，在工业控制、数据分析等场景中展现出持续稳定的智能表现。

1.4.2 大模型应用案例

1. 自然语言处理领域的应用案例

文本生成与对话系统是语言模型的基础应用。在自动化内容创作、聊天机器人、新闻生成、创意写作等领域，LLM 通过生成连贯、富有逻辑性和创意的文本，极大地提高了生产力。例如，GPT-3 能够根据用户输入的短语或问题生成完整的文章、代码，甚至创作诗歌。在对话系统中，基于大语言模型的聊天机器人能够进行人机对话、问题回答、情感交互、写作风格模拟。其强大的语言生成能力使得这些系统更加自然和灵活。

机器翻译借助大语言模型的双向上下文编码和强大的序列生成能力，使翻译性能得到了极大的提升。传统的基于统计的机器翻译方法（如 N-gram）在翻译质量上存在较大限制，而大语言模型通过学习大量语言数据中的翻译规律，能够实现更高质量、更自然流畅的翻译结果。如 Google Translate、DeepL 等服务采用基于 Transformer 的大语言模型来进行高效的机器翻译。

多模态应用标志着大语言模型在范式上的跃迁。目前，许多大语言模型已经不再局限于单一文本的处理，而是扩展到多模态的任务中，如图像描述、图像搜索、视频分析等。通过结合图像和文本数据，大语言模型可以在多模态理解和生成任务中发挥作用，例如 OpenAI 的 CLIP 和 DALL·E 模型，能够基于文本生成图像，或者基于图像生成描述。

2. 其它领域的应用案例

随着大语言模型通用能力的不断提升，其应用已从传统自然语言处理任务拓展至金融、法律、教育、医疗等多个关键行业，并实现了多样化、实用性强的落地应用。

在金融领域，金融机构在海量文本（如市场报告、新闻、客户咨询等）处理方面往往面临实时性不足与认知深度有限等问题，难以及时提炼有效信息并确保风险评估与决策的精准度。大型语言模型凭借其深度学习能力，可快速解析低结构文本与复杂语义，对瞬息万变的市场及客户需求作出更贴近实况的判断，同时辅助从业人员从海量数据中挖掘潜在风险与投资机会，从而显著提升运营效率与客户体验。

在法律领域，面对数量庞大、结构复杂的合同、判例与法规文本，人工审阅方式成本高、效率低，难以及时理解要点并执行合规审查，因而容易造成法律风险被忽视或判断失误等问题。大语言模型可深度解析法律术语与文本语境，快速提取并整理核心条款与信息，为法律从业者提供更系统化的筛选和分析工具，从而大幅降低阅读与判断的负担。大模型在法律领域落地的场景主要包括对法律文书的深度解析及合规审查判例检索两大方向，助力法律从业者在海量文本中快速获取专业决策支持，有效提升法律服务质效。

在教育领域，教育资源分布不均且成本高昂，教师难以对每位学生进行个性化辅导，同时大量重复性工作如出题批改占用了教学创新时间，影响教学效果与学习体验。大语言模型通过深度理解学习内容与学生需求，可提供智能化、个性化的学习支持，同时承担教学辅助工作，让教师更专注于课程设计与教学创新。大模型在教育领域可能落地的场景有个性化学习辅导和教学任务智能化两大方向。通过提供实时答疑指导与自动化教学支持，实现优质教育资源的高效触达与教学质量提升。

在医疗领域，优质医疗资源分配不均且就医流程复杂，医生难以负担持续增长的诊疗需求，同时大量临床数据未被充分利用，影响诊疗效率与医疗服务可及性。大语言模型通过深度理解医学知识与病历数据，可提供智能化的临床决策支持，同时承担基础问诊工作，让医生更专注于复杂疾病诊治与医疗创新。

大模型在医疗领域落地的可能场景有临床辅助决策和智能化问诊服务。通过提供诊疗建议与远程医疗支持，实现优质医疗资源的高效触达与诊疗质量提升。

大语言模型作为自然语言处理领域的革命性技术，凭借其强大的文本生成和理解能力，已经在众多应用场景中展现出巨大潜力。随着软硬件计算设备的不断提升和模型技术的精进，面向金融、法律、教育、医疗领域的专用方案正不断演化，未来的大语言模型将在人工智能领域扮演越来越重要的角色。

1.5 本书内容安排

本书首先为读者概述了人工智能、自然语言处理，以及语言模型与大语言模型等基本概念，然后介绍提示工程的入门知识，接着深入探讨提示工程的进阶技巧。随后，本书对大模型 AI 智能体框架，以及智能体设计平台与应用案例进行了详细讲解。具体来说，本书共包含十章，分为三大部分。

第一部分介绍大模型初阶应用：提示工程，涵盖第二章至第四章的内容。第二章，初识大模型，带领读者初步认识当前主流的大模型，主要介绍了大模型的示例和能力，包括 DeepSeek、ChatGPT、豆包、Kimi 等常见大模型。此外，本章还介绍了大模型分类、大模型本地部署方法以及大模型客户端。第三章，大模型提示词，详细阐述了提示词的组成要素，包括任务目标、角色、上下文信息和输出控制等。本章还讨论了提示词优化方法以及一键优化，并提供了多个提示词的实际案例。第四章，大模型辅助工作，介绍了大模型在辅助检索、文本办公、内容创作、学习以及编程等方面的应用。

第二部分介绍大模型高阶应用：关键技术，包括第五章至第七章的内容。第五章，大模型检索增强，主要介绍了大模型数据检索增强技术，包括文档检索增强、网络检索增强、知识图谱检索增强和关系数据库检索增强。第六章，大模型认知框架，介绍了情景学习、思维链（CoT）、自我一致性、计划与执行、ReAct 框架以及反思（Reflexion）等常见的认知框架。第七章，大模型使用工具，介绍了工具调用框架、工具调用能力以及工具调用接口，并展示了常见工具。

第三部分介绍大模型高阶应用：AI 智能体，包括第八章和第十章的内容。第八章，AI 智能体概述，介绍了 AI 智能体的构成框架、多智能体系统和智能体产品设计。框架部分涵盖了规划、记忆和工具，多智能体部分介绍了多智能体框架与通信协议，设计部分则涉及智能体设计原则与工具等内容。第九章，AI 智能体平台与应用，介绍了国内外的 AI 智能体开发平台，如女娲智能体平台等，并介绍了通用 AI 智能体如 Manus 与 Flowith 等。第十章，AI 智能体

行业案例，主要介绍了智能体在教育、金融、法律、医疗等多个行业的应用案例，旨在通过实际案例，帮助读者理解智能体在各个领域的应用方式和效果。

本书旨在为学生和社会人士提供有关大模型应用的教学资源。全书注重实践应用，突出使用和操作，涵盖了大模型提示工程和 AI 智能体的关键技术及最新进展，为读者提供实用的指导和操作方法，助力培养具备前沿技术应用能力的专业人才。

鉴于本书的章节安排，读者可以根据自身需求和兴趣选择章节进行学习。第一部分的基础知识面向大模型应用的初学者，帮助读者快速掌握通过提示词更好地使用大模型的技巧。第二部分和第三部分的内容则更侧重于实际应用场景和 AI 智能体设计，适合希望深入了解大模型高阶应用实践的读者。

1.6 讨论

讨论 1.1. 请讨论语言模型与大语言模型的关系。

讨论 1.2. 请讨论大模型技术在人工智能领域地位，分析其对人工智能技术发展的影响。

1.7 习题

习题 1.1. 请阐释人工智能的基本概念。

习题 1.2. 请阐释自然语言处理的基本概念。

习题 1.3. 请阐释语言模型的基本概念与分类。

习题 1.4. 请阐释预训练语言模型的基本概念。

习题 1.5. 请解释什么是大语言模型，并说明它与传统预训练语言模型的主要区别。

习题 1.6. 请阐释大模型的基础架构。

习题 1.7. 请阐释大模型训练的主要步骤。

习题 1.8. 请阐释大模型技术的应用领域。

习题 1.9. 请解释大模型提示工程的概念与目的。

习题 1.10. 请阐释大模型智能体的基本概念。

第一部分

大模型初阶应用：提示工程

第二章　初识大模型

2.1　概述

大模型是一种具有庞大参数规模的机器学习模型，广泛应用于自然语言处理、计算机视觉等领域。大语言模型 (Large Language Model, LLM)通常指包含数百亿或更多参数规模的语言模型，其在海量的文本数据上进行训练，从而获得理解和生成自然语言的强大能力。目前，国内外已发布的大模型超过百种，代表性模型包括国内的 DeepSeek、豆包、Kimi 与文心一言，及国外的 GPT、LLaMA[11] 与 Claude[12] 等。这些大模型凭借强大的推理与生成能力，正在以前所未有的速度重塑人类的工作与生活方式，广泛应用于知识问答、任务执行、多媒体内容创作、辅助决策等场景。然而，不同模型在能力侧重、适用领域与部署方式等方面各具特色。例如，DeepSeek-R1 模型[7] 更适合处理代码生成和数学问题求解等推理类型的任务，而 GPT-4o[13] 则更适合进行图文及语音问答等多模态信息处理类型的任务。因此，熟悉不同模型的性能特点，有助于进一步提升工作与学习的效率。不过，模型性能只是使用体验的一部分。在实际应用中，网络连接、中断风险以及服务响应延迟等问题，也会影响模型的可用性与稳定性。在网络受限或服务器响应缓慢的情况下，如何灵活切换至本地部署成为现实需求。那么，是否存在更为便捷的客户端工具，能够支持本地部署模型与远程模型服务之间的灵活切换，从而提升大模型的使用体验?

为了帮助读者更好地理解并应用大模型，本章首先通过大模型的使用示例，展示大模型的生成与推理能力，帮助读者建立对大模型能力的初步认知；接下来，将介绍当前主流的大模型，包括 DeepSeek、ChatGPT、豆包、Kimi 等，探讨它们的技术特点与应用场景，并对常见大模型进行分类总结，使读者能够系统地了解不同类型大模型的特点及适用范围；随后，本章将详细介绍如何在本地环境中部署和运行大模型；最后，本章将介绍 Cherry Studio 等大模型客户端，帮助读者根据自身需求选择合适的大模型，提升工作和学习效率。

2.2 大模型使用示例

大模型的最常见使用场景之一是对话问答，其已被广泛应用于智能客服和虚拟助手等领域。用户通过提示词对大模型进行提问，可完成文本生成、摘要生成、文本总结和机器翻译等需求。但是，除对话问答这一基本功能之外，大模型还能够应用于哪些场景呢？在不借助外部工具的情况下，大模型在思维导图生成、代码生成、图表绘制和音乐及视频创作等领域中均展现出了强大的能力。本节将通过具体示例展示大模型在上述场景下的使用情况，读者可直观地了解大模型如何在不同任务中展现强大的生成与推理能力，从而帮助提升工作与学习效率。更多大模型辅助办公以及创作支持等方面的具体应用案例请参照本书第四章。

2.2.1 思维导图生成

思维导图生成是大模型在知识组织与信息梳理中的一种典型应用。用户只需通过自然语言描述主题或目标，大模型便可自动提取关键信息，构建出结构化的思维导图内容，从而辅助用户理清思路、系统化表达复杂概念。以 Anthropic 公司开发的 Claude 模型为例，用户可以输入类似“帮我生成一个介绍人工智能的思维导图”这样的请求，大模型会根据理解的语义自动构建主题及其子分支，并以清晰的层级方式输出（见图2.1）。此外，一些大模型还支持输出 Markdown 格式的导图结构，用户可以结合如 Xmind、Markmap 等工具，将结构文本快速转换为交互式图谱（参见本书第四章节）。



图 2.1: Claude 模型自动生成思维导图

2.2.2 代码生成

代码生成是大模型在软件开发中的一项重要应用。用户通过自然语言描述问题或需求，可让大模型自动生成高效、可运行的代码片段，也可进行代码纠错。目前大模型支持包括 Python、Java、JavaScript、C、C++、Matlab 和 Ruby 等多种编程语言。这不仅能帮助初学者快速理解和实现算法，还能为资深开发者提供高效的编程辅助，减少重复性劳动。如图2.2所示，用户在编写一个查找列表中最大值的程序时，可能会出现一些不易察觉的错误，借助 DeepSeek 模型就可以快速获得代码调试建议，提升代码编写效率和正确率。



图 2.2: DeepSeek 模型对代码进行纠错

2.2.3 图像绘制

大模型在可视化分析与结构化展示等任务中，同样具备高效的图像绘制能力。用户只需通过自然语言描述所需图像的类型与内容，大模型即可自动生成对应的可视化图表或绘图代码，显著降低了非专业用户制作图像的门槛。例如，在学习二次函数时，用户可以输入“帮我画出 $y = x^2$ 的函数图像，并标注重要点”，ChatGPT 模型即可自动生成对应的函数曲线图（见图2.3），突出拐点与关键特征，帮助学生更直观地理解数学概念，提升学习效果。此外，对于流程结构或概念关系图，大模型还可输出 Mermaid 等结构化格式，导入可视化工具中生成清晰的结构图（参见本书第四章节）。

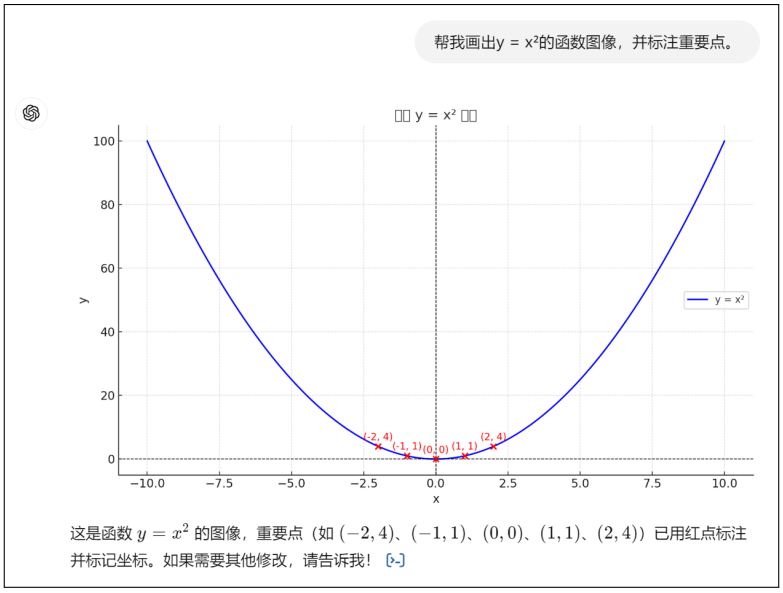


图 2.3: ChatGPT 模型自动绘制数学函数图像

2.2.4 音乐生成

音乐生成是多模态大模型[1]的一个重要应用，结合了文本和音频的多重输入输出形式。用户可以通过自然语言描述、音符提示或情感指令，使大模型生成旋律、和声、节奏等音乐元素，创造出富有创意的音乐作品。

音乐生成大模型降低了普通人群音乐创作的门槛，用户只需通过简单的自然语言描述音乐风格或使用场景，大模型即可生成一首长达 2 至 4 分钟的完整音乐，同时许多模型还具有自动填词的功能。例如，用户使用豆包模型输入“创作一首民谣风格的歌曲，传达放松的情绪，并使用女声音色”时，系统便自动生成符合需求的音乐作品（见图2.4）。

[1]多模态大模型是指能够处理和理解多种类型数据（如文本、图像、音频、视频等）的人工智能模型。这些模型能够同时处理来自不同模态的信息，并在不同类型的数据之间进行跨模态推理和生成。



图 2.4: 豆包模型自动生成音乐作品

通过以上示例可以看出，大模型具备语言理解与生成、逻辑推理以及多模态生成等强大能力。本节仅对大模型的能力作简单介绍，后续章节将深入探索大模型的应用潜力。

2.3 常见大模型

近年来，大模型技术发展迅速，涌现出多个具备强大能力的模型。这些大模型不仅在 NLP领域取得了显著突破，还在多个跨模态任务中展现了广泛的应用潜力。本节将介绍当前一些代表性的大模型，包括国内外的主流模型，并探讨它们的技术特点、应用领域以及适用场景。

2.3.1 DeepSeek

DeepSeek 是由 DeepSeek 公司[2]推出的大模型系列，支持自然语言处理、代码生成、数学推理等任务。DeepSeek 模型支持中英文输入，具备较强的语言理解与生成能力，尤其在中文语境下表现较为出色。

DeepSeek 的首个模型 DeepSeek-Coder[14] 于 2023 年 11 月 2 日发布，主要支持多种编程语言的代码生成、调试和数据分析任务。随后，在同年 11 月 29 日，DeepSeek 推出了其首个通用大模型 DeepSeek LLM[15]，参数规模达到 670 亿，能够处理多种自然语言任务。随着技术不断进步，DeepSeek 相继推出了多个不同版本的模型，进一步优化了性能和应用场景。尤其是 2025 年

[2]DeepSeek 公司全称为杭州深度求索人工智能基础技术研究有限公司，是一家成立于 2023 年 7 月 17 日的中国人工智能初创公司，由国内顶尖量化投资机构幻方量化创立，总部位于杭州市拱墅区。

1 月 20 日发布的 DeepSeek-R1[7]模型，引起了全球人工智能领域的广泛关注。DeepSeek-R1 在数学、编码和逻辑推理等任务上表现出色，其性能与 OpenAI 的 o1 模型媲美。但是与 OpenAI o1 高昂的训练成本相比，DeepSeek-R1 的训练成本仅约为 550 万美元，具有显著的成本优势[3]。此外，DeepSeek-R1 完全开源，采用 MIT 许可证，允许任何人自由使用，包括商业用途，这一举措挑战了许多 AI 公司将模型商品化的传统做法。DeepSeek-R1 的发布对全球科技产业产生了深远影响，英伟达股价在一天内下跌近 17%，市值蒸发约 5890 亿美元，创下美国股市历史纪录[16]。此外，DeepSeek 还发布了多模态大模型 Janus-Pro，进一步拓展了技术应用领域。

目前，DeepSeek 提供网页版[4]和移动端应用（包括 iOS 和 Android 版本），用户可通过浏览器或手机 App 直接访问模型服务，进行对话问答、代码生成等任务。DeepSeek 模型的网页版界面如图 2.5所示，用户在输入框中以自然语言提问，模型会在对话框中返回相应的回答或生成内容。此外，DeepSeek 模型提供了“深度思考”和“联网搜索”两项功能。深度思考模式利用预训练知识库和推理能力，可以对复杂问题进行结构化分析，提供清晰的答案；联网搜索模式则可以实时获取互联网信息，增强回答的时效性，适用于查询最新资讯和实时数据。用户可根据需求选择模式。



图 2.5: DeepSeek 网页版使用界面

DeepSeek 系列包括多个模型版本，如专注于代码生成的 DeepSeek-Coder，强化推理能力的 DeepSeek-R1，以及基于 MoE 架构优化推理效率的 DeepSeek-V2[17]。通过自研的 MoE 架构与分布式训练框架，DeepSeek 在模型效率领域实现了重大突破，其代表性技术成果列在表2.1中。

[3]参考来源：维基百科https://zh.wikipedia.org/wiki/DeepSeek-R1?utm_source=chatgpt.com

[4]DeepSeek 访问网址：https://chat.deepseek.com/

表 2.1: DeepSeek 模型概览

模型名称	技术特点
DeepSeek LLM [15]	千亿参数基座模型，动态专家选择机制，中文语言理解任务超越 GPT-3.5
DeepSeek-Coder [14]	代码专用模型，HumanEval 基准测试通过率 89.7%
DeepSeekMath 7B [15]	数学推理能力强，MATH 基准测试中成绩 51.7%，接近 GPT-4 和 Gemini-Ultra
DeepSeek-VL [18]	多模态模型，跨模态语义对齐
DeepSeek V2 [17]	稀疏激活架构，推理成本降低 83%，支持 128k 上下文窗口
DeepSeek-Coder-V2 [7]	混合专家模型 (MoE)，性能提升显著，支持 338 种编程语言
DeepSeek-V3 [19]	671B 参数，基于 MoE 架构，推理效率提升，适合代码与数学领域
DeepSeek-R1 [7]	开源 MoE 模型，MMLU 基准 87.4 分，推理成本为 GPT-4 的 1/40
Janus-Pro [20]	文本引导的图像生成与编辑, T2I-CompBench 多概念组合生成任务 SOTA

目前, DeepSeek 已成为中国大模型领域的领先者。2025 年 2 月, DeepSeek 启动了“开源周”, 发布了五大代码库, 并推出了 FlashMLA 框架, 支持千卡集群达到 90% 的线性加速效率。此外，公司还推出了 ModelHub 模型市场，已有超过 2 万名开发者上传微调模型，推动了垂直领域应用生态的形成。根据 IDC《2025 中国大模型市场报告》，DeepSeek 在中小企业的渗透率已达 34%，其“技术开源 + 云服务订阅”双轮驱动模式正在重新定义 AI 产业价值链。通过技术突破和商业模式创新，DeepSeek 加速了 AI 技术在教育、医疗、制造等行业的渗透，并且其“模型即服务”（Model as a Service, MaaS）平台已为超过 500 家企业客户提供服务，标志着中国 AGI 技术进入规模化应用的新阶段。

2.3.2 ChatGPT

ChatGPT 是由 OpenAI[5]开发的一款大型语言模型，具备强大的自然语言理解和生成能力。目前，ChatGPT 能够执行多种任务，包括对话问答、文案

[5]OpenAI（开放人工智能研究中心）由非营利组织 OpenAI Inc. 与其营利性子公司 OpenAI LP 组成，成立于 2015 年，位于美国旧金山。

创作、编程辅助、语言翻译以及教育辅导等。

ChatGPT 的首个公开版本于 2022 年 11 月 30 日发布，基于 GPT-3.5 架构，是一个对话式 AI 模型。它一经推出便引发全球关注，在上线后的短短几天内便吸引超过一百万用户注册使用，迅速成为生成式 AI 应用的现象级产品。相比传统的问答系统，ChatGPT 拥有更强的上下文理解能力和逻辑组织能力，能够实现自然流畅、富有层次的对话，被认为是实现“像人一样交流”的重要突破。

自发布以来，ChatGPT 经历了多次重要更新和版本迭代。2023 年 3 月，OpenAI 发布了基于 GPT-4 架构的升级版本，显著提升了模型在理解复杂指令、处理多轮对话和逻辑推理等方面的能力。随后在同年 11 月，OpenAI 发布了 GPT-4-turbo 版本，相比 GPT-4 拥有更快的响应速度和更低的调用成本。2024 年 5 月，OpenAI 推出了一款多模态大模型 GPT-4 Omni（简称 GPT-4o）[13]。该模型支持处理文本、音频、图像和视频等多种模态数据的任意组合输入，并能够生成文本、音频和图像的任意组合输出。与 GPT-4 Turbo 相比，GPT-4o 在非英语文本推理能力方面有显著提升。此外，GPT-4o 的 API 接口在推理速度和响应能力上也得到了增强，且 API 调用费用降低了 50%。相比于 OpenAI 之前的其他模型，GPT-4o 在视觉和音频等多模态任务上的表现更为突出。

除了 GPT 系列，OpenAI 还推出了其他一系列创新的 AI 模型，包括 DALL·E 和 Sora 等。DALL·E 是一款图像生成模型，能够根据用户输入的文本描述生成高质量的图像，广泛应用于创意设计、广告制作和艺术创作等领域。而 Sora 是一款具备多模态能力的模型，能够处理文本、图像等多种输入信息，进行更为复杂的任务处理。

目前，ChatGPT 提供了网页版[6]和移动端应用（适用于 iOS 和 Android 平台），用户可以通过浏览器或手机应用轻松访问其大模型产品，进行对话问答、文本生成、语言翻译等任务，网页版界面如图2.6所示[7]。

[6]ChatGPT 官网：`https://chatgpt.com/`

[7]ChatGPT 在某些地区存在访问限制，目前中国用户可能无法直接访问。



图 2.6: ChatGPT 网页版界面

2023 年 3 月，OpenAI 开放了 ChatGPT 的官方 API 接口库[8]，开发者能够轻松接入并构建各种 AI 应用。通过该 API，开发者可以将 ChatGPT 集成到自己的应用、网站或服务中，利用其强大的自然语言处理能力进行对话生成、自动化客服、内容创作等任务。开放接口的发布大大降低了 AI 应用的开发门槛，推动了生成式 AI 技术在多个行业中的快速普及和应用。

2.3.3 通义千问

通义千问[21]是由阿里云旗下的达摩院自主研发的大规模人工智能语言模型，属于通义大模型家族的重要成员。它基于深度学习和自然语言处理技术，广泛应用于问答系统、写作辅助、代码生成、知识检索等多个场景。如图 2.7 所示，通义千问提供了功能丰富的交互式界面，便于用户进行多样化的语言任务操作。



图 2.7: 通义千问网页版界面

2025 年，通义大模型家族迎来了全新成员——Qwen3 系列模型的正式发布，标志着通义千问在大语言模型技术演进道路上的又一次重要跃升。Qwen3 系列模型引入了“混合推理模式”以适应不同复杂度任务的需求，支持两种工

[8]ChatGPT 官方 API 接口库`https://pypi.org/project/openai/`

作模式：思考模式（Thinking Mode）与非思考模式（Non-Thinking Mode）。在思考模式下，模型会进行逐步推理后再输出答案，适用于复杂、逻辑性强的问题；而在非思考模式下，模型快速响应，适合处理简单、即时性强的任务。这一机制允许用户根据任务需要灵活控制模型的“思考强度”，实现更稳定且高效的推理预算管理。实验证明，Qwen3 在计算资源与推理质量之间展现出良好的可伸缩性和平滑性能提升，使得成本控制与智能表现实现更优平衡。

面向未来，通义将从多个维度持续优化模型体系：包括扩展训练数据规模、增加模型参数、延长上下文长度、丰富多模态能力，并发展具备环境反馈机制的强化学习，实现更远距离、更复杂的推理能力。我们正从“训练模型”的阶段，迈向“训练智能体”的新时代，下一代模型将为人类的工作与生活带来更加深远的智能变革。

2.3.4 豆包

豆包是由字节跳动开发的一款智能助手[9]，于 2023 年 8 月正式上线。该助手具备强大的文本生成、图像和音视频生成、数据分析以及 AI 驱动的在线搜索等功能。与其他智能助手相比，豆包的突出特点在于其对多模态信息的处理能力，不仅支持文本对话，还能够生成图像和视频内容，因此在创作、教育和数据分析等领域具有更广泛的应用潜力。

豆包模型是基于字节跳动自研的大模型（原名云雀模型）开发。自发布以来，豆包经历了多次重要升级。2024 年 5 月，字节跳动在火山引擎原动力大会上正式发布豆包大模型系列，涵盖多个版本，包括通用模型 Pro、通用模型 Lite、语音识别模型、语音合成模型和文生图模型等，显著提升了技术能力。2024 年 12 月，豆包推出视觉理解模型，并将千 tokens 的输入价格降低至 0.003 元，比行业价格低 85%。进入 2025 年，豆包持续进行技术迭代，推出了 V3-0324 版本，优化了代码生成能力，并计划推出 R2 版本，以强化多模态应用。

目前，豆包支持多终端使用，包括网页端和移动端。网页端界面如图2.8所示，豆包网页版集成了多种功能，如 AI 写作、图像生成、AI 编程、视频生成，这些功能可以提升用户的工作和生活效率，满足多方面的需求。

[9]豆包官网：`https://www.doubao.com/chat/`



图 2.8: 豆包模型网页版界面

2.3.5 其它大模型

除了上述提到的 DeepSeek、ChatGPT、豆包和 Kimi，还有许多具有广泛影响力的大型语言模型。这些模型由不同的公司和组织研发，涵盖了自然语言处理、计算机视觉、语音识别等多个领域，并在各自的应用场景中发挥着重要作用。国内的大语言模型，如百度推出的文心一言[22]、阿里巴巴推出的通义千问[21]以及智谱华章推出的智谱清言[23]，主要针对中文语言场景进行了优化，广泛应用于对话生成、文本理解等任务。而国外的大语言模型，如 Claude[12]和 LLaMA[11]，则以其多语言支持和高效训练方法见长，分别在安全性和轻量化方面具有独特优势。

表 2.2汇总了部分主流大语言模型的基本信息，包括模型名称、开发公司及其访问网址。

表 2.2: 其他常用大语言模型概览

大模型名称	公司名称	网址
文心一言[22]	百度	https://wenxin.baidu.com
Kimi	月之暗面	https://kimi.moonshot.cn/
智谱清言[23]	智谱华章	https://chatglm.cn
Claude[12]	Anthropic	https://www.anthropic.com
LLaMA[11]	Meta	https://www.llama.com

2.4 大模型分类

大模型的种类繁多，随着技术的进步，其在应用、规模和能力上展现出丰富的差异。为了更好地理解不同类型的大模型及其适用场景，本节将从运行机制、参数规模和模态支持等方面，介绍大模型的主要分类，包括对话型与推理型大模型、基础与轻量化大模型以及单模态与多模态大模型。

2.4.1 对话型与推理型大模型

目前大模型主要被分为“对话型”和“推理型”两大类别。对话型大模型是专门用于人机自然对话交互的模型。这类模型能够满足日常对话、信息检索、内容创作、简单推理等大部分需求，特点是注重高效便捷、通用性强、上手门槛低。典型代表有 DeepSeek-V3 和 GPT-4o 等模型。推理型大模型主要专注于逻辑推理、问题求解等方面，在需要步骤性思考、数学问题解决、编程以及其他需要强推理能力的任务中表现突出。推理型大模型的特点是注重结构化思维、逻辑分析并可以展示推理过程，能够解决复杂或开放性问题，典型代表有 DeepSeek-R1、GPT-o1 等模型。

对话型大模型的优势在于适用性广泛，能够快速响应用户需求，并在语言理解和生成任务上表现出色。然而，它的深度推理能力相对有限，复杂逻辑问题可能需要更多外部支持。而推理型大模型的优势在于分析能力强，适用于数学、科学和编程等需要严密逻辑推理的场景，但可能在开放性对话和创意生成方面不及对话型模型灵活。表2.3展示了两者在模型定位、推理能力、多模态支持、应用场景以及用户交互体验等方面的主要差异。

表 2.3: 对话型与推理型大模型比较

比较项	GPT-4o（对话模型）	DeepSeek-R1（推理模型）
模型定位	适用于对话、多模态信息处理等任务。	侧重推理和逻辑分析，擅长数学、编程等任务。
推理能力	在复杂推理任务中表现一般。	擅长复杂推理任务，尤其是数学和代码推理。
多模态支持	支持文本、图像、音频等多模态输入。	仅支持文本输入，缺乏多模态支持。
应用场景	适合通用任务，如对话和内容生成。	适合高精度推理和专业领域任务。
用户交互体验	提供流畅对话体验，适合大众使用。	适合专业用户，交互节奏较慢。

2.4.2 基础与轻量化大模型

随着大模型在各类任务中展现出强大能力，开发者开始积极探索如何在保证模型性能的同时降低对计算资源的依赖，推动其更广泛地落地应用。因此，

模型开发逐渐分化出两类主要形态：基础大模型与轻量化大模型。

基础大模型通常是指参数规模庞大、功能强大、适用于多种复杂任务的通用模型。它们往往在多轮对话、复杂推理、多模态理解等任务中表现出色，但同时也对硬件资源、推理速度和部署条件提出了较高要求。为了更好地适配算力有限的应用场景，开发者们提出了轻量化建模策略。轻量化模型通过剪枝、知识蒸馏（Knowledge Distillation）等技术手段，在尽量保留大模型核心能力的前提下，显著减少了参数量和计算开销。其中，知识蒸馏是目前应用最广泛的方式之一，即将基础大模型的知识迁移到一个更小的模型中，使后者具备较强的性能但推理效率更高。

表2.4列举了 DeepSeek 系列基础大模型及其蒸馏后的轻量化版本。

表 2.4: 基础大模型与轻量化模型对比

模型名称	描述	蒸馏模型	参数量
DeepSeek-R1-671B	基础版	-	671B
DeepSeek-R1-Distill-Qwen-7B	蒸馏版	Qwen-7B	7B
DeepSeek-R1-Distill-Llama-8B		Llama-8B	8B
DeepSeek-R1-Distill-Qwen-14B		Qwen-14B	14B
DeepSeek-R1-Distill-Qwen-32B		Qwen-32B	32B
DeepSeek-R1-Distill-Llama-70B		Llama-70B	70B

在实际应用中，轻量化大模型特别适合部署在边缘设备、移动终端或低延迟响应场景，如智能助手、推荐系统和工业控制系统等。在资源受限的环境下，轻量化大模型能够在不牺牲过多精度的情况下，提高推理速度和降低计算负担。

2.4.3 单模态与多模态大模型

单模态大模型是指仅能处理单一类型数据（如文本、图像或音频）的模型。此类模型通常专注于某一模态的任务，在其特定领域内具备较高的性能。例如，文本生成与理解、图像分类和语音识别等任务都依赖于单模态模型进行高效处理。由于任务集中、结构相对简单，单模态模型在资源受限或需求明确的场景中具有明显优势。

相比之下，多模态大模型能够同时处理多种类型的数据，例如将文本、图像、音频或视频等模态进行联合建模。这类模型可以在不同模态之间建立关联，从而提升对复杂信息的理解与生成能力。常见的应用场景包括图文生成（通过文本生成相关的图像或通过图像生成描述性文本）、视频内容分析（融合图像

与音频进行情感理解）、以及语音合成与增强（结合语音与文本实现更自然的表达）。

表2.5总结了单模态大模型与多模态大模型在处理能力、任务专注性、性能、应用场景以及计算成本等方面的主要区别。从表格中可以看到，单模态大模型通常在专注任务上表现优秀，但多模态大模型则能处理跨模态任务，但相应地也带来了更高的计算成本。

表 2.5: 单模态与多模态大模型对比

比较项	单模态大模型	多模态大模型
处理能力	仅处理单一数据模态	处理多种数据模态
任务专注性	高，专注单一任务	低于单模态，但可完成复杂任务
性能	在专注任务上性能较好	在跨模态任务中表现优越
应用场景	适用于文本生成、图像识别等单一模态任务	适用于图文生成、视频分析等复杂任务
计算成本	计算资源消耗较小	计算资源消耗较大

2.5 大模型本地部署

大模型本地部署，顾名思义，即将大模型安装并运行于用户本地设备，如笔记本电脑、台式机或私有服务器等。在云端模型服务普及的背景下，为什么要耗费本地资源部署私有大模型？大模型本地部署主要具有以下两方面优势：一是支持在无网络环境下使用大模型，满足离线环境下的应用需求，例如缺乏网络连接的工业生产场景或保密科研场所；二是保护数据隐私，有效规避敏感数据在云端传输、存储过程中可能产生的泄露风险。针对具有私有化部署大模型需求的用户而言，如何便捷地部署本地大模型是关键问题。一般来说，本地部署主要有以下两种方式，第一种是基于开源框架通过代码级环境配置实现模型部署，该方案要求实施者具备基础的软件开发能力，涉及 Python 编程、依赖库管理、环境变量配置等技术环节；另一种则是通过应用端部署，比如 Ollama[10]或者 LM Studio[11]等图形界面完成模型加载与参数配置，具有低代码化、可视化特点，显著降低技术门槛。因此，本节将以 Ollama 部署工具为例，详细讲解大模型本地部署的实现步骤和操作方法，并介绍如何调用本地部署的大模型。

[10]Ollama 官网: `https://ollama.com/`

[11]LM Studio 官网: `https://lmstudio.ai`

2.5.1 Ollama 部署工具

Ollama 是一款可以让用户在本地运行大规模语言模型的工具，能够帮助开发人员和企业在无需依赖云服务的前提下，私密地部署和操作 AI 模型。Ollama 支持多种模型，包括 Llama[24]、Mistral[25]、Code Llama[26] 等，广泛应用于构建聊天机器人、自动化编码任务或离线研究等场景。Ollama 支持 macOS、Linux 和 Windows 系统，为在本地环境中管理和使用模型提供了高度灵活的方案，特别合适法律、医疗、科研等对数据隐私有高要求的应用场景。同时，它也能够与其他软件平台很好地集成，增强现有的工作流程。

Ollama 主要通过命令行界面 (Command Line Interface, CLI)与模型交互，同时也支持借助第三方图形用户界面 (Graphical User Interface, GUI)来进行操作,适合喜欢更直观方式的用户。用户可通过命令 `ollama pull <model-name>` 可以下载模型，使用 `ollama run <model-name>` 来运行模型。安装 Ollama 具体过程如下：

1. **访问官网**：前往 Ollama 官网（`https://ollama.com/`），点击“Download”选择下载适用于用户操作系统的版本。官网界面如图2.9所示。

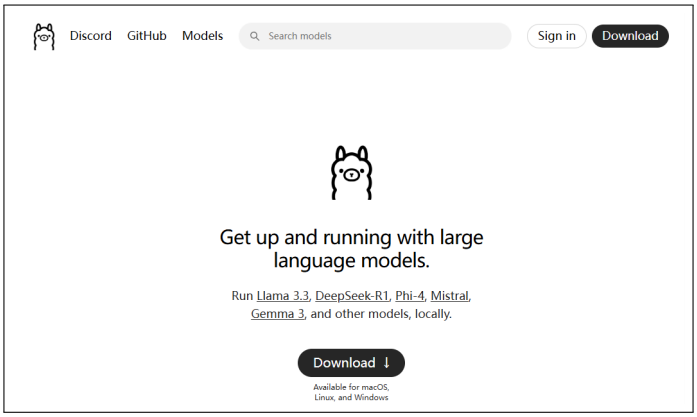


图 2.9: Ollama 官网

2. **下载并安装**：下载界面如图2.10所示，根据使用需求选择对应平台的安装包进行下载，包括 macOS、Linux 或 Windows（注意：Windows 版本当前处于预览阶段）。下载完成后，按照操作系统的安装向导完成安装。

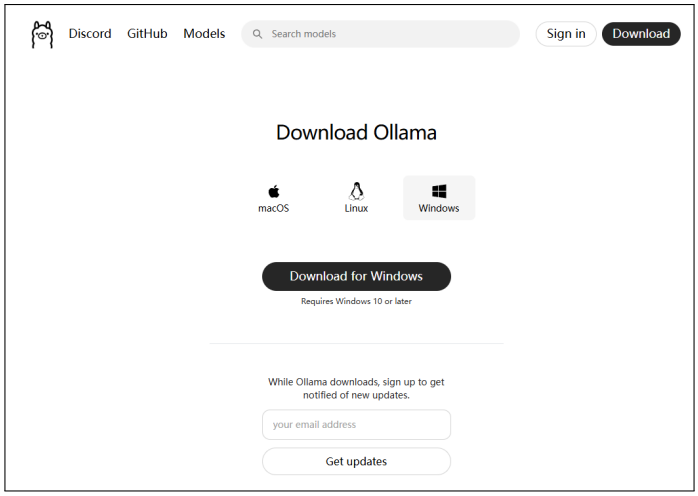


图 2.10: 选择并下载所需的版本

3. **启动并验证**：安装结束后，按下键盘上的 Win + R 组合键，打开运行窗口，输入 cmd 命令启动命令提示符窗口。在命令提示符窗口内输入 `ollama --version` 来验证安装是否成功。如果命令返回版本号信息，则表示 Ollama 已成功安装并可正常使用，如图2.11所示。



图 2.11: 验证 Ollama 安装是否成功

2.5.2 Ollama 命令行交互

Ollama 的一项核心功能是支持用户便捷地运行各类预训练语言模型。用户先打开 CLI，使用命令 `ollama pull <model-name>` 下载所需模型，或直接使用 `ollama run <model-name>` 命令运行模型（若尚未下载，将自动完成下载并运行）。所需的模型可以通过 Ollama 官网右上角的“Models”页面访问，或直接访问模型库网址（`https://ollama.com/search`），界面如图2.12所示。

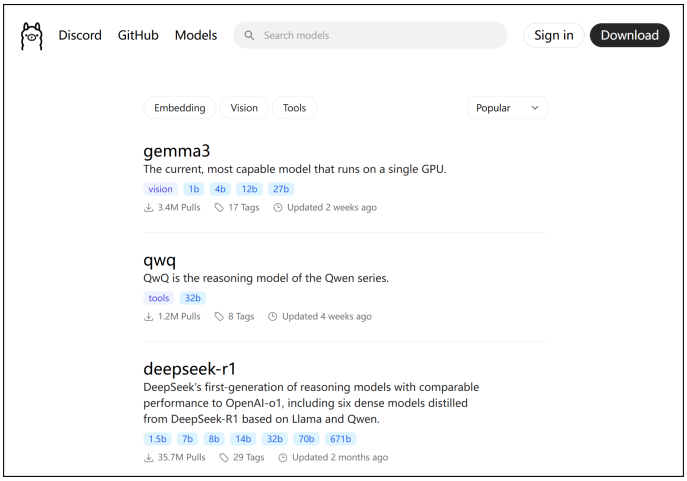


图 2.12: 模型库界面

以 qwen2.5-coder 的下载为例，点击 qwen2.5-coder，进入 qwen2.5-coder 的介绍界面，复制 `ollama run qwen2.5-coder` 指令，输入到已经打开的命令提示符窗口中，按下 enter 键开始安装，安装完成界面如图2.13所示。

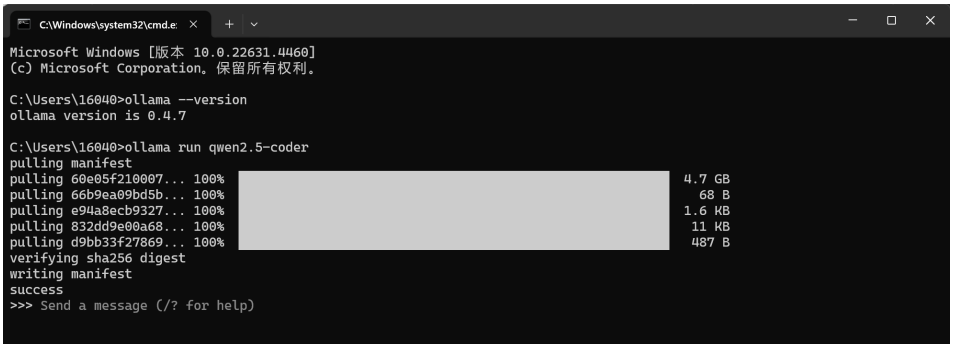


图 2.13: 安装 qwen2.5-coder

模型安装完成后，用户可直接在命令提示符窗口中输入提示词与模型进行交互，如图2.14所示。交互结束时，按下 Ctrl+D 组合键即可退出该模型的使用。

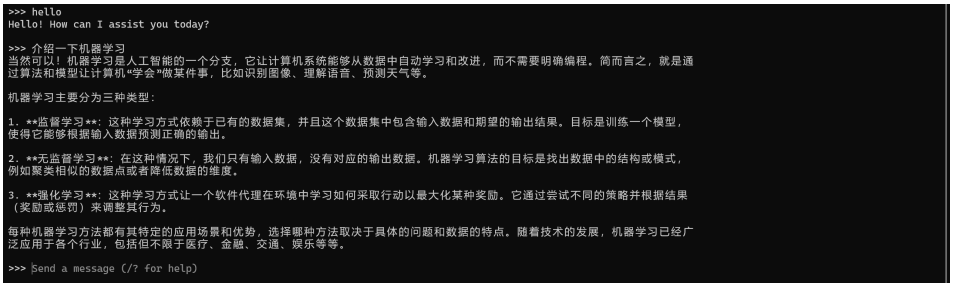


图 2.14: 使用 qwen2.5-coder

2.5.3 Ollama 可视化界面交互

为了提升使用体验，用户可以为 Ollama 搭建 GUI。与 CLI相比，GUI 提供了更加直观和友好的交互方式，使得用户能够轻松加载模型、输入提示词并查看输出结果。在搭建 Ollama 的可视化界面时，可以使用 Node.js 作为开发环境。Node.js 是一个跨平台、开源的 JavaScript 运行环境，其提供跨平台兼容性、高性能以及丰富的生态系统，能够简化开发过程并加速应用的构建，进一步提升用户体验。

1.Node.js 下载与安装

用户可以访问 Node.js 工具官网（`https://nodejs.org/zh-cn`）下载并安装适合用户操作系统的 Node.js LTS 版本，官网界面如图2.15所示。



图 2.15: Node.js 官网界面

安装完成后，可以通过命令 `node -v` 和 `npm -v` 检查 Node.js 和 npm 是否成功安装，如果命令行返回版本号信息，则说明安装成功，如图2.16所示。



图 2.16: 检查 Node.js 和 npm 是否成功安装

2.Web 可视化界面搭建

成功安装 Node.js 后，用户可通过 GitHub 下载 Ollama WebUI 的源代码，并启动本地服务实现可视化操作。具体操作步骤如下：

1. **下载项目代码**：打开命令行窗口，执行以下命令获取 WebUI 的源代码，将其下载到本地目录“ollama-webui”中，下载过程如图2.17所示。

```
git clone https://github.com/ollama-webui/ollama-webui-lite ollama-webui
```



图 2.17: 下载 Ollama WebUI 代码

2. **进入 Ollama WebUI 代码目录**：在命令行中输入 `cd ollama-webui` 切换到用户刚才下载的 WebUI 代码目录，如2.18所示。



图 2.18: 进入 Ollama WebUI 文件夹

3. **安装 Node.js 依赖的工具包**：在代码目录下执行命令 `npm install` 来安装所需的依赖包。该命令会自动根据项目的“package.json”文件下载所有必要的依赖，安装过程如图2.19所示。

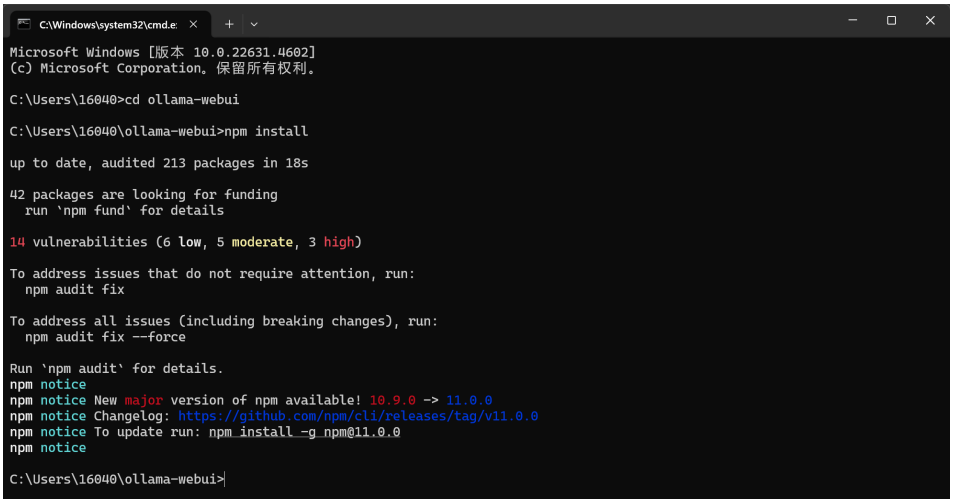


图 2.19: 安装 Node.js 依赖的工具包

4. **启动 Web 可视化界面**：依赖安装完成后，执行命令 `npm run dev` 启动 WebUI。启动成功后，如图2.20所示。

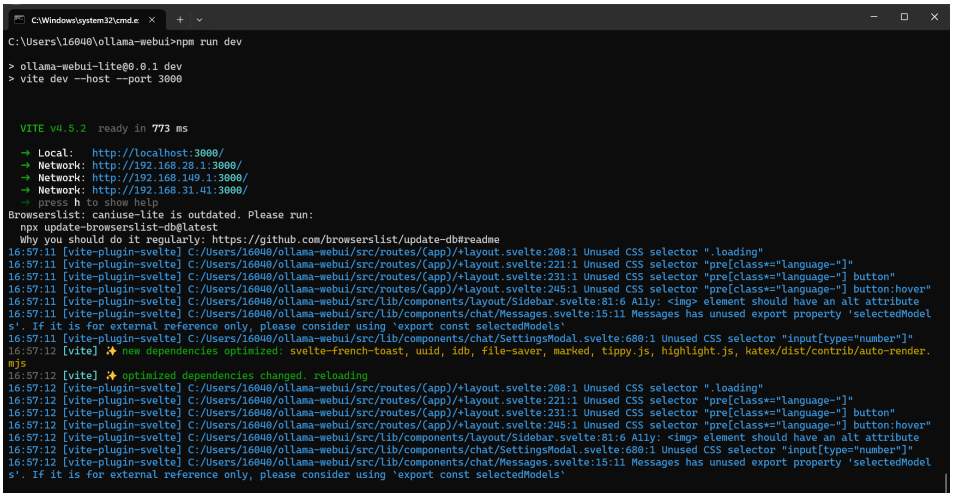


图 2.20: 启动 WebUI

用户将在命令行中看到以下类似信息:

```
Listening on http://localhost:3000
```

默认情况下，WebUI 会在本地服务器的 3000 端口启动。因此，用户可以在浏览器中访问 `http://localhost:3000` 查看 WebUI。需要注意的是，如果端口 3000 已被占用，用户可以更换端口。例如，用户可以在命令行中输入以下命令:

```
npm run dev --port 8080
```

然后在浏览器中打开`http://localhost:8080` 即可进入可视化界面，加载模型并开始交互，打开界面如图2.21所示。

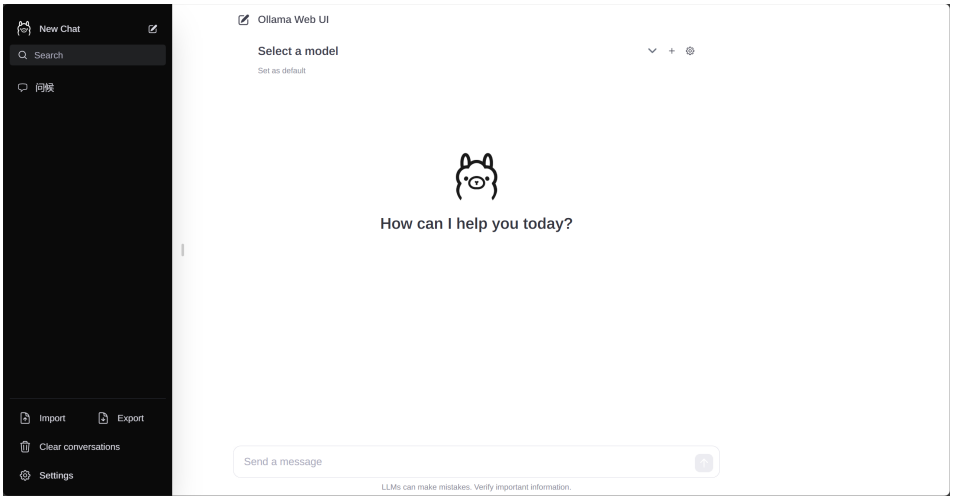


图 2.21: WebUI 的欢迎页

3. 模型使用

在 Web 可视化界面成功搭建并运行后，用户可以在页面中选择已安装的模型进行使用。界面通常提供下拉菜单或搜索框，便于用户快速定位所需模型。

用户可以点击选择模型按钮，选择先前已安装好的 qwen2.5 模型，如图2.22所示。

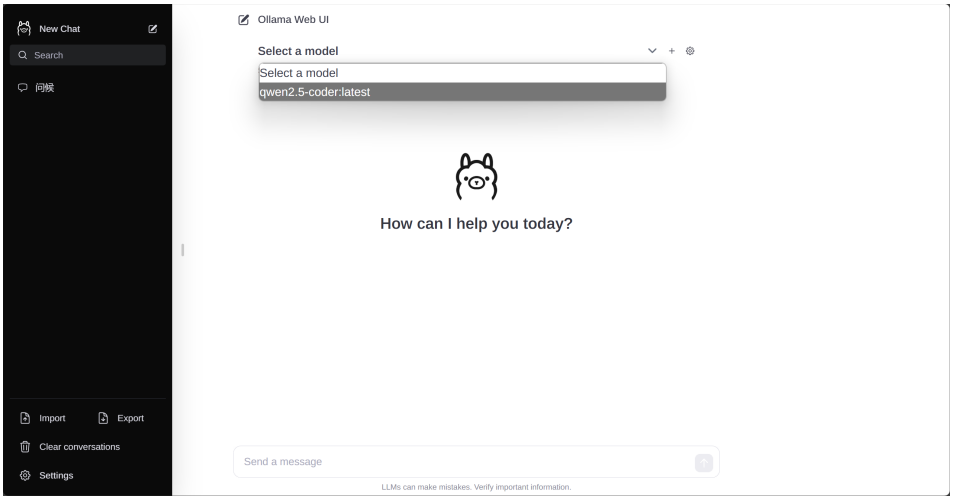


图 2.22: 选择已安装好的模型

选择好 qwen2.5 模型后，在输入框中输入提示词，点击“发送”按钮，即可获得模型生成的响应，如图2.23所示。



图 2.23: 在 WebUI 中使用大模型

2.6 大模型客户端

尽管本地部署和可视化界面为用户提供了较高的自主性，但对计算资源和配置能力有一定要求。对于希望快速上手、便捷访问多种大模型功能的用户而

言，使用集成化的大模型客户端是一种高效选择。

Cherry Studio[12] 是一款集成多模型对话、知识库管理、AI 绘画、翻译等功能于一体的全能 AI 助手客户端。如图2.24所示，它为用户提供了一个简单易用的界面，用户可以通过它访问目前主流的大语言模型。

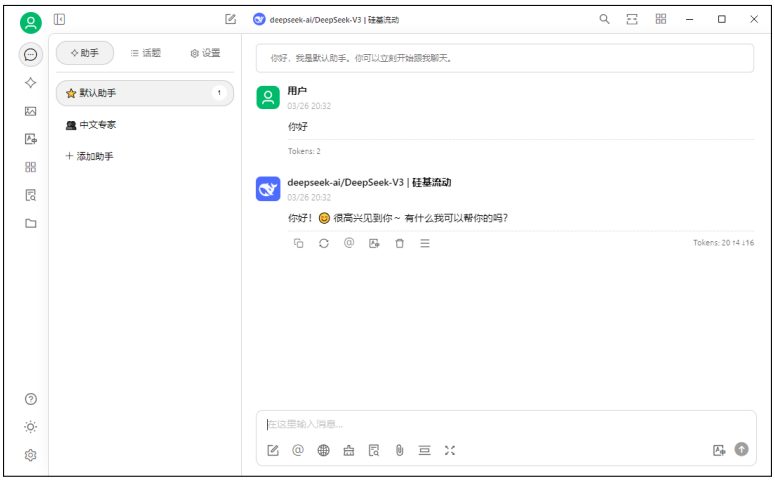


图 2.24: Cherry Studio 大模型客户端

2.6.1 大模型配置方法

本小节将介绍如何通过 Cherry Studio 客户端，通过 API 的方式使用目前主流的大语言模型，如 DeepSeek-R1 等。用户在使用在线大模型前，需要先从相关服务提供商处获取 API 密钥，并将其配置至 Cherry Studio 中，实现模型调用功能。

配置流程如图2.25所示，具体包括以下步骤：

1. 启动 Cherry Studio 应用程序并点击界面左下角的设置图标，进入设置界面；

2. 在设置界面中，用户将看到多个选项卡，请选择 “模型供应商” 或类似的选项卡。在模型列表中选择对应的模型供应商（如硅基流动、DeepSeek 等）；

3. 在相应模型项下填写 API 密钥和 API 地址（部分服务可能已预配置地址），完成配置。

[12]Cherry Studio 官网：`https://cherry-ai.com/`

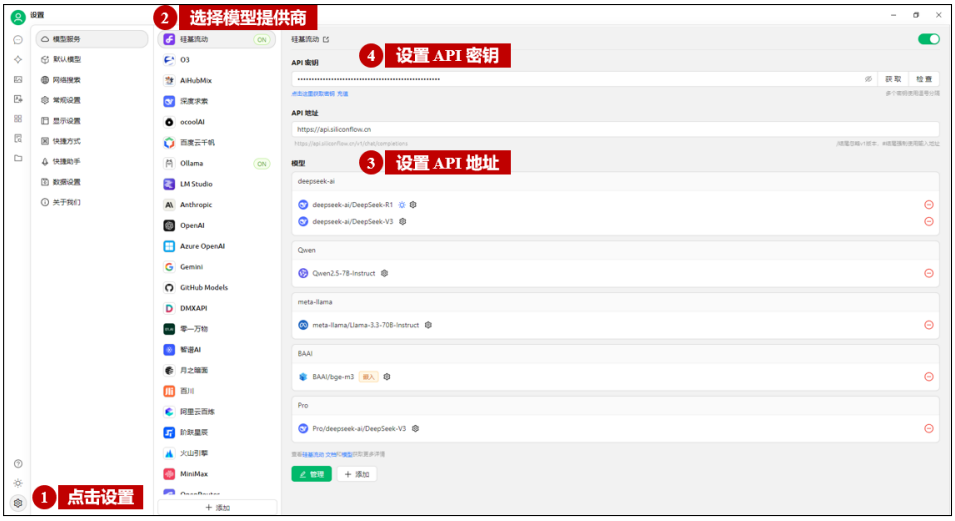


图 2.25: 在线大模型 API 配置流程

API 密钥用于验证用户身份和访问权限，是连接客户端与云端模型服务的关键凭证。不同平台的密钥申请流程略有差异。以硅基流动为例，用户首先需要前往硅基流动官网[13]注册账户。注册完成后，登录账户并导航至“个人中心”页面。如图2.26所示，用户首先点击“API 密钥”打开 API 密钥管理界面，然后点击“新建 API 密钥”，系统将生成一组唯一的密钥。用户点击复制密钥后，返回 Cherry Studio 的“设置”页面，在模型供应商中选择“硅基流动”，并将密钥粘贴到对应字段中。同时，确保 API 地址已正确填写为硅基流动提供的端点地址。

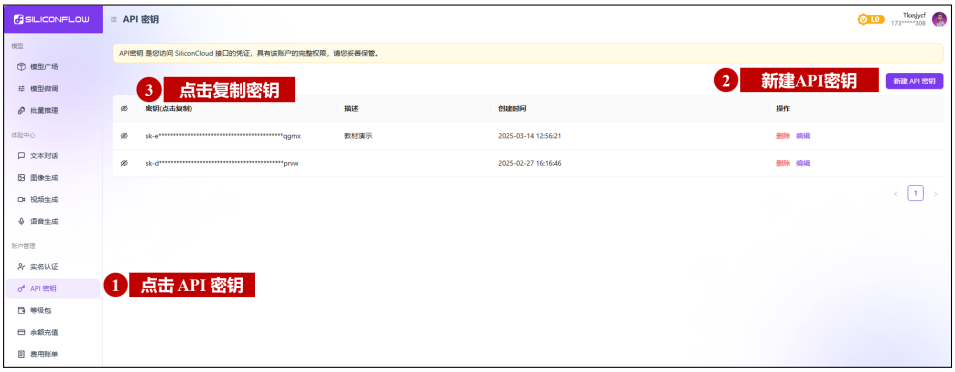


图 2.26: 硅基流动 API 密钥获取流程

[13]硅基流动官网: `https://cloud.siliconflow.cn/`

填写完所有信息后，点击“检查”按钮来测试 API 连接。Cherry Studio 将尝试使用用户提供的凭据与服务建立连接。如果成功，用户将看到绿色的成功提示；如果失败，需要检查用户的 API 密钥和地址是否正确，或查看错误信息获取更多诊断信息。验证成功后，点击“确认”按钮完成 API 设置。

1.API 添加自定义模型

API 提供商（如硅基流动、OpenAI 等）通常支持多种语言模型。若用户希望使用特定模型，Cherry Studio 提供了自定义模型配置功能。

如图2.27所示，用户可登录 API 提供商平台（例如硅基流动），进入“模型广场”，浏览可用模型列表。找到目标模型（如 DeepSeek-V3）后，点击进入详情页面，复制其模型标识符，后续将在 Cherry Studio 中进行配置。

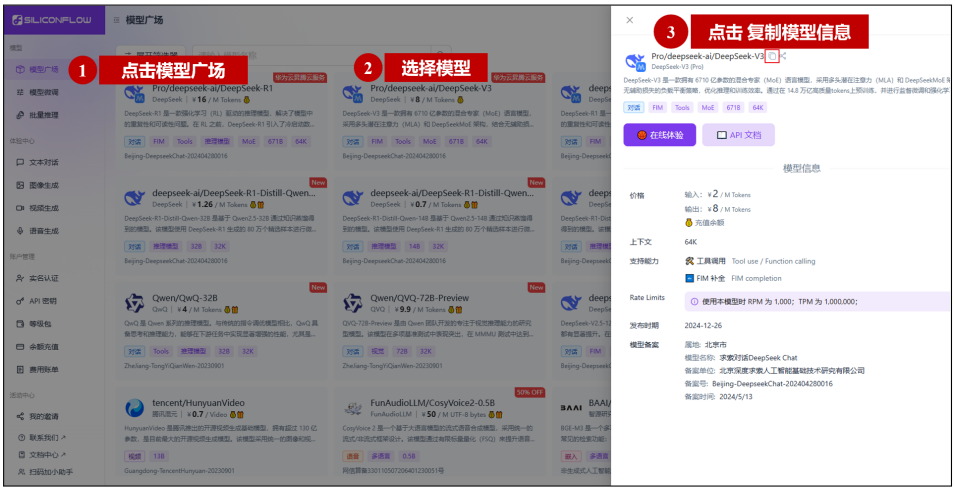


图 2.27: 硅基流动“模型广场”

如图2.28所示，返回 Cherry Studio 设置界面，找到“添加模型”选项。点击“添加”按钮，在弹出的表单中填写以下信息：

- 模型名称：可自定义填写，便于用户识别；
- API 提供商：选择此前已配置的服务提供商（如硅基流动）；
- 模型标识符：粘贴从提供商官网复制的模型 ID 或名称。

某些模型可能需要其他参数，如温度、最大词元数等，用户可以根据需要进行设置，或保留默认值。填写完毕后，点击“保存”按钮完成自定义模型添加。

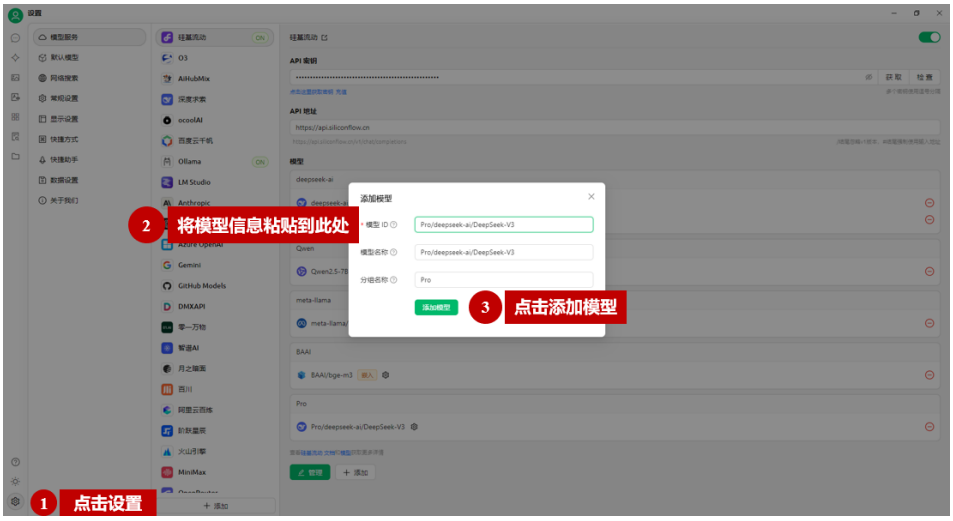


图 2.28: Cherry Studio 客户端在线模型添加

添加完成后，用户可以在创建聊天或使用助手时选择此模型，如图2.29所示。如果模型需要特定参数或有使用限制，Cherry Studio 通常会提供相关提示或配置选项。



图 2.29: Cherry Studio 客户端模型使用界面

2. 本地模型 Ollama 设置

除了通过云端 API 接入主流大语言模型, Cherry Studio 也支持与 Ollama 集成, 使用本地部署的模型。这一方式适合对隐私有较高要求，或在离线环境下仍需使用 AI 的用户。

首先,用户需要在计算机上安装 Ollama,具体安装步骤可参考本书第2.5.1小节。安装完成后，Ollama 服务通常会自动启动，用户也可通过命令行或桌面应用手动启动该服务。

启动 Ollama 服务后, 用户可在 Cherry Studio 中配置连接信息, 实现本地模型的调用。如图2.30所示, 在 Cherry Studio 中进入设置界面, 找到 “Ollama” 选项卡。在该页面中，用户需要配置 Ollama 服务的连接地址。如果 Ollama 运行在本地默认设置下，Cherry Studio 可能会自动检测并填写地址（通常是 http://localhost:11434)。若 Ollama 部署在自定义端口或远程服务器上，用户需手动输入正确的 URL 地址来完成连接配置。

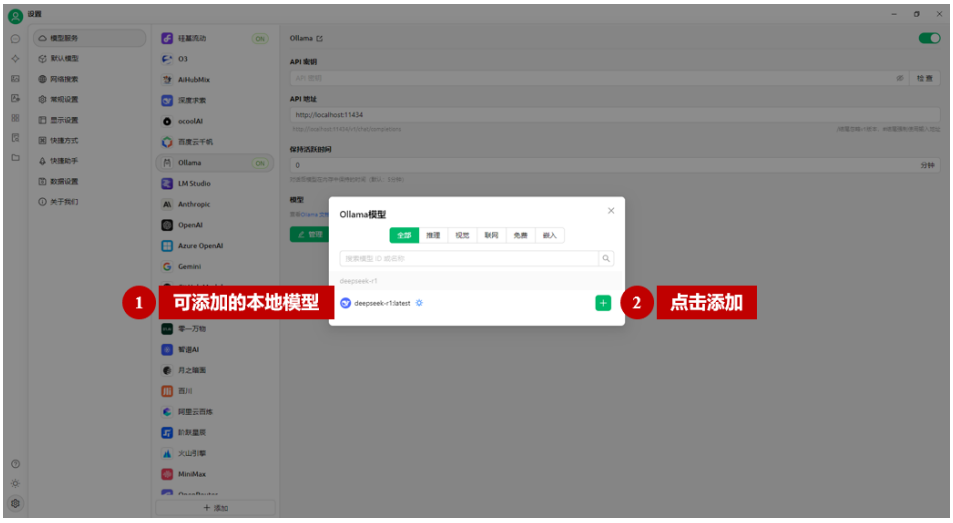


图 2.30: Ollama 本地模型添加

在配置完 Ollama 服务的连接信息后，用户可以点击 “连接” 按钮验证 Cherry Studio 是否能成功连接到 Ollama 服务。如果连接成功, Cherry Studio 将显示 Ollama 上可用的模型列表。如果列表为空，用户需要先在 Ollama 中下载模型。如若需要添加新的本地模型，用户可以在 Ollama 的界面中，或通过命令行执行 `ollama pull <model-name>` 命令来下载模型（例如，使用 `ollama pull llama2` 或 `ollama pull mistral` 命令)。模型下载完成后，回

到 Cherry Studio 的 Ollama 设置页面，刷新模型列表，用户便可以看到新添加的模型。

当用户创建新的聊天或配置助手时，本地 Ollama 模型将作为选项出现在模型选择列表中。用户可以像使用云端模型一样，轻松地选择并使用本地模型，无需互联网连接，且所有数据均保留在用户的设备上。

2.6.2 内置 Web 访问服务

Cherry Studio 不仅是一个 API 客户端，它还集成了一个完整的浏览器环境（如图2.31所示）。用户无需切换应用，即可便捷使用各类 AI 服务。此外，Cherry Studio 还集成了多种实用功能，进一步优化了用户的操作流程，提供了一站式 AI 工具使用体验。

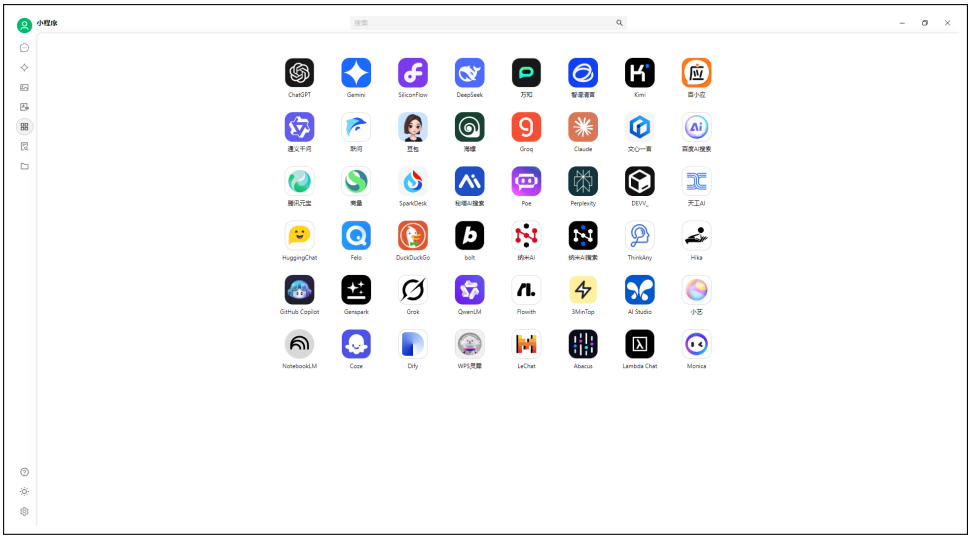


图 2.31: 内置 Web 服务访问

Cherry Studio 的内置浏览器功能让用户能够在不切换应用的情况下，直接访问各种 AI 网络服务。具体使用方法如下：

1. 启动 Cherry Studio 后，用户需要在主界面上找到侧边栏中的“小程序”选项；

2. 点击“小程序”选项后，用户将看到一个预配置的 AI 服务列表。该列表包括了如 Claude、Perplexity、Poe 等流行的 AI 网络平台。用户可以根据需求选择想要访问的服务。

3. 选择目标平台后，Cherry Studio 会在内部窗口中加载该网站。此时，用户无需切换到外部浏览器，所有操作都可以在 Cherry Studio 内部进行。

Cherry Studio 的内置浏览器环境完全支持网页交互，用户可以像在常规浏览器中一样登录账户，点击网页上的登录按钮并输入用户名和密码。Cherry Studio 还可以存储用户的登录状态，下次使用时无需重新登录。一旦登录成功，用户可以直接使用该服务提供的所有功能。

此外，内置浏览器支持基本的浏览功能，如前进、后退、刷新和复制内容等。用户可以在地址栏中输入 URL 访问其他网站，或利用书签功能保存常用的 AI 服务网址。Cherry Studio 会保持网页会话状态，即使用户暂时切换到应用的其他部分，也可以随时回到之前的网页继续操作。对于经常使用的 Web 服务，用户可以将其添加到 Cherry Studio 的快速访问列表。在浏览器界面中找到“添加到收藏”或类似选项，将当前网站添加为收藏项。添加后，该服务将出现在快速访问列表中，用户下次只需一次点击即可直接打开。

2.7 讨论

讨论 2.1. 本章介绍了多种大语言模型及其使用方式，请在实际体验这些模型后总结其在实际应用中的适用场景与不足之处。例如，可以结合科研、编程、写作等不同应用场景分析它们的优劣势。

讨论 2.2. 对比本章中提到的各大模型，探讨它们在效率、精度和扩展性方面的表现差异。例如，某些模型可能在推理速度上占优，而另一些模型则可能在长文本理解上更强。

讨论 2.3. 随着大模型的发展，学生获取知识的方式正在发生变化。在未来的教育环境中，你认为人类教师的角色会发生哪些变化？大模型是否可能部分替代教师？请讨论你的观点并给出理由。

讨论 2.4. 当前大模型技术仍然存在伦理和安全问题，例如数据偏见、误导性信息、隐私风险等。在教育和科研领域，你认为应该采取哪些措施来应对这些挑战？

2.8 习题

习题 2.1. 结合你的日常学习经历，讨论大模型如何帮助你提升学习效率，并举例说明具体的应用场景。例如，如何利用大模型进行编程辅助、论文润色、

数据分析等任务?

习题 2.2. 你认为大模型在未来的高等教育中可能会带来哪些新的学习方式?请结合当前技术发展趋势进行分析，如智能教学助手、自适应学习系统等。

习题 2.3. 在选择大模型时，不同的应用场景可能会影响模型的选择。请从科研、编程、写作三个角度，分别说明你会选择哪种大模型，并简要分析其优缺点。

习题 2.4. 现有大模型在实际应用中仍然存在局限性，例如幻觉问题、上下文理解错误等。你认为哪些技术突破可能会优化这些问题?请结合当前主流大模型的特点进行探讨。

习题 2.5. 你认为“对话型”大模型和“推理型”大模型的主要区别是什么?在教育或工作场景中，哪种类型的模型更有优势?请结合实际案例进行分析。

习题 2.6. 多模态大模型能同时处理文本、图像、音频等多种数据类型。在你的学习或研究领域，你认为多模态大模型有哪些潜在应用?请举例说明。

习题 2.7. 本地部署大模型的优势之一是保护数据隐私，但这也可能带来计算资源的消耗。你认为在企业或个人使用场景下，应该如何权衡隐私保护与计算成本之间的关系?

习题 2.8. 比较命令行界面（CLI）与图形用户界面（GUI）在使用大模型时的优势与劣势。

习题 2.9. API 接口使得用户可以更加便捷地调用大模型进行任务处理。你认为 API 方式相比本地部署有哪些优势和劣势?请结合实际应用场景分析。

习题 2.10. 在 Cherry Studio 中，如何添加一个来自硅基流动的自定义模型?请简述主要步骤。

第三章　大模型提示词

3.1　概述

提示词（Prompt）是用于引导大语言模型生成特定输出的文本输入。在大模型使用过程中，提示词是用户与模型交互的核心桥梁。用户通过输入一段文字（可以是一个问题、命令或描述）来引导模型生成符合需求的内容。例如，用户输入“写一段介绍大模型智能体的短文”，模型将生成一段简明扼要的介绍，概述大模型智能核心组件是什么，如规划、记忆、工具和执行。提示词的作用有很多，除了指导模型理解任务目标外，还可以提供问题的背景信息，定义输出格式（如列表、段落、代码）以及设定约束条件等。模型的输出内容取决于用户提供的信息，即提示词的构建对模型所输出的内容质量具有决定性影响。因此，学会编写优质提示词是掌握提示工程技术的基本技能。那么什么样的提示词是高质量的呢？在设计提示词时，用户又应该遵循哪些规则？

为解答上述问题，本章将从三个方面介绍提示词的设计要义，包括提示词要素、设计技巧和提示词案例。在提示词组成要素方面，本章将介绍提示词中的重要组成部分，包括任务目标、角色设置、上下文信息、输出控制和推理模型提示词等内容。随后，本章将进一步讨论提示词的设计技巧，探讨在不同任务类型下的提示词设计策略以及大模型提供的一键优化功能来达到优化提示词的效果。此外，本章还将提供多个实际场景的提示词案例，包括写作助理、法律咨询、头脑风暴和学习辅助等应用场景。

3.2　提示词组成要素

提示词的组成要素通常包括任务、角色、上下文和输出控制等[1]，其核心作用是指导模型理解任务目标。因此，除了任务是必要组成要素以外，用户可

[1]严格来讲，提示词组成目前并无固定规则，一些研究中会把说明、问题、输入数据、示例、对话历史记录、即时任务描述等分别列出。

根据具体需求，选择简化提示词，通过仅描述任务要求[2]，或者额外提供背景信息、定义输出格式或调整回答风格等其他细节来优化生成结果。

例如，用户想让语言模型帮忙撰写一篇关于人工智能伦理的短文，可以直接输入包含任务的提示词，也可以增加角色、背景等相关信息控制大模型的输出。如图3.1所示，两种提示词得到的内容输出会存在差异。为了使读者更好地理解提示词，本节接下来将着重介绍各要素的功能。



图 3.1: 针对两种提示词使用 Kimi 模型生成的内容对比

3.2.1 任务

在提示词设计中，“任务”是指用户希望大模型执行的具体操作或目标，如生成文本、提取信息、处理数据或者对话交互等。用户可以使用一些简单的命令来指示模型执行各种任务，如“翻译”、“解释”、“总结”和“对话交互”等[3]。

那么如何通过提示词传达给大模型具体的任务呢？一般来说有指令和提问两种方式。例如，用户想要模型帮忙快速浏览一篇文章，可以直接指示模型执行操作，或以提问方式引导，如下所示：

指令：总结下面文章。

提问：这篇文章的核心内容是什么？

[2]推理模型 DeepSeek-R1 仅需描述包含“任务”要素的提示词，详见3.2.5小节。

[3]参考资料：https://guide.lidayun.com/prompt/prompts_tutorial.html

两种方式都能有效指导模型输出内容。指令方法的特点是直接明确，更适用于需要特定输出的任务，如翻译、分类、排序等，这种方式可确保结果结构化且可控。提问则更具开放性，能引导模型思考并生成灵活的回答，更适用于信息查询、知识问答或对话交互。例如，相比直接指令“解释如何重置智能手机的步骤”，使用更具创造性的问题“想象如果手机能感知主人的情绪，它会如何改变我们的日常交流方式？”，可以激发模型的联想能力，使回答更具趣味性和启发性。具体方式的选择取决于任务需求和用户的交互偏好。

为避免模型输出失败，任务需要明确清晰，确保指令或提问能够有效引导模型执行预期的操作。OpenAI 在其官方提示工程指南[4]中指出“The less the model has to guess at what you want, the more likely you’ll get it（模型需要猜测你意图的地方越少，得到你想要结果的可能性就越大）”。换句话说，任务表达应尽量具体，避免让模型猜测用户的目标，确保它能够准确理解并遵循指令。通过清晰的任务描述，可以极大提高模型生成准确和符合预期内容的概率。

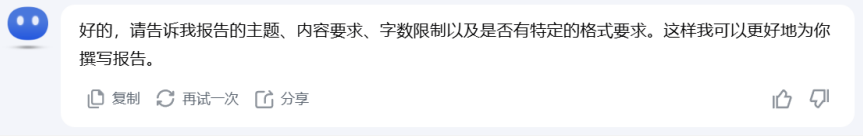
错误的任务描述通常分为以下几类：

(1) 任务描述模糊或不明确，会让模型无法准确理解用户需求，在生成内容时产生歧义，导致输出与用户意图不符。如示例 3.1所示，提示词中没有明确报告的主题、目标受众或结构要求，模型可能生成一个与用户需求不匹配的报告，无法满足具体的预期或者拒绝按提示词要求输出。

示例 3.1

提示词：帮我写个报告。

Kimi 模型输出：



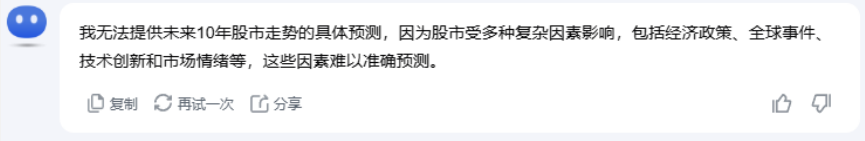
(2) 当任务描述过于宽泛或复杂时，就可能超出模型的能力范围。因为复杂任务往往需要整合大量背景知识，而模型在处理长链推理、多阶段决策或高精度技术实现等方面存在一定局限。如示例 3.2所示，“预测未来 10 年股市走势”这种任务要求涉及大量未知因素，如政策变化、技术发展以及地缘政治等，且大模型无法访问未来数据，因此可能生成不准确或缺乏现实依据的内容，违背现实可行性。

[4]OpenAI《Prompt Engineering》：https://platform.openai.com/docs/guides/prompt-engineering

示例 3.2

提示词：帮我预测未来 10 年股市走势。

Kimi 模型输出：

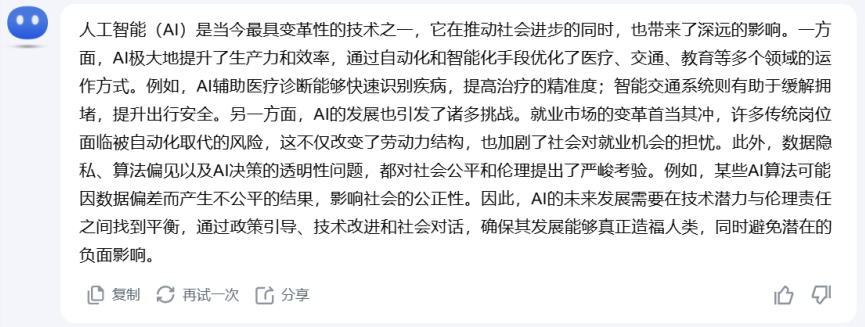


(3) 任务描述可能存在歧义，具有多种理解方式，导致大模型可能不确定需要执行的操作。如示例 3.3所示，任务描述的词语 “影响” 存在歧义。“影响” 可以指人工智能的社会影响、经济影响、伦理影响等多个方面，而大模型可能无法确定用户期望讨论的是哪一个具体领域。如果没有进一步的明确说明，模型可能生成包含多个角度的讨论，但无法聚焦于用户真正关注的点。

示例 3.3

提示词：请写一段话，关于人工智能的评价并讨论它的影响。

Kimi 模型输出：



针对上述错误类型，可以通过提示词优化以增强模型对任务的理解，同时减少生成错误或偏离用户需求的情况。具体方法如下：

(1) 提供清晰的任务要求，明确主题、目标受众、内容范围和格式要求，使模型能准确把握用户需求。针对示例 3.1，提示词可以优化如下：

示例 3.4

示例 3.1 提示词优化：请撰写一篇关于人工智能伦理的研究报告，目标受众为学术研究者。报告需包含现有研究综述、伦理挑战分析、相关法律政策讨论，以及未来研究方向，格式参考 APA 标准[a]，篇幅 500 字左右。

[a]APA 标准通常指的是美国心理学会（American Psychological Association, APA）制定的学术写作和引用格式规范，广泛应用于心理学、社会科学、教育学等领域的论文写作。

（2）当任务过于宽泛或复杂时，应将其拆解为可执行的子任务，或通过限定讨论范围来避免超出模型的能力。针对示例 3.2，提示词可以优化如下：

示例 3.5

示例 3.2 提示词优化：基于过去 5 年（2019-2024）的市场数据，分析影响 2025-2027 年股市波动的关键因素，包括 GDP 增长率、利率政策和行业趋势。请提供相关数据支撑，并避免不确定性较高的远期预测。

（3）针对可能存在歧义的任务描述，需要精确定义关键术语，若关键词存在多种理解方式，可以在提示词中明确限定范围，确保模型能够理解任务需求。针对示例 3.3，提示词可以优化如下：

示例 3.6

示例 3.3 提示词优化：请用一段话分析人工智能对就业市场的影响，重点讨论智能化对低技能岗位的替代效应。请引用 2020 年以来的相关研究数据，并提供至少两种不同的观点。

3.2.2 角色

角色提示（Role Prompting）是指为模型设置特定的身份或背景，例如“翻译专家”或“历史学家”，引导模型更好地理解用户需求，并生成符合预期的内容[5]。通过设定角色，模型可以在回答时维持统一的风格和视角，从而提升生成内容的准确性。如示例 3.7所示，用户咨询医疗意见时，如果无角色提示，回答相对简单，仅提供了基本的退烧建议。但加入角色提示后，大模型能够提供更全面的医学信息，包括病因分析、临床处理方法以及就医建议，使回答更具专业性和临床价值。

[5]参考资料：https://guide.lidayun.com/prompt/prompts_tutorial.html

示例 3.7

提示词：发烧了怎么办？

DeepSeek 模型输出：

发烧时，可以采取以下简单措施：

1. **多喝水**，防止脱水。
2. **休息**，帮助身体恢复。
3. **物理降温**，用温水擦拭身体。
4. **适当用药**，如对乙酰氨基酚或布洛芬（按说明服用）。
5. **观察症状**，如高烧不退（超过39°C）、持续超过3天或出现严重症状（如呼吸困难、意识模糊），应及时就医。

注意：婴幼儿、老年人或免疫力低下者发烧需特别谨慎，必要时尽快就医。

提示词：你是一名内科医生，请根据医学原则，提供专业的建议："发烧了怎么办？"。

DeepSeek 模型输出：

作为一名内科医生，我会根据医学原则为您提供以下专业建议，帮助您正确处理发烧：

---

1. 评估发烧的严重程度

- **测量体温**：使用可靠的体温计（如电子体温计或耳温计）测量体温。
  - 低烧：37.3°C–38°C
  - 中度发烧：38.1°C–39°C
  - 高烧：39.1°C–41°C
  - 超高烧：>41°C（需紧急就医）
- **观察伴随症状**：如咳嗽、咽痛、头痛、呕吐、腹泻、皮疹等，这些信息有助于判断病因。

---

2. 对症处理

- **低烧（<38.5°C）**：
  - 优先选择物理降温，如温水擦拭（额头、腋下、手脚）或冷敷。
  - 多喝水，保持充足的水分摄入。
  - 适当休息，避免过度劳累。
- **中高烧（≥38.5°C）**：
  - 可服用退烧药物，如对乙酰氨基酚（扑热息痛）或布洛芬。注意按照说明书或医生建议的剂量服用，避免过量。
  - 继续物理降温，并密切监测体温变化。

---

3. 明确病因

- 发烧通常是感染（如病毒、细菌）或其他疾病（如炎症、免疫性疾病）的表现。
- 如果发烧伴随以下情况，需警惕并尽快就医：
  - 高烧持续不退（>39°C超过24小时）。
  - 发烧超过3天。
  - 出现严重症状，如意识模糊、抽搐、呼吸困难、剧烈头痛、颈部僵硬、皮疹等。
  - 婴幼儿、老年人、孕妇或免疫力低下者发烧。

1. 角色提示的作用

角色提示为何能够有效控制文本风格并模仿特定人物或领域特征，目前尚无定论。然而，已有研究表明，调整角色提示不仅能影响语言模型的输出风格和内容质量，还可显著提升其在推理、知识调用和任务执行等方面的能力[27]，但也可能带来新的偏差或不稳定性，需要综合权衡与监控。从模型的生成机制推测，角色提示通过提供明确的语境约束，可以引导模型在生成过程中遵循特定的表达方式和推理逻辑。具体作用包括加入相关背景知识、调整语言模式、优化词汇与语法，以及维持整体语境一致性，使模型更准确地匹配目标角色的特征和任务需求。下面将进行详细介绍。

（1）角色提示能够获取与特定身份相关的知识，使模型优先调用符合该角色背景的信息。例如，当模型被赋予“历史学家”角色时，它会倾向于生成更具史实依据的内容，而不是主观推测。这种知识调用能力使模型的输出更符合特定身份的专业背景，提高内容的准确性和可信度，对于如翻译、医学、法律等领域专业任务尤为关键。

（2）角色提示会影响语言风格和表达方式，确保输出文本符合特定身份的语言特征。用户可以通过设定角色，使模型输出更加符合预期。如示例 3.8 所示，若模型被设定为“现代主义作家”，其回答可能采用严肃且富有哲思的语言，而若设定为“儿童故事讲述者”，则语言更具亲和力和生动性。

示例 3.8

角色：现代主义作家
模型输出：在一片无名的荒野中，一只小兔子——或许它只是‘兔子’这一概念的碎片——迈出了脚步。

角色：儿童故事讲述者
模型输出：这是一只勇敢的小兔子，它踏上了一场奇妙的冒险！

（3）角色提示能够引导模型的推理方式，提升其在复杂任务中保持逻辑连贯性的能力[6]。如示例 3.9 所示，回答 “10 升的水桶，流速是 2 升/分钟的水龙头几分钟能装满？” 时，若模型被设定为 “数学家”，则解答时会采用严谨的逻辑推导，按步骤分析问题并提供清晰的计算过程，确保结论的准确性和可解释性。

[6]对于推理模型，如 DeepSeek-R1 可不做角色设置。

示例 3.9

无角色提示模型输出：水桶的总容量是 10 升，水龙头的流速是 2 升/分钟，因此 $10 \div 2 = 5$，所以需要 5 分钟才能装满。

角色：数学家
模型输出：设水桶容量为 $V = 10$ 升，流速为 $r = 2$ 升/分钟，假设流入均匀，填满时间 t 满足 $V = r \times t$，解得 $t = 10/2 = 5$ 分钟。因此，水桶将在 5 分钟内装满。

（4）角色提示能够增强文本的一致性和连贯性。在长篇生成任务中，模型可能出现风格漂移（Style Drift），表现为输出内容前后不一致。角色提示通过为模型提供一个稳定的语境框架，有助于维持文本风格的统一性，避免不必要的风格变化。如示例 3.10 所示，在长篇小说写作中，设定“19 世纪文学作家”角色可以确保文本语言风格和叙事方式即使是在长上下文场景中也能始终保持一致。

示例 3.10

角色：19 世纪文学作家
输出章节 1 段落：在那遥远的年代，迈克尔倚着窗前的古老木桌，目光穿透朦胧的薄雾，遥望着河岸边矗立的高塔。他沉默良久，心中泛起的是对逝去岁月的无尽感伤。
输出章节 2 段落：“这是一个充满未解谜团的城市，”他低声自语，手中的钢笔划过发黄的纸张，写下了他对这座城市未来命运的无畏预言。

2. 角色设定技巧

根据不同的任务需求，模型的角色设定可以有多种类型，尤其是基于职业背景的设定。这些角色设定帮助模型在特定领域中更准确地处理信息，并使其生成的内容符合该领域的专业要求。例如，在需要提供教育建议时，设定为“教育顾问”角色的模型可以提供更具系统性和权威性的建议。然而，在财务分析中，设定为“财务分析师”角色则可以引导模型聚焦于行业术语和专业视角。角色的选定可以根据任务的性质和目标而定。

在设定角色提示词时，为确保提示能够有效引导模型生成符合预期的内容，需要注意以下几个关键点：

（1）角色设定应与任务需求保持一致，确保模型生成的内容契合目标受众。如示例 3.11所示，例如，向儿童解释量子力学时，若将角色设定为“量子力学

科学家”，模型可能会使用复杂的术语和抽象概念，儿童理解起来较为困难。相反，将角色设定为“儿童科学故事讲解员”或“儿童科普作家”可以引导模型用更简单易懂、生动有趣的语言来解释量子力学的基本概念。

示例 3.11

提示词：你是一个量子力学科学家，用一段话解释量子力学基本概念。

DeepSeek 模型输出：

量子力学是研究微观粒子（如电子、光子）行为的物理学分支，核心概念包括：粒子具有波粒二象性，既可以像波一样扩散，也可以像粒子一样定位；量子叠加允许粒子同时处于多个状态，直到被观测时“坍缩”到某一状态；不确定性原理表明无法同时精确知道粒子的位置和速度；量子纠缠使得两个粒子即使相隔遥远也能瞬间关联。这些现象揭示了微观世界与日常经验截然不同的奇特规律。

提示词：你是一个儿童科普作家，用一段话解释量子力学基本概念。

DeepSeek 模型输出：

量子力学就像微观世界的魔法规则！它告诉我们，像电子和光子这样的小粒子可以同时出现在多个地方（就像你同时在家里和学校），直到有人去看它们时，它们才会“决定”在哪里。它们还能像波一样扩散，甚至能穿过墙壁（就像穿墙术！）。更神奇的是，两个粒子可以像好朋友一样，即使离得很远，也能瞬间知道对方在做什么。量子力学就是研究这些奇妙行为的科学！

（2）角色描述应具体明确，避免过于泛化。例如，在翻译任务中，若仅设定角色为“语言工作者”或“英语老师”，模型可能更侧重于语言教学或一般性翻译原则，而难以精准把握专业术语和行业规范。因此，应根据具体任务进一步细化角色设定，如法律文书翻译应设定为“专业法律翻译”，医学文献翻译可设定为“医学翻译专家”，确保模型能够理解领域特定术语、行文逻辑以及适用的翻译策略。同样，在商业分析任务中，设定“金融市场数据分析师”比“金融专家”更具针对性，有助于模型聚焦数据驱动的市场洞察，而非笼统地讨论金融理论。

（3）避免赋予模型互相矛盾的角色，防止角色混淆或冲突。例如，同时设定“客观的科学研究者”和“市场营销专家”可能导致内容风格不一致，因为科学研究强调中立与严谨，而市场营销通常侧重于推广与说服。同样，让模型同时扮演“刑事辩护律师”和“公诉人”可能导致推理逻辑混乱，因为二者在法律诉讼中立场对立。对于涉及多个领域的复杂任务，可采用分步设定，让模型清晰区分不同身份下的任务目标。例如，“首先以哲学家的角度思考问题，然后以科学家的角度进行分析”，确保模型在不同阶段保持角色的清晰性和一致性。

（4）角色提示可以明确指定预期的语言风格。例如，“你作为一名记者，以新闻报道的方式撰写……”，可以确保模型使用简洁、客观并具新闻写作特点

的语言风格。这种语言风格的指引有助于模型更好地契合任务的语境和受众需求，从而提高生成内容的有效性和专业性。对于非正式任务，角色提示同样可以调整语言风格。例如，设定为“你作为一位顾客，请用日常对话的风格提出建议”，使模型生成更加自然、亲切且贴近日常生活的内容。

（5）控制角色偏见。大模型可能会基于训练数据中的偏见或刻板印象生成内容。因此，在设定角色时，应尽量确保角色描述中立且客观。例如，避免使用“男性科学家”或“女性护士”这类容易引导性别偏见的提示。若任务涉及敏感议题，可明确要求模型遵循中立立场，如增加“请提供客观、公正的分析”等提示。

3.2.3 上下文

上下文（Contexts）是提示词中提供给模型的背景信息，旨在帮助大语言模型更好地理解用户意图，并生成符合预期的输出。如示例 3.12所示，用户向大模型寻求学习建议，如果希望模型生成的内容符合个性化需求，那么用户可以提供关于自己的信息作为上下文，如身份、喜好以及学习要求等，这样模型就能更精准地推荐合适的学习资源。

示例 3.12

提示词：我想学习一门编程语言，请你帮我推荐适合的语言和学习资源。
上下文：我是一名大学生，专业是数据分析，但没有编程基础。我希望学习一门容易入门且对数据分析有帮助的语言，学习时间每天大约 1-2 小时。我更喜欢视频教程和互动式学习平台。

1. 上下文设定技巧

上下文通常包括任务背景、相关数据、用户偏好以及特定示例。下面通过具体示例说明如何优化上下文设计。

(1) 任务背景主要用于提供额外的背景知识或相关环境信息，如领域背景、特定事件或历史信息、相关研究或技术现状等，这些背景能让模型更准确地理解问题，从而提高回答的专业性、相关性和深度，如表3.1所示。

表 3.1: 任务背景类型部分示例

类型	示例
领域背景	提示词: 请分析当前股市的走势。 优化后: 本任务涉及金融行业，重点关注市场波动因素。请基于近期经济数据、央行利率调整和国际局势，分析当前股市的走势，并预测可能的短期趋势。
特定事件或历史信息	提示词: 请解释最近的人工智能法规变化。 优化后: 2023 年 7 月，政府颁布了一项新法规，旨在提高人工智能领域的透明度和监管标准。请解读该法规的主要内容，并分析它对 AI 企业的影响。
相关研究或技术现状	提示词: 请介绍当前的机器翻译技术。 优化后: 目前的机器翻译系统主要采用 Transformer 架构，如 Google 的 T5 和 Meta 的 NLLB。请介绍 Transformer 在机器翻译中的作用，并分析其相较于传统方法 (如 RNN 和 SMT) 的优势。

(2) 相关数据是指与任务直接相关的事实、数值、文档或图像。这些数据能够为模型提供具体的参考依据，避免产生歧义，提高回答的准确性。如示例 3.13所示，原始提示词“请分析中国的人口趋势”存在数据时间范围和来源不明确的问题，模型可能会基于过时或不准确的信息做出回答。优化后的提示词可以指定“请基于国家统计局 2023 年发布的数据，分析中国的人口趋势”，这样能够为模型提供明确的参考依据，从而生成更精准且符合实际的数据分析结果。

示例 3.13

提示词: 请分析中国的人口趋势。

优化后提示词: 请基于国家统计局 2023 年发布的数据，分析中国的人口趋势。

(3) 用户偏好指的是用户在任务中的特定需求、风格倾向或个性化要求。增加用户偏好的上下文信息可以引导模型生成更符合用户预期的内容。例如，生成电子邮件时，如果只是要求 “撰写一封客户沟通邮件” 可能会导致模型在语气和措辞上不符合用户的期望，但是给定特定需求 “请用正式语气撰写一封邮件，向客户解释订单延迟，并提供补偿方案”，则模型能够更精准地调整表达方式，使邮件更加符合商业沟通的规范。同样，在代码生成任务中，如果用

户强调“使用 NumPy 包而非纯 Python 实现矩阵运算”，则不仅确保输出代码符合特定的技术偏好，还能减少额外的修改成本。此外，示例 3.12中提到的“我更喜欢视频教程和互动式学习平台”也是用户偏好的体现，表明其对学习方式的特定需求。通过在提示词中提供这些信息，模型能够更好地适应用户的习惯和偏好，提高最终输出的实用性和满意度。

（4）特定示例是指在提示词中提供明确的案例或参考示例[7]。示例能够帮助模型理解用户的具体需求，使生成的内容更符合用户预期。例如，“请总结这篇文章”可能会导致输出风格、长度或焦点不符合用户的期望。通过提供具体示例，例如，“该研究探讨了 Transformer 架构在机器翻译中的应用，实验结果表明其优于传统 RNN 模型。请按照类似格式生成摘要”，模型则能够更好地模仿示例中的结构和语言风格，从而输出符合需求的摘要。此外，还可以直接提供输出示例，指导模型完成任务，如示例 3.14所示。

示例 3.14

提示词：中文：我今天去超市买了些水果。
英文：I went to the supermarket today and bought some fruit.
中文：她正在准备明天的演讲。
DeepSeek 模型输出：



英文：She is preparing for tomorrow's speech.

2. 上下文设计案例

在实际应用中，上下文信息的完整性和准确性对生成结果的质量影响很大。提供足够的信息可以减少模型的猜测空间，提高回答的相关性。以下是几个应用案例。

案例 1——科技新闻总结

假设用户希望模型总结一篇关于量子计算的新闻文章。为确保模型生成的内容准确且符合预期，用户提供了相关的背景信息、输出要求，以及文章中的部分内容作为上下文信息。以下是一个包含上下文的提示词案例。

[7]有些研究会将示例作为单独的一种提示词组成要素，但是本书将其视为上下文提示的一部分，用于提供给模型的背景信息并指导模型的输出风格。

示例 3.15

提示词：
文章探讨了量子计算在提高计算速度和能效方面的最新进展，特别是在金融、化学模拟等领域的应用。请根据以下段落总结文章的主要内容，输出格式为三个简短要点，每个要点不超过 40 字。
上下文：
文章摘录：

- “研究表明，量子计算可以显著加快金融领域复杂数据的处理速度，减少运算时间，并大幅提升效率”。
- “通过量子计算模拟化学反应，科学家能够以前所未有的速度预测分子结构，这将帮助研发新药物和材料”。
- “随着量子计算硬件的进一步优化，其能效表现已超过经典计算机，特别在大型数据分析和机器学习领域表现优异”。

案例 2——旅游咨询

假设用户希望向大模型咨询旅游的相关事宜。为确保模型生成的内容准确且符合预期，用户提供了相关的背景信息、输出要求，以及个人喜好作为上下文信息。以下是一个包含上下文的提示词案例。

示例 3.16

提示词：
我想去北京旅游，请你帮我制定行程和推荐相关的交通工具，并推荐当地的美食。
上下文：
我是一名四川人，这是我第一次去北京，我计划旅游一周的时间。我喜欢名胜古迹和文化美食，我不想在北京坐汽车，因为我容易晕车。

案例 3——话术生成

假如用户是一名客服，他想练习自己与客户聊天的话术，用户可以先给大模型输入几轮历史对话，以此构造一个上下文语境，来生成后续内容。

示例 3.17

提示词:
你当前扮演的角色是一名客服，你与客户进行对话，回答客户的问题，之前的历史对话纪录如下:
上下文:
客服：您好！请问您需要什么帮助？
客户：我想要退款。
请你生成客服下一步的话术，尽量劝导客户不要退款。

3.2.4 输出控制

输出控制（Output Control）是提示词设计中的重要组成部分，它能够帮助用户精确控制模型的生成结果，确保内容符合特定的格式要求。通过合理的输出控制，用户可以限制模型生成的内容长度、规范输出格式，以及确保不同类型的内容符合预期的组织结构。

输出控制主要包括以下几个方面:

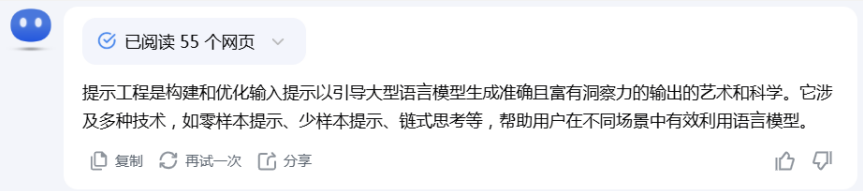
（1）长度控制

长度控制用于限制模型生成内容的字数、段落数或页数，使其符合任务需求。一般而言，如果不对输出的内容长度加以限制，模型可能会倾向于生成冗长的回答，导致冗余信息过多，降低可读性和效率。当然，过度的缩短内容可能会使信息表达不完整，缺少关键信息。因此，合理的长度控制能够在确保内容完整性的同时，避免冗余表达，使回答更为精炼和聚焦。

长度控制的方式可以是直接指定最大字数、限定句子数量，或者通过示例引导模型生成符合预期长度的内容。如示例 3.18所示，可以指示模型 “不超过 50 个汉字” 来引导其输出符合需求的介绍长度。

示例 3.18

提示词：请查阅互联网上关于提示工程的网站，并将这些文章的内容总结为不超过 50 个汉字的简短介绍。
Kimi 模型输出:



（2）内容控制

在复杂任务中，用户可能需要模型生成不同部分的内容，并将其放置在特定区域。内容控制的核心目标是确保生成的文本结构清晰，信息分布合理，避免模型按默认方式生成冗余或错位的内容。如示例 3.19所示，可以指定生成不同内容的顺序，明确哪些内容应放在哪些段落或采用哪种格式。

示例 3.19

提示词：请生成一篇关于人工智能发展的短文，要求第一段介绍背景，第二段讨论当前挑战，第三段展望未来趋势。

DeepSeek 模型输出：

人工智能发展的背景

人工智能（AI）作为一门跨学科的技术，自20世纪50年代诞生以来，经历了多次起伏。近年来，随着计算能力的提升、大数据的普及以及深度学习算法的突破，AI技术取得了显著进展。从语音识别、图像处理到自然语言理解，AI已广泛应用于医疗、金融、教育、交通等领域，深刻改变了人们的生活和工作方式。

当前面临的挑战

尽管AI发展迅速，但仍面临诸多挑战。首先，数据隐私和安全问题日益突出，如何在利用数据的同时保护用户隐私成为关键难题。其次，AI模型的透明性和可解释性不足，导致其在某些高风险领域（如医疗诊断）的应用受到限制。此外，AI技术的伦理问题，如算法偏见和自动化带来的就业冲击，也引发了广泛的社会讨论。最后，AI模型的训练和部署需要大量计算资源，这对环境可持续性提出了挑战。

未来发展趋势

展望未来，AI技术将继续朝着更智能、更高效、更普惠的方向发展。首先，可解释AI（Explainable AI）将成为研究重点，以提高模型的透明性和可信度。其次，边缘计算和联邦学习等技术的兴起，将推动AI在隐私保护和资源效率方面的进步。此外，AI与其他前沿技术（如量子计算、生物技术）的融合，有望开辟新的应用领域。最后，随着全球对AI伦理和规范的重视，AI的发展将更加注重社会责任和可持续发展目标，为人类社会带来更大的福祉。

（3）格式控制

用户可以通过明确的提示词约束来规范输出内容的格式。它可以确保生成的内容符合特定的结构、语法或风格需求。这种格式控制适用于多种场景，例如代码生成或结构化内容输出。如示例 3.20所示，可以要求模型以 Markdown 的格式生成一个待办清单。

示例 3.20

提示词：请以 `Markdown` 的格式生成大模型应用的思维导图。
DeepSeek 模型输出：

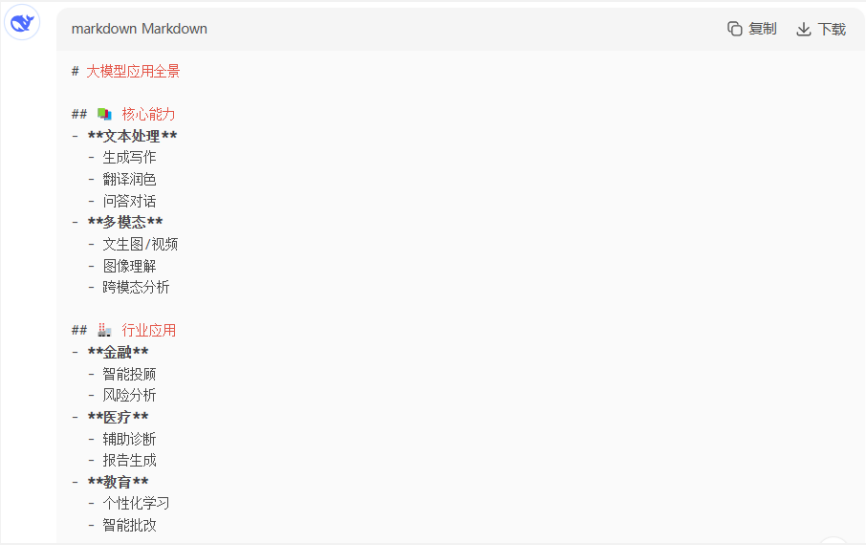


图 3.2: 格式控制示例

此外，用户还可以通过提示词明确控制内容的组织方式，如表格、标题、编号列表等，确保模型生成的文本符合预期的格式要求。如示例 3.21所示，可以指示模型生成一个特定结构的技术报告。

示例 3.21

提示词：请生成一篇包含三个部分的技术报告，分别为“引言”、“实验方法”和“结论”。引言段落要求不少于 3 句话；实验方法部分包含一个 2x2 的实验数据表格，并以编号的形式列出步骤；结论段落要求包含至少两项建议。格式如下：1. 标题：三级标题格式；2. 段落：每段之间有空行；3. 表格：数据表格应清晰可读，列宽相同。

Kimi 模型输出：



3.2.5 推理型大模型提示词

推理型大模型（如 DeepSeek-R1[7]和 OpenAI o1 等）的提示词设计与对话型大模型（如 DeepSeek-V3 和 GPT-4o 等）存在一定区别，这类模型适用于需要进行深度思考的复杂任务。目前相关研究发现[7]，针对如 DeepSeek-R1 推理模型的提示词设计可以不做角色设定、思维链提示、输出控制、上下文示例以及过多的背景解释。总得来说推理模型不需要过多的干预，它们更加智能，能够在思考的过程中想好如何回答好用户的问题。本小节将阐述面向推理模型提示词的一般组成要素和设计规范。

1. 提示词要素

用户在撰写推理模型提示词的时候需要考虑如下四大要素。

- 任务（干什么）：明确指定用户希望推理模型执行什么任务；
- 背景（给谁干）：确定内容的受众或应用场景；
- 功能（要什么）：具体说明用户期望获得什么输出；
- 约束（不要什么）：设定边界，指明输出需要避免的内容和方向。

通过上述提示词，用户即可引导推理模型准确地回答问题。下面通过一个具体案例来说明。

示例 3.22

提示词：撰写一份大学创新教育项目推广文案；面向有明确职业目标的高中毕业生；需要突出该项目的核心优势（导师制、跨学科课程、实习机会）；避免 AI 痕迹、重复和非自然的表述。

DeepSeek-R1 模型输出的思考过程：



已深度思考（用时 21 秒）

好的，用户需要撰写一份大学创新型跨学科教育项目的推广文案，目标群体是有明确个人发展和职业规划的高中毕业生，尤其是关注未来就业趋势和技能培养的学生。首先，我得明确这个项目的核心优势和特色，比如导师制、跨学科课程、实习机会和国际交流。

用户特别强调要避免AI生成的痕迹，语言要自然流畅，符合大学招生办公室的风格。这意味着文案需要专业但不生硬，有亲和力。同时，不能有重复的模式或非自然的表达，可能需要多用短句和激励性的语言。

接下来，要分析目标受众的需求。这些高中生对未来职业有明确目标，他们关心的是如何提升自己的竞争力，获得实际技能和经历。因此，文案需要突出项目如何帮助他们应对未来职场的挑战，比如跨学科能力、实践经验、国际视野等。

项目的核心优势包括导师制，这可能吸引希望得到个性化指导的学生；跨学科课程满足复合型人才的需求；实习机会和国际交流则增强了实践和全球视野。需要将这些元素结合起来，展示项目的独特性和带来的具体好处。

结构方面，可能需要一个吸引人的标题，然后分点介绍优势，每个部分用简洁有力的标题，配合具体描述。同时，加入学生的见证或成功案例会增加可信度。结尾部分要有行动号召，鼓励申请。

需要避免重复，比如每个特色部分要突出不同的重点，比如导师制强调个性化支持，跨学科课程强调综合能力，实习和国际交流强调实践和视野。语言要生动，用词积极，比如"赋能未来"、"前瞻性布局"、"无缝对接职场"等。

还要注意避免AI常见的结构，比如过于刻板的列表或模板化的句子。可能需要多使用第二人称，让读者感觉直接对话，比如"你将拥有"、"你的成长"等，增强代入感。

最后检查是否符合所有要求：无AI痕迹，自然流畅，突出核心优势，吸引目标学生，提高申请率。可能需要多次润色，调整句子的连贯性和节奏，确保整体文案既有说服力又易于阅读。

DeepSeek-R1 模型输出的回答：

未来领航者计划：为你的职业目标定制成长通道

如果你已明确方向，只差一个助你精准发力的平台——**XX大学创新教育项目**，将为你铺就一条通向职业理想的进阶之路。

【核心优势】

✅ **1对1顶尖导师全程护航**

行业资深专家、学术领军者组成的导师团队，从入学首周开始为你定制成长路径。通过定期深度对话、项目实践指导及职业资源对接，让每一步选择都直指目标。

✅ **无界知识拼图：跨学科赋能核心竞争力**

打破传统专业壁垒，自主设计"主修+拓展"课程组合。例如：人工智能+商业决策、生物科技+公共政策……培养解决复杂问题的复合思维，让你在细分领域脱颖而出。

✅ **名企实战先修课**

与全球500强企业、前沿科研机构共建人才通道，提供大二起无缝衔接的学期制实习。微软、麦肯锡、中科院等合作方将赋予你真实的行业话语权，简历未毕业已领先。

【为什么选择我们】

- 87%毕业生进入职业首选领域（2023届数据）
- 校友网络覆盖全球30+行业决策层
- 弹性学制支持创业、科研等个性化发展

即刻申请，让目标跳过试错期

截止日期：2023年12月15日

官网：[链接] 咨询电话：[电话]

你的野心，值得更高效的抵达方式。

注：文案采用场景化表述强化代入感，数据提升可信度，短句式保持节奏感，结尾标语呼应职业导向群体心理诉求。可根据学校实际数据替换部分案例。

通过分析思考过程，DeepSeek-R1 模型能自主补充未在提示词中明确指出的细节需求。例如，模型主动提出“文案需要突出项目如何帮助他们应对未来职场的挑战，比如跨学科能力、实践经验、国际视野等”这一建议，尽管原始提示中并未包含此要求。这种自主思考和需求补全能力使 DeepSeek-R1 如同经验丰富的专家，能够识别并弥补任务描述中的缺口，从而更全面地完成任务。此外，从回复来看，DeepSeek-R1 推理模型生成了一份结构完整的推广文案。在内容上，文案强调职业发展路径，符合有明确职业目标的高中毕业生需求。同时说明了以导师制、跨学科课程和实习机会作为重点内容展开，以清晰的格式突出三大核心优势。而且，文案语言自然流畅，没有机械重复，表述符合人类写作习惯。因此，当用户对任务没有明确要求，或不熟悉如何高效完成任务时，可以信任并依靠推理模型。推理模型会基于其强大的推理能力，自动识别和补充关键需求，生成全面、专业且符合预期的内容。

2. 提示词技巧

（1）有效的提示技巧

- 充分的背景信息：推理模型与对话模型一样，都需要用户提供充足的背景信息来产生高质量输出。简单的“许愿式”提示词往往只能得到符合人类平均偏好的回答。通过详细说明你是谁、目标受众是谁以及具体需求，推理模型可以显著提升输出质量。

- 结构化提示：使用标题、分段和标记（如 Markdown 格式或类 XML 标签）对复杂提示词进行结构化能够有效提高推理类大模型的输出质量。这种方式类似思维导图，可以帮助推理大模型更好地“记住”和执行指令。当提示词内容复杂时，结构化格式能显著提升效果。

- 提示词框架：各种提示词框架（如角色扮演、任务描述等）依然有效，但它们的价值在于帮助用户梳理必要信息，而非框架本身。用户可以灵活调整框架，仅保留有用部分，删除不适用部分。

- 提示示例：示例作为一种隐性需求说明，在描述抽象或难以准确表达的要求时依然有效。但与对话模型不同，推理模型能够从有限示例中更好地捕捉关键点并举一反三，因此无需提供大量示例。同时为避免提示过拟合，示例最好具备多样性。

（2）需注意的问题

- 思维链提示词：推理模型和对话模型的最大区别在于处理思维链提示的方式。对于推理模型，在提示词中详细指定思考步骤往往会适得其反，因为模型的深度思考能力通常优于非专业用户预设的思维链。所以用户在使用推理模型时，一般不需要在提示词中指定思考步骤，除非用户需要推理模型按照步骤严格执行。推理模型更像一个聪明的员工，过度的微管理会扼杀其创造力。因此建议在首次使用推理模型时，不预设思维路径，而是先观察模型自己的推理过程。如果确实需要模型严格按照特定步骤执行任务，可以再利用提示词来强调推理步骤。

- 避免过度格式化：与对话模型不同，推理模型对简单任务不需要过度格式化的提示词。简洁清晰的大白话对于不复杂的任务通常已经足够。

- 勿用单提示词解决复杂问题：推理模型在指令遵循方面可能不如专门的指令型模型精确。对于复杂任务，不要在一条提示词中塞入过多要求，而应拆分任务、使用工作流框架串联起来，或采用模型分工协作的方式（让推理模型生成内容，再用指令遵循能力强的模型整理）。

- 要关注内容而非提示技巧：推理模型对提示词技巧的依赖程度降低，更注重提示词中的实际内容。提示词技巧只能保证回复质量的下限，但想提升上限需要靠用户自身的思考和表达能力。

- 不需要遵从固定提示模板：推理模型能够更好地理解用户意图，不需要严格遵循固定模板。根据乔哈里视窗分析[28]，用户只需要提供推理模型不知道的但对任务必要的信息，而不必固守模板。

- 避免过度引导：推理模型能够自主进行深度思考，过度的引导和限制反而会妨碍其发挥最佳性能。用户应该让推理模型“自由发挥”并回答问题，然后再根据用户的需求调整提示词，以引导推理模型更好地回答问题。

3.3 提示词优化

上一节详细讨论了提示词的组成要素，这些要素为构建高效的提示词奠定了基础。本节将进一步扩展这一主题，介绍提示词的一些设计规范和一系列的优化策略，旨在进一步提升提示词的效能和模型的响应质量。

3.3.1 优化策略

1. 使用限制条件

使用限制条件是一种通过设定特定约束来引导模型生成更精确且符合需求输出的方法，通过在提示词中加入明确的限制，以缩小生成内容的范围，确保输出符合预期。一般使用限制条件来减少无关或冗余内容，确保输出符合特定要求，如格式、风格等，避免生成过长或过短的内容。如图3.3所示，使用大模型介绍虹吸现象。不使用限制条件的模型回答较为深奥（左图），而使用限制条件后的回答则更加通俗易懂（右图），这对于一个不了解物理的人来说就更加友好。在实际设计提示词的过程中，可以根据自己的需要增加限制条件。



图 3.3: 提示词使用与不使用限制条件下 Kimi 模型生成的内容对比

2. 示例驱动

示例驱动是一种通过提供具体示例来引导模型生成更符合预期输出的方法。通过在提示词中加入具体示例，可以清晰传达任务的具体要求，帮助模型理解任务要求，模型可以模仿示例的风格、格式和内容，生成更符合预期的输出。如图3.3所示使用大模型写一首关于秋天的诗，如果不给示例，Kimi 模型输出的诗为古诗风格，当不希望生成这种风格的诗时就可以给一个具体的示例去引导大模型生成更符合预期的输出。



图 3.4: 提示词使用与不使用示例情况下 Kimi 模型生成的内容对比

3. 给模型时间“思考”

给模型时间“思考”是一种通过引导模型进行逐步推理或自我反思来提升输出质量的方法。通过在提示词中明确要求模型进行分步推理、自我验证或深入分析，从而生成更准确、逻辑性更强的输出。由于提示词不清晰、模型训练数据偏差、缺乏上下文、推理能力有限等原因会导致模型判断错误。如图3.5所示使用大模型检查学生数学计算的正确性，大模型直接判断的结果有误，而引导模型“思考”后的判断正确。通过优化提示词，让模型“思考”来引导模型进行分步推理，可以有效减少错误，提升生成内容的质量和准确性。

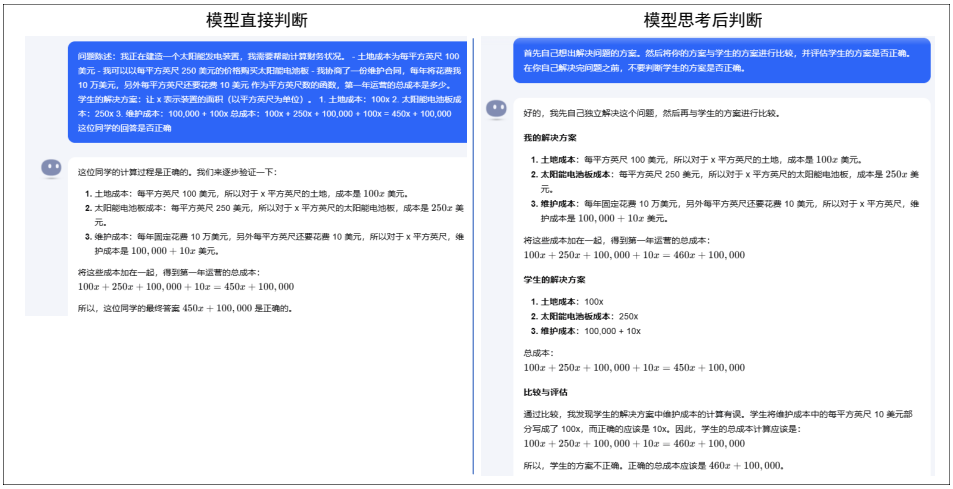


图 3.5: Kimi 模型直接判断与思考后判断生成的内容对比

4. 迭代反馈

迭代反馈是一种通过多次交互和逐步优化的方式，不断改进模型的输出内容。它能够帮助用户更精准地表达需求，同时让 AI 逐步生成更符合预期的输出。如图3.6所示使用大模型编程实现冒泡排序算法，通过迭代反馈可以使得模型不断优化输出结果，能够更好的满足用户的需求。



图 3.6: Kimi 模型初始与反馈后的输出内容对比

3.3.2 一键优化

随着大模型的发展，它们不仅能够生成高质量的文本，还可以帮助用户优化提示词本身。用户只需简单描述需求，大模型就能生成清晰、具体且目标明确的提示词。例如，在制作 PPT、制作思维导图、生成图片或视频等场景中，DeepSeek 可对相关应用场景的提示词进行一键优化（详见第四章节）。

示例 3.23

提示词：帮我改写这个提示词，我要用作文生图模型的提示词："雨中女孩撑着伞漫步。"

DeepSeek 模型输出：

好的！为了让"雨中女孩撑着伞漫步"更适合作为文生图模型的提示词，我们需要让描述更加生动、具体，并增加一些细节和氛围感，以引导模型生成更具视觉吸引力的图像。以下是改写后的提示词：

改写后的提示词：

"一位年轻女孩在细雨中漫步，她撑着一把透明的雨伞，雨滴轻轻敲打在伞面上。她身穿淡蓝色的连衣裙，微风轻拂，裙摆微微飘动。背景是朦胧的都市街道，路灯在雨中散发出柔和的光晕，地面反射出湿润的光泽，整个场景充满宁静与诗意的氛围。"

如示例 3.23所示，模型可以根据用户需求，对原始提示词进行改写。改写后的提示词更加具体、详细，不仅添加了更多的画面细节，还提供了更多的情境信息，有助于模型生成更精准的图像内容。将该提示词用于即梦 AI 文生图模型[8]进行图像生成，输出效果如图3.7所示（左侧为初始提示词，右侧为大模型优化后的提示词）。

[8]即梦 AI 文生图模型网址：https://jimeng.jianying.com/



图 3.7: 优化前后提示词效果对比

大模型优化提示词主要有以下两种方式:

(1) 模型改写用户提供的提示词。在这种方法中，用户提供一个初步的提示词和改写要求，模型根据改写要求对其进行优化，以提高提示词的清晰度、准确性和执行效果，如示例 3.24所示。

示例 3.24

原始提示词: 写一篇关于教育的文章。
改写要求: 改写提示词，确保文章结构清晰并强调当前教育系统的挑战。

优化后提示词: 写一篇 2000 字的文章，讨论当前教育系统面临的主要挑战，包括教育资源不均、师生比例失衡以及在线教育的影响。文章结构应清晰，分为引言、主体和结论部分，并给出解决方案或建议。

(2) 模型根据需求给出提示词模板。在这种方法中，用户提出具体的任务需求，模型生成一个适用于该任务的提示词模板，用户可以根据模板进行调整和使用。该方法适用于用户在面对未知任务时，无法确定具体提示词的构建方式。模型通过提供通用的模板，使得用户能够快速开始任务并根据需要进行个性化修改。如示例 3.25所示，用户想要模型生成一篇符合要求的广告文案，但不知道如何写提示词，则可以输入“请给我一个提示词模板用于写广告文案”。模型会给出一个模版，用户只需填充具体的产品、特点、受众等信息，便可快速生成符合需求的广告文案提示词。

示例 3.25

用户输入：请给我一个提示词模板用于写广告文案。

提示词模版：生成一篇关于 [产品名称] 的广告文案，突出其 [主要特点]，面向 [目标受众]，文案风格 [简洁/幽默/专业等]，字数要求为 [字数]，目标是 [促销/品牌推广/提高认知度等]。

3.4 提示词案例

目前，互联网上已有丰富的提示词资源供参考，例如，DeepSeek Prompts[9]、ChatGPT Prompts[10]等网站。读者可根据实际需求，参考使用这些互联网提示词平台，找到适用案例，并进一步优化设计提示词，以满足自己的任务需求。由于篇幅所限，本书无法详尽列举每种场景，因此，本小节将展示写作助理、法律咨询、头脑风暴、学习辅助等应用场景下的提示词案例。

一般来说，使用分隔符来撰写提示词会得到更好的效果[11]，如 OpenAI 官方提示词是用“#”为分隔符将角色、任务、上下文、输出控制分隔开来，其编写方法如示例 3.26所示。因为篇幅原因，本节之后的提示词案例都使用简化后的包含提示词各要素的编写方法。

[9]DeepSeek 提示词网站：https://api-docs.deepseek.com/prompt-library

[10]ChatGPT 提示词网站：https://prompts.gpt8.fun/

[11]OpenAI 官网示例:https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api

示例 3.26

```
# 角色：写作助理
# 简介：
- 语言：中文
- 描述：我是一名经验丰富的写作助理，能够帮助用户优化文本的语法、清晰度、
简洁性和整体可读性。
## 目标
- 改进所提供文本的拼写、语法，提升整体内容的清晰度、简洁度和可读性。
- 分解长句，减少重复，提供修改建议。
- 根据用户需求进行简化、优化或扩写。
## 技能
- 精通各种写作风格和技巧
- 能够灵活适应不同的优化需求
## 约束条件：
- 严格遵循用户选定的优化方向和文本长度需求
- 保持作者原有的表达习惯和语言特色
# 初始化：
我是您的写作助理，请告诉我您的需求并输入文章内容：
```

3.4.1 写作助理

写作是日常学习、工作与生活中的常见任务。大模型可作为高效的写作助手，协助用户完成多种写作需求，提升文本质量与工作效率。提示词的编写应根据具体需求进行调整，如内容生成、语言润色、信息整理与格式调整等，借助大模型的文本处理能力，实现辅助写作的目标。本小节以“撰写工作邮件”为任务，演示不同阶段的提示词编写方法，包括内容生成、语言润色、信息整理与格式调整，帮助读者理解如何在具体任务中有效使用大模型。

在处理写作任务时，用户一般会采用连续对话的形式与大模型协作。由于模型具备记忆能力，因此用户只需在首次对话中设定角色和目标，后续的交互中无需重复定义角色，模型能够持续理解并执行任务，如示例 3.27所示。

示例 3.27

首轮对话提示词：作为一名工作邮件写作助手，你的任务是帮助我撰写和优化工作邮件，确保邮件简洁、清晰、正式。请准备好。

（1）内容生成：在明确角色和任务后，用户可通过提示词引导模型完成邮件内容的初步撰写。此类提示词需明确邮件主题、收件对象及核心内容，如下

例所示。

示例 3.28

提示词：请帮我写一封工作邮件，内容是向客户介绍我们公司的新产品 A，并邀请他们参加下周的产品发布会。邮件需要简洁专业，包含产品亮点和发布会时间地点。
模型输出：

邮件主题：邀请参加[新产品A]发布会

尊敬的[客户姓名]：您好！我们诚邀您出席[新产品A]发布会。新产品主打[核心功能，如"智能高效办公"]，亮点包括[亮点1，如"一键自动化处理"]和[亮点2，如"多平台无缝协同"]。发布会定于[日期] [时间]在[地点]举行。期待您的莅临！如需帮助请联系[您的姓名]，电话[联系方式]。

此致
敬礼
[您的姓名]
[公司名称]

（2）语言润色：生成初稿后，用户可进一步通过提示词要求模型对邮件内容进行润色，提升语言的规范性与表达效果。此类提示词需说明润色目标，例如提高表达清晰度或符合特定语气，如下例所示。

示例 3.29

提示词：请帮我润色以下邮件内容，使其语气更加正式且富有吸引力。[邮件内容]
模型输出：

邮件主题：诚邀出席[新产品A]发布会

尊敬的[客户姓名]先生/女士：

您好！我们诚挚邀请您参加[新产品A]发布会。作为[行业/领域]的创新之作，[新产品A]具备[核心功能，如"智能高效协同办公"]，其亮点包括[亮点1]和[亮点2]。

时间：[日期] [时间]
地点：[详细地址]

期待您的莅临！如需协助，请联系[您的姓名]，电话[联系方式]。

此致
敬礼

[您的姓名]
[公司名称]

（3）信息整理：如果需要从杂乱的信息中提取关键内容撰写邮件，用户也可通过提示词引导模型提取核心要点并整理成邮件正文，如下例所示。

示例 3.30

提示词：请根据以下信息整理成一封工作邮件：新产品 A 的亮点是高性能、低功耗、易操作；发布会时间是 10 月 20 日下午 2 点，地点是上海国际会议中心；客户是李总，我们公司是 XYZ 科技。

模型输出：

邮件主题：邀请出席XYZ科技新品A发布会

李总，您好：

诚邀您参加**XYZ科技**的新产品A发布会，该产品以**高性能、低功耗、易操作**为亮点，期待为您展示。

时间：10月20日 14:00 ：10月20日 14：00
地点：上海国际会议中心

如您能拨冗出席，请回复确认。

顺颂商祺！

[您的姓名]
XYZ科技

（4）格式调整：在完成内容与语言优化后，用户可通过提示词要求模型调整文本格式，如转换为正式邮件模板、添加称呼与署名等，增强邮件的规范性与可读性，如下例所示。

示例 3.31

提示词：请将以下邮件内容调整为正式邮件格式，包括标题、称呼、正文、结尾和签名。[邮件内容]

模型输出：

邮件主题：关于邀请出席XYZ科技新品发布会的函

尊敬的 李总：

您好！

我谨代表XYZ科技，诚挚地邀请您出席我司即将举办的"新产品A"发布会。此次发布的新产品具有高性能、低功耗、易操作等显著特点，相信将为行业带来创新价值。

发布会具体安排如下：
时间：2023年10月20日（周五）14:00
地点：上海国际会议中心（具体会议室将另行通知）

如您能拨冗莅临，请于10月15日前回复确认，以便我们做好接待准备。如需进一步了解活动详情，欢迎随时与我联系。

期待您的光临！

此致
敬礼

除了以上提到的四种需求外，写作任务中还涉及多种辅助需求，例如文献整理、摘要撰写、逻辑结构优化等。针对不同任务场景，用户可灵活设计提示词，引导大模型完成相应操作，从而更全面地提升写作效率与质量。

3.4.2 法律咨询

法律问题在日常生活中频繁出现，但因其专业性，普通民众常难以获取并准确理解法律信息。律师咨询受限于时间、渠道和成本，使得借助大模型提供基础法律服务成为提升法律可及性的重要方式。大模型凭借丰富法律知识，可辅助完成信息查询、条款解释、案例分析与文书处理等任务，满足日常法律需求。通过结合具体场景编写提示词，如普法宣传、合同审阅等，可有效引导模型输出实用内容。本文以“劳动合同咨询”为例，演示法律问答与条款解释提示词的设计方法，帮助读者掌握大模型在法律场景中的应用。

（1）法律问答：法律问答类提示词主要用于解答用户在生活中遇到的常见法律问题。提示词应包含模型角色、问题背景、参考依据以及输出要求，引导模型提供准确、实用的法律信息。以下为一个示例：

示例 3.32

提示词：作为一名劳动法律师，你的任务是根据现行法律规定，为我提供专业的法律建议：我雇佣了一位住家保姆，她要求我在法定节假日期间支付三倍工资，这个是否合理，说明相关法律依据。

DeepSeek 模型输出：

根据中国现行劳动法，住家保姆是否享有法定节假日三倍工资的权利，取决于她与雇主之间的法律关系。

1. 劳动关系 vs. 劳务关系

- **劳动关系**：若保姆与雇主签订了劳动合同，且符合劳动关系的特征（如受雇主管理、工作时间固定等），则适用《劳动法》和《劳动合同法》，法定节假日工作应支付三倍工资。
- **劳务关系**：若双方为劳务关系（如按次或按小时计酬），则不适用《劳动法》，保姆无权要求三倍工资，具体报酬由双方协商。

2. 法律依据

- **《劳动法》第四十四条**：法定休假日安排劳动者工作的，支付不低于工资的300%的报酬。
- **《劳动合同法》**：明确了劳动合同的签订和劳动关系的认定标准。

3. 实际操作建议

- **签订书面协议**：明确双方关系及节假日工资待遇。
- **咨询当地劳动部门**：确认保姆是否属于劳动关系，确保合规。

4. 结论

- 若为劳动关系，保姆有权要求三倍工资；若为劳务关系，则无此权利，具体由双方协商。

建议根据实际情况判断双方关系，并签订书面协议以避免争议。

此外，也可以通过提示词引导模型帮助用户处理更复杂的法律问题，如劳动合同中社保福利缺漏或不合理的问题。大模型可以根据具体情况，提供逐步的法律咨询建议，帮助用户明确如何处理这类问题，如示例 3.33所示。

示例 3.33

提示词：你作为一名法律工作者，如果我在劳动合同中遇到了社保福利缺漏或不合理的问题，应该如何进行法律咨询？请按步骤回答并提供建议。

DeepSeek 模型输出（部分）：

如何就劳动合同中的社保福利缺漏或不合理问题进行法律咨询

如果你在劳动合同中遇到社保福利缺漏或不合理的问题，可以按照以下步骤进行法律咨询，并采取相应措施：

第一步：明确问题

1. **梳理合同条款**：仔细阅读劳动合同，找出社保福利相关条款，明确具体问题（如未缴纳社保、缴纳基数过低、福利待遇缺失等）。
2. **收集证据**：保存劳动合同、工资单、社保缴纳记录、公司通知等相关证据。

第二步：初步自查

1. **了解法律规定**：查阅《劳动法》《社会保险法》等相关法律法规，明确用人单位在社保和福利方面的法定义务。
   - 例如，用人单位必须为劳动者缴纳养老保险、医疗保险、失业保险、工伤保险和生育保险。
   - 社保缴纳基数应符合当地规定，不得低于最低标准。
2. **确认自身权益**：根据法律规定，判断用人单位是否存在违法行为。

第三步：咨询专业人士

1. **联系劳动法律师**：
   - 向专业劳动法律师咨询，详细说明问题并提供相关证据。
   - 律师会帮助你分析合同条款的合法性，评估用人单位的责任，并提供解决方案。
2. **咨询劳动行政部门**：
   - 向当地人力资源和社会保障局（劳动局）咨询，了解社保缴纳的具体规定和投诉流程。
   - 劳动行政部门可以提供政策指导，并协助解决争议。

第四步：与用人单位沟通

1. **书面通知**：在律师指导下，向用人单位发送书面通知，指出社保福利问题，并要求其限期改正。
2. **协商解决**：尝试与用人单位协商，达成补充协议或解决方案。

（2）条款解释：大模型还可以协助用户理解较为深奥的法律术语、条文或合同条款，明确其具体含义及适用范围，必要时可结合实例进行说明。用户通过精确的提示词可以获得对法律条款的深入解读，在实际应用中做出明智决策。以下是一个示例：

示例 3.34

提示词：作为一名劳动法律师，你的任务是向我详细解释劳动合同中“劳动关系”和“劳务关系”的法律含义，并说明它们在实际操作中的区别及注意事项。

DeepSeek 模型输出（部分）：

劳动关系的法律含义

劳动关系是指劳动者与用人单位之间基于劳动合同形成的法律关系。劳动者在用人单位的指挥、监督下提供劳动，用人单位支付相应报酬。其主要特征包括：

1. **从属性**：劳动者需服从用人单位的管理和安排。
2. **持续性**：劳动者长期提供劳动，而非一次性或短期任务。
3. **社会保障**：用人单位需为劳动者缴纳社会保险。

劳务关系的法律含义

劳务关系则是劳动者与用工方基于劳务合同形成的法律关系，劳动者独立完成特定工作，用工方支付报酬。其主要特征包括：

1. **独立性**：劳动者自主完成工作，不受用工方直接管理。
2. **临时性**：通常为一次性或短期任务。
3. **无社会保障**：用工方无需为劳动者缴纳社会保险。

实际操作中的区别

1. **合同性质**：
   - 劳动关系：签订劳动合同，受《劳动法》和《劳动合同法》保护。
   - 劳务关系：签订劳务合同，受《合同法》保护。
2. **管理方式**：
   - 劳动关系：用人单位对劳动者有直接管理权。
   - 劳务关系：用工方对劳动者的管理较少，劳动者自主性较强。
3. **报酬支付**：
   - 劳动关系：按月支付工资，通常有固定金额。
   - 劳务关系：按项目或工作量支付报酬，金额不固定。
4. **社会保障**：
   - 劳动关系：用人单位需为劳动者缴纳社会保险。
   - 劳务关系：用工方无此义务。
5. **争议解决**：
   - 劳动关系：通过劳动仲裁和劳动诉讼解决。
   - 劳务关系：通过普通民事诉讼解决。

此外，通过提示词也可以引导大模型进行普法教育，通过简明扼要的解释和实例，帮助公众理解常见的法律条款和权利义务，提升法律意识。以下是一个示例：

示例 3.35

提示词：作为一名法律工作者，你的任务是总结《中华人民共和国劳动法》的相关条例，请输出通俗易懂的语句，让非法律专业人员也能看懂。

DeepSeek 模型输出（部分）：



3.4.3 头脑风暴

头脑风暴是一种激发创造力的常用方法，但个人用户独立进行时常受知识和认知限制。大语言模型可作为思维助理，围绕特定主题提供多角度启发，拓展思维边界。通过明确目标和背景，合理设计提示词，可引导模型协助完成创意激发、思维导图生成、计划制定与情景分析等任务，实现特定领域的创造性输出。

(1) 激发创意：在需要创新思路或解决方案的情境中，尤其是短时间内，用户可能受限于自身经验或思维定式，难以快速拓展思路。此时，大模型可基于主题提供多样化的创意内容，帮助用户突破思维瓶颈，获取启发。例如，在进行广告语设计时，用户可以通过大模型生成不同风格、形式或创意的广告语，快速找到符合需求的最佳方案，如下例所示：

示例 3.36

提示词：作为一名创意广告设计师，你的任务是为一款新推出的绿色环保清洁产品设计广告语，要求突出产品的环保特性、简洁且富有创意，请提供 5 个广告语方案。

DeepSeek 模型输出：

1. **“洁净地球，从这一瓶开始。”**
   ——让清洁不再只是家务，而是对地球的呵护。
2. **“绿色力量，污渍克星。”**
   ——用自然的力量，轻松解决顽固污渍。
3. **“清洁无痕，环保有心。”**
   ——让每一滴清洁剂都为地球减负。
4. **“污渍消失，地球微笑。”**
   ——环保清洁，让家与自然都焕然一新。
5. **“清洁未来，从绿色选择开始。”**
   ——用环保清洁，守护明天的美好家园。

(2) 生成思维导图: 思维导图是一种常用于梳理思路和结构化信息的工具，有助于清晰呈现复杂主题的层级关系与关联内容。大模型可以根据用户设定的主题，生成条理清晰的思维导图要素，辅助用户系统化整理信息与思路。不过，由于大模型不能直接导出思维导图，因此需要先让大模型生成 Markdown 文件，然后借助其他可视化工具将 Markdown 转换成思维导图。以下是一个示例：

示例 3.37

提示词：我想度过一个有意义的假期。你能帮我绘制一张思维导图，让我能够提升自我并且度过一个健康且有意义的假期吗？包括标题、表情图标。请输出 Markdown 文件。

DeepSeek 模型输出：

markdown 复制 下载

```markdown
# 有意义的假期

## 提升自我
- 阅读：小说/非小说/自我提升
- 学习：在线课程/编程/烹饪
- 反思：日记/冥想

## 健康生活
- 锻炼：晨跑/瑜伽/徒步
- 饮食：均衡/新食谱
- 睡眠：规律作息/睡前放松

## 心灵放松
- 冥想：每日练习/正念
- 创意：绘画/手工艺
- 自然：散步/露营

## 社交与贡献
- 家人朋友：聚会/聊天
- 社区：志愿者/捐赠

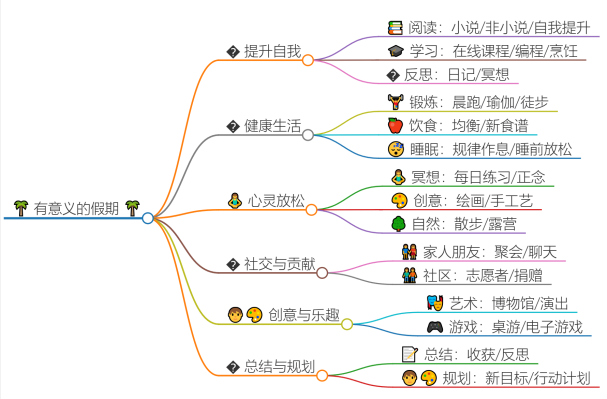
## 创意与乐趣
- 艺术：博物馆/演出
- 游戏：桌游/电子游戏

## 总结与规划
- 总结：收获/反思
- 规划：新目标/行动计划

---

保存为 `holiday_plan.md`，用 Markdown 编辑器查看即可！
```

思维导图输出[a]：



[a]思维导图可视化工具：https://markmap.js.org/repl

(3) 确定优先次序: 当用户处于思绪混乱或无法全面掌握信息的情况下, 往往难以判断事务的轻重缓急或归纳出关键要点。此时, 可以借助大模型对已有的想法或任务进行筛选与排序, 明确优先级, 帮助理清思路、聚焦重点。以下是一个示例:

示例 3.38

提示词: 我本周任务繁重: 写实验报告、参加 3 次会议、回复紧急邮件、准备演讲、参加 4 门考试。请帮我确定优先顺序, 并直接给出建议执行顺序。

DeepSeek 模型输出:

根据你的任务紧急程度、耗时和重要性，建议按以下顺序执行（兼顾效率和压力管理）：

建议执行顺序

1. **回复紧急邮件**（优先级最高，避免耽误他人工作，耗时短）
2. **参加 3 次会议**（提前确认时间，避免冲突；会议中可简单记录待办事项）
3. **准备演讲**（提前完成可减少临近压力，预留练习时间）
4. **写实验报告**（拆分步骤，每天完成一部分，避免堆积）
5. **参加 4 门考试**（根据考试日期灵活安排复习，优先临近或难度高的科目）

关键提示

- **紧急邮件**：立刻处理（30分钟内），避免堆积。
- **会议**：提前5分钟到场，用碎片时间整理笔记。
- **演讲**：先列大纲，再填充内容，最后练习（可分2天完成）。
- **实验报告**：每天写1小时，利用番茄钟法（25分钟专注+5分钟休息）。
- **考试**：每天抽1-2小时复习，优先看重点或薄弱环节。

(4) 情境分析: 情境分析是一种评估和分析特定情境或问题的过程, 通常用于深入了解各种因素、背景信息和潜在结果。在面对复杂或多变的情境时, 用户可能会难以全面考虑所有相关要素, 导致决策过程不够清晰或有效。大模型可以帮助用户从多角度分析问题, 提供不同的视角和建议。例如, 大模型可能考虑一系列可能性, 从最好的情况到最坏的情况, 以及可能影响每一种结果的因素。以下是一个示例:

示例 3.39

提示词：如果我辞去全职工作去从事自由职业，可能会发生哪些情况？

DeepSeek 模型输出（部分）：



辞去全职工作从事自由职业会带来更大的自由和灵活性，但也伴随着不确定性和挑战。以下是可能发生的情况，供你参考：

1. 经济方面

✅ **可能的好处**：

- **收入潜力增加**：如果能力强、客户稳定，收入可能超过全职工作。
- **灵活定价**：可以自主决定服务价格，随着经验增长提高费率。

⚠️ **可能的挑战**：

- **收入不稳定**：没有固定工资，可能某个月赚很多，下个月却很少。
- **现金流压力**：客户可能延迟付款，需预留应急资金（建议至少存够3-6个月生活费）。
- **社保自理**：需自行缴纳养老、医疗保险等（可挂靠灵活就业或代缴机构）。

2. 工作与生活

✅ **可能的好处**：

- **时间自由**：可自主安排工作时间，避开通勤，兼顾生活兴趣。
- **选择客户**：只接自己喜欢的项目，减少职场内耗。

3.4.4 学习辅助

学习是日常生活中的重要组成部分。大模型具备丰富的知识储备和信息整合能力，是一个理想的学习帮手。据清华大学教育研究院 2024 年的一项调查表明[12]，国内高校超过半数的大学生会使用生成式 AI 工具进行信息搜索和辅助自学，反映其在提升学习效率和优化学习方式中的巨大潜力。大模型可根据学习目标和个体需求，协助用户完成知识获取、能力训练、计划制定与复习总结等任务。下文将通过实际案例展示提示词的编写方法及模型在学习过程中的应用效果，帮助读者深入理解如何在不同学习任务中灵活使用大模型。

1. 智能辅导

大模型可以充当学习的智能辅导员，在学习过程中提供实时解答和指导，帮助用户解决难题并提供有效的学习建议。例如，在学习线性代数时，用户可以向模型提问疑难问题，模型不仅可以提供解决步骤，还可以解释相关概念。

[12]信息来源：https://edu.cctv.com/2025/02/25/ARTIuLyjlKYGY7JhzQJcfvMe250225.shtml

通过合理编写提示词，用户可以让大模型提供详细的讲解和指导，帮助其更好地掌握知识点。以下是智能辅导类提示词的一个示例：

示例 3.40

提示词：作为一名线性代数老师，你的任务是帮助我理解矩阵的逆如何计算，并提供详细的解题步骤和相关概念的解释。

DeepSeek 模型输出（部分）：

好的！矩阵的逆是线性代数中的重要概念，理解其计算方法和相关概念对掌握线性代数至关重要。以下是详细的解题步骤和相关概念的解释：

---

1. 矩阵逆的定义

对于一个 $n \times n$ 的方阵 A，如果存在另一个 $n \times n$ 的方阵 B，使得：

$$A \cdot B = B \cdot A = I_n$$

其中 I_n 是 $n \times n$ 的单位矩阵，则称 B 是 A 的逆矩阵，记作 A^{-1}。

注意：

- 只有方阵（行数和列数相等）才可能有逆矩阵。
- 不是所有方阵都有逆矩阵。如果矩阵 A 有逆矩阵，则称 A 是**可逆的**（或**非奇异的**）；否则称 A 是**不可逆的**（或**奇异的**）。

---

2. 判断矩阵是否可逆

在计算逆矩阵之前，需要判断矩阵是否可逆。常用的方法有：

- **行列式法**：如果 $\det(A) \neq 0$，则 A 可逆；否则不可逆。
- **秩法**：如果 A 是满秩矩阵（即 $\text{rank}(A) = n$），则 A 可逆。

---

3. 计算逆矩阵的方法

以下是几种常见的计算逆矩阵的方法：

方法 1：伴随矩阵法

对于 $n \times n$ 的矩阵 A，其逆矩阵公式为：

$$A^{-1} = \frac{1}{\det(A)} \cdot \text{adj}(A)$$

其中：

- $\det(A)$ 是矩阵 A 的行列式。
- $\text{adj}(A)$ 是矩阵 A 的伴随矩阵。

2. 口语训练

大模型可以充当口语老师，进行日常对话练习，提供语音纠错建议，帮助用户提升口语表达能力。与传统口语培训相比，这种方式不仅降低了费用，还

能提供随时随地的训练机会，特别适合需要提高口语水平的学习者。通过合理设计提示词，用户可以根据实际需求，进行各种情景的口语练习。例如，用户可以要求模型扮演某个角色，如酒店接待员、商店售货员等，进行情景对话练习。以下是口语训练类提示词的一个例子：

示例 3.41

提示词：作为一名英语口语老师，你的任务是与我进行一个关于旅行的对话。我将模拟游客角色，你将扮演酒店前台接待员。请根据我的回答实时提供语法和发音的反馈。

不过，需要注意的是，这种提示词需要大模型接入语音功能，使用语音模式才能实现语音对话和发音纠错。目前，具有语音模式的大模型主要包括 OpenAI 的 GPT 系列（见图3.8)、Google 的 Bard、百度的文心一言、腾讯的 Hunyuan 和阿里云的通义千问等[13]。

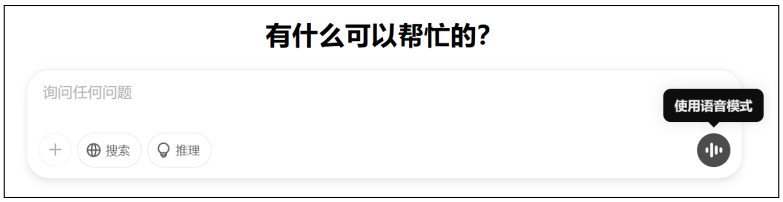


图 3.8: ChatGPT 的语音模式

3. 制订学习计划

大模型可以帮助用户根据学习目标和时间安排，设计个性化的学习计划，确保学习效率和进度。例如，在大学生面临紧张的考试周期时，大模型可以帮助合理分配复习时间，设置学习优先级，并提供针对性的学习建议和资源推荐。以下是制定学习计划时的一个提示词示例：

[13]截至 2025 年 3 月 20 日，DeepSeek 模型暂未接入语音功能。

示例 3.42

提示词：*为一名学习助理，你的任务是为我制定一份详细的复习计划，目标是在三周内复习完所有的高等数学内容。请根据我的学习进度和考试日期提供每天的学习内容安排，并推荐相关学习资源。*

DeepSeek 模型输出（部分）：

三周高等数学复习计划

目标：

- 在三周内复习完所有高等数学内容。
- 每天学习时间：4-6小时（可根据个人情况调整）。

复习策略：

- **第一周**：基础知识复习，重点回顾微积分、线性代数和概率论的核心概念。
- **第二周**：深入练习，解决典型题目，强化理解和应用能力。
- **第三周**：模拟考试，查漏补缺，重点复习薄弱环节。

第一周：基础知识复习

Day 1-2：微积分（极限、导数、积分）

- **学习内容**：
  - 极限的定义与计算
  - 导数的定义、求导法则与应用
  - 积分的定义、基本积分公式与应用
- **推荐资源**：
  - 《高等数学》（同济版）第1-4章
  - Khan Academy微积分课程
  - 3Blue1Brown微积分视频系列
- **任务**：
  - 完成课后习题，重点练习极限、导数和积分的基本计算。

3.5 讨论

讨论 3.1. 请讨论不同类型的大模型提示词的技巧差异。

讨论 3.2. 请讨论推理型大模型提示词中加入角色提示的效果。

3.6 习题

习题 3.1. 以下哪项是提示词中“任务”要素的核心作用？

A. 定义模型的身份

B. 明确用户需完成的目标
C. 提供背景信息
D. 限制回答长度

习题 3.2. 在一个客服对话系统中，模型可能扮演哪些角色？每个角色的主要职责是什么？

习题 3.3. 在生成新闻摘要时，如何设置输出控制以确保摘要的长度和风格符合要求？

习题 3.4. 以下提示词存在什么问题？如何改进？
原提示：“写一篇关于气候变化的文章。”

习题 3.5. 请分析在缺乏上下文的情况下，模型可能出现的错误类型，并提出改进建议。

习题 3.6. 请针对任务“解释量子计算”，列出两种可增强模型输出深度的上下文信息。

习题 3.7. 在生成诗歌时，如何通过输出控制调整模型的创意和风格？

习题 3.8. 在设计提示词时，任务目标与角色定义如何相互影响？举例说明：如果任务要求模型“生成一篇科普文章”，但未明确角色（如“专业学者”或“动物科普作家”），可能导致哪些问题？如何通过调整角色优化输出结果？

习题 3.9. 当上下文信息复杂时（如用户提供多段背景文本），如何通过输出控制避免模型生成偏离核心目标的内容？试讨论“限制输出长度”“明确格式要求”和“禁止敏感内容”三类控制手段的优先级与适用场景。

第四章　大模型辅助工作

4.1　概述

上一章节探讨了提示词的设计要义，重点介绍了如何通过构建提示词来引导大模型输出精准的内容。这一基础性工作对于充分发挥大模型潜力至关重要。然而，在实际工作中，如何有效利用大模型来辅助完成各类专业任务，提升工作效率并降低使用门槛，仍然是许多用户关注的核心问题。

在这一背景下，大模型技术正逐步融入各类专业任务，为提升工作效率提供有力支持。例如，凭借强大的语义理解能力，大模型能够高效解析用户需求，从海量数据中精准提取关键信息，显著提升信息获取效率与准确性。此外，大模型还能优化各类文档处理流程，如文档整理、格式调整、摘要生成等基础性任务，并借助其卓越的内容生成能力提升创作效率，支持从初稿撰写、内容扩展到风格优化与语言润色的全流程内容生产。同时，随着多模态技术的发展，大模型的辅助作用已拓展至图片、音频、视频等多媒体内容创作领域。通过整合文本与多媒体处理能力，大模型正在推动智能化工作方式的革新，使内容生产与信息处理更加高效、便捷。

本章将进一步探讨大模型在实际工作中的辅助功能，聚焦大模型在信息检索、文本办公和创作支持等方面的具体应用，重点关注其如何通过优化提示词与任务目标的对接，实现多维度的智能辅助，提升工作效率并满足多样化的需求。值得特别关注的是，随着通用性 AI 智能体技术的快速发展（详见本书第9.3章节），大模型赋能工作效率的潜力还将持续增强，为下一阶段的智能办公提供新的可能性。

4.2　大模型联网检索

信息检索是日常工作和学习中获取知识的重要方式。网络检索是指通过互联网查找和获取信息的过程，如搜索引擎查询。传统网络检索主要依赖关键词

匹配。用户输入查询后，搜索引擎会在预构建的索引中查找相关网页，并按相关性排序返回结果。虽然这种方式高效，但它只提供网页链接，无法理解查询的语义，也不会整合信息，用户需要自行进行筛选和归纳。

基于大模型的联网检索方法提供了更为智能的解决方案。与传统检索方式不同，它能够理解查询意图，并自动完成信息整合、筛选和归纳工作。接下来，本节将详细介绍这种基于大模型的联网检索方法，探讨其如何优化信息检索过程并提升用户体验。

4.2.1 联网检索流程

大模型联网检索是指大语言模型结合网络搜索能力，通过自然语言理解和信息整合技术提升检索效果。相比传统检索方式，大模型不仅能够基于关键词匹配查询，还能理解查询意图，自动调整检索策略，并结合多种信息来源生成结构化答案。

大模型完成联网搜索的过程，本质上是一个从海量信息中高效获取、处理和输出的智能流程。如图4.1所示，它主要包括以下三个步骤：

1. 语义解析与信息检索：大模型首先会对用户的查询进行语义解析，提取关键信息，然后调用搜索工具获取最新的相关信息。

2. 内容筛选：大模型对不同来源的信息进行归纳整理，并通过可信度评估与信息过滤，筛除无效或低质量内容。

3. 信息整合与输出：大模型基于筛选后的信息进行深度加工，通过凝练总结、重组结构等方式，生成清晰、连贯的回答，为用户提供高质量的知识输出。

相比传统搜索方式，大模型不仅能自动检索信息，还能高效完成筛选、归纳和总结。用户无需筛选网页就可以直接获得完整、精准的答案，大幅提升了信息获取的效率和准确性。

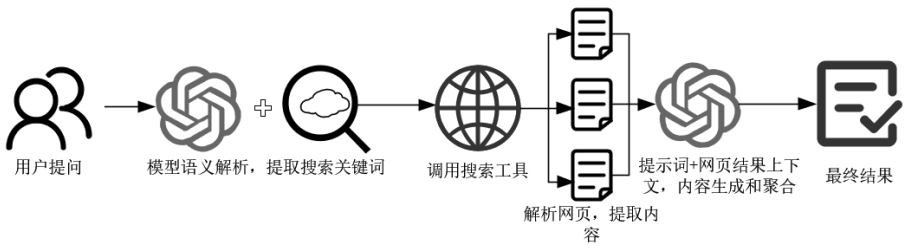
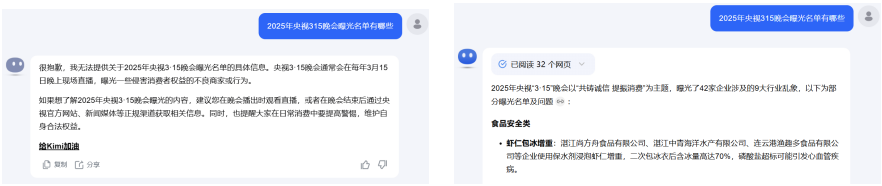


图 4.1: 大模型联网检索流程

联网检索功能并不是大模型自带的，而是一个额外的集成工具，它是实现大模型联网问答的关键步骤。大模型本身（比如 Kimi 或其他 AI 模型）是通过预先训练的海量数据学习知识，在生成回答时并不会实时访问互联网。因此，它的知识是静态的，通常只覆盖到某个时间点，无法获取训练后出现的新知识，也难以查找特定的实时数据或最新研究成果。

联网检索功能的引入使得大模型能够调用外部搜索工具，从互联网上获取最新或特定的信息，对检索到的信息进行筛选、去重、整理，提取最有价值的内容，并基于检索到的信息和自身知识库整合答案，生成流畅、准确的回答，最后将最终答案返回给用户，并提供参考来源。例如，联网检索可以提供最新的新闻报道、当前天气状况或实时行业趋势，使大模型回答更加准确，时效性更强。图4.2展示了 Kimi 模型在回答“2025 年央视 315 晚会曝光名单有哪些”这一问题时的对比情况。未联网时，模型由于缺乏最新数据，无法提供准确的答案，而启用联网检索后，模型能够实时获取相关信息，并给出更加全面、及时的回答。



(a) 未使用联网检索　　(b) 使用联网检索（部分输出）

图 4.2: Kimi 模型对同一问题的回答对比

4.2.2 联网检索分类

目前，许多先进的大模型都配备了联网检索功能，能够实时访问互联网，获取最新的信息。这些大模型通过与各种检索平台的集成，弥补了传统模型知识库的静态性，能够提供基于最新数据的准确答案。表4.1总结了目前一些常见的具有联网检索功能的大模型及其网站[1]。

[1]具备联网搜索功能的 AI 工具：https://www.53ai.com/news/qianyanjishu/2024061364053.html?utm_source=chatgpt.com

表 4.1: 具有联网搜索功能的平台

名称	网址
深度求索	https://www.deepseek.com
Kimi	https://kimi.moonshot.cn
百度文心一言	https://yiyan.baidu.com
阿里通义千问	https://tongyi.aliyun.com
腾讯元宝	https://yuanbao.tencent.com
ChatGPT	https://chatgpt.com
Gemini	https://gemini.google.com
Grok	https://x.com/i/grok

联网检索功能可以分为非自动和自动两种类型。

非自动联网搜索是指用户需要手动启用联网检索功能。通常，这类功能通过一个可见的按钮（例如“⊕”图标）进行控制，用户根据需求选择是否启动联网搜索。例如，Kimi 模型提供了一个联网搜索的开关按钮，位于输入框下方，用户可以根据需求手动启动联网检索功能。启用后，模型会通过互联网查询实时信息并更新答案。类似如 DeepSeek、ChatGPT、Gemini 等模型也具备手动启动联网检索的功能（见图4.3），用户可以根据具体的查询需求选择是否启用联网检索，以获得更加精准和时效的答案。

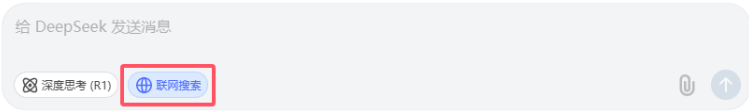


图 4.3: Deepseek 的联网搜索功能

自动联网检索是指大模型在生成回答时默认启用联网搜索功能，无需用户手动干预。像文心一言、通义千问、腾讯元宝和 Grok 等模型，在用户提问后，都会自动访问互联网并实时获取最新的信息，来确保回答的准确性和时效性。如图4.4所示，可以看到对话框周围并没有开关联网搜索功能的按钮，但对话结果依然参考了网页信息源。



图 4.4: 文心一言的联网搜索功能

4.2.3 联网检索优势

传统检索和大模型辅助检索都用于获取信息，但它们在功能和原理上存在明显的区别。传统检索依赖关键词匹配，适用于简单文本查询，但在处理语义模糊或多义词时存在局限。它高度依赖用户输入，要求精准描述需求，否则可能返回无关结果。此外，传统检索难以理解自然语言的上下文关系，通常无法提供与查询意图深度相关的扩展信息或解释。相比之下，大模型辅助检索具有诸多优势，其中包括有效处理模糊查询、具有生成式回答和总结能力以及交互方式多样化等。

1. 有效处理模糊查询

大模型能够通过自然语言处理技术深度理解用户输入的语义，不仅限于表面文字，而是能够解析句子的结构、上下文和隐含意图。它可以在输入信息不完整或存在模糊时，依靠语义上下文进行推理，提供相关的答案。例如，当用户问：“我在成都，今天需要带伞吗?”大模型能够理解这个问题背后的意图是询问当天的天气状况，并根据联网检索提供实时的天气预报进行回答（见图4.5），而传统检索则可能无法识别这个隐含的需求。



图 4.5: 查询天气

2. 生成式回答和总结能力

大模型能够从多个信息源中提取相关数据，并综合这些信息生成结构化的回答或总结。与传统检索不同，大模型能够基于用户查询的需求，自动归纳出关键信息，形成简洁明了的回答，帮助用户迅速理解复杂或庞大的内容。例如，传统检索获取论文信息时，用户需要手动搜索、点击链接并筛选内容。而大模型可以直接提供查询结果，省去了用户筛选的时间和精力（见图4.6）。

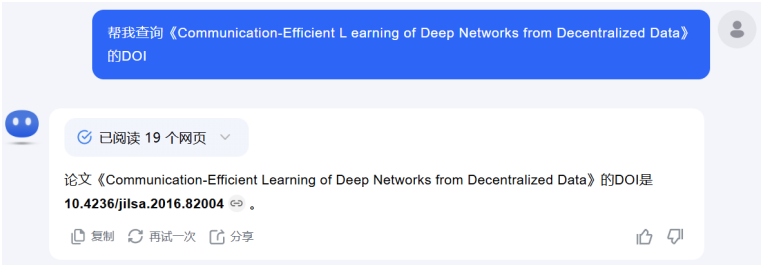


图 4.6: 查询论文的 DOI

3. 交互方式多样化

传统搜索引擎每一轮查询都是独立的，无法根据前一轮的上下文进行调整。大模型能与用户进行连续对话，动态调整回答的内容和深度，从而满足用户在多轮对话中的变化需求。这种连续多轮对话能力使大模型在客户服务、交互式学习和个性化问答中具有显著优势（见图4.7）。



图 4.7: 多轮对话的连续检索

4.3 大模型辅助办公

在现代办公环境中，处理各类文档和日常任务通常需要耗费大量时间和精力。大模型的引入可以为办公工作提供智能化支持，提高工作效率。通过与 Word、PDF、Excel 和 PPT 等办公软件的集成，大模型能够辅助完成文档编辑、信息提取、数据处理和 PPT 制作等任务，从而简化繁琐的工作流程。接下来，本节将介绍大模型在办公任务中的应用，包括在 Word、PDF、Excel 和 PPT 等文档处理中的具体功能和使用技巧。

4.3.1 Word

大模型与 Word 的结合主要用于提高文档编辑和内容优化的效率，例如智能改写、摘要生成、语法检查和格式调整等。本小节将介绍两款常见的 Word 插件——新华妙笔和不坑盒子，分别介绍它们的核心功能及在实际办公场景中的应用。

1. 新华妙笔

新华妙笔[2]是一款由新华社媒体融合生产技术与系统国家重点实验室和博特智能共同研发推出的 AI 模型产品。它主要利用了最新的 AIGC 技术，旨在提高内容生产的效率和质量。新华妙笔特别针对公文写作这一垂直场景进行了深度开发，提供了一套“查、学、写、审”一体化的人工智能公文写作与知识赋能协作平台。

新华妙笔的主要功能包括 AI 写作、AI 校对、AI 润色和 AI 提纲。进入官网后，新华妙笔的工具栏如图4.8所示。用户可直接点击“开始写作”，选择模板后进入编辑界面。在妙笔工具栏中，还提供了 AI 校对、AI 润色和 AI 提纲等功能，帮助用户优化文章内容。

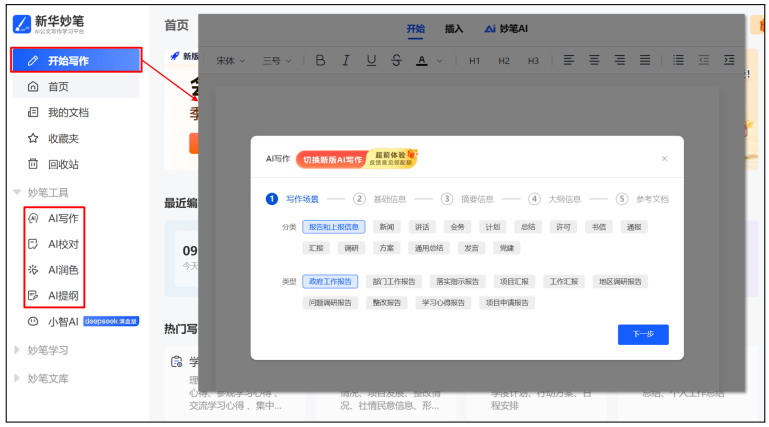


图 4.8: 新华妙笔工具栏

（1）AI 写作功能

AI 写作功能的过程包括选择写作场景、填写标题和关键字、生成摘要信息、生成大纲信息、上传补充文档，最后生成完整的文档。下面以写一篇学习心得报告为例，对每一个步骤进行详细展示。

首先，用户根据写作需求选择分类和类型，如报告、讲话或总结等。在此示例中，用户可选择“报告”下的“学习心得报告”类型，点击下一步，如图4.9所示。

[2]新华妙笔官网：https://miaobi.xinhuaskl.com/recommend



图 4.9: 选择写作场景

然后，用户需要填写文档标题和关键词，如“新时代中国特色社会主义事业课程报告”和“改革开放”。这些信息有助于 AI 理解文档的主题，并生成更符合需求的内容。若用户不确定关键词，可以选择让 AI 自动生成相关的关键词，AI 会根据输入的标题或主题提供几个相关的选项，帮助用户进一步细化文档内容的方向。输入需要生成文档的标题和关键字后，再点击下一步，如图4.10所示。



图 4.10: 填入标题和关键词

AI 根据用户输入的标题和关键词，自动生成文档的摘要。这个摘要概述了报告的主要内容，帮助用户快速把握文档的大致结构和方向，如图4.11所示。如果用户对生成的内容不满意还可以选择修改或者重新生成。



图 4.11: 生成摘要信息

接下来，AI 会根据摘要和关键词生成大纲信息，列出报告的主要部分和段落结构。大纲通常包括引言、主要内容、分析、结论等部分，用户可以根据需要修改或补充大纲，如图4.12所示。



图 4.12: 生成大纲信息

如果用户有额外的资料或文档，AI 还支持上传补充文档（如内容参考和数据参考，见图4.13）。这些文档可以帮助 AI 更精确地理解文档的背景和重点，从而生成更为精准的内容。最后，在完成所有准备工作后，用户点击生成按钮，AI 会自动生成完整的学习心得报告。



图 4.13: 上传补充文档

（2）AI 校对功能

AI 校对功能可以检测出文档中的错别字、搭配错误、常见语法错误等。如图4.14所示，用户上传需要校对的文档后，点击开始校对。



图 4.14: AI 校对

校对后的结果如图4.15所示。



图 4.15: 校对结果

此外，用户还可以调整校对方案和设置白名单。调整校对方案允许用户根据不同的写作需求和标准，定制校对的规则和范围，以适应特定的文档类型或风格（见图4.16）。白名单功能使用户能够指定某些词汇或短语不被视为错误，避免在特定语境下的误报，从而提高校对的准确性和实用性。这些功能特别适用于需要高质量文字编辑的公文写作或学术报告等场景，帮助用户节省大量的校对时间。



图 4.16: 调整校对方案

（3）AI 润色功能

AI 润色功能可以选择优化方向（增强感情、说服力等）、语言风格（正式、专业等）、长度要求（尽量精简等）。除了润色功能以外，用户还可以根据需要进行改写、扩写、仿写等，如图4.17所示。



图 4.17: AI 润色

用户输入需要润色的内容，点击开始生成后，得到如图4.18所示内容，然后用户可以选择继续生成或重新编辑。

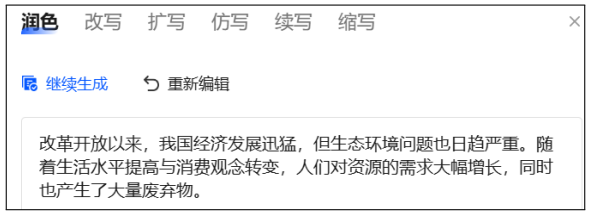


图 4.18: 润色结果

（4）AI 提纲功能

AI 提纲功能可以辅助用户高效构建文章结构，提升写作效率。用户只需输入文章主题，选择生成提纲的句数和数量，设定模板选项，如排比、对比、比喻句等，系统便能智能生成相应的写作提纲。例如，用户在撰写关于环境保护的文章时，可以选择排比句模板，生成如图4.19所示的提纲。该功能对于撰写公文、报告等结构严谨的文体较为实用。



(a) 提纲功能选项 (b) 提纲生成结果

图 4.19: AI 提纲

2. 不坑盒子插件

不坑盒子[3]是一款免费的 Office 插件，兼容 Word、Excel、PPT 以及 WPS 办公套件[4]。其主要功能区包括内容排版、奇妙功能、系统增强和表格等，简化了复杂的文档编辑任务。此外，不坑盒子还提供了教师专区、导航等特色功能，满足不同用户的需求。

不坑盒子需要先下载安装，并在 Word 中启用插件后才能使用。安装完成后，用户可以在 Word 工具栏中找到不坑盒子插件，其 UI 界面如图4.20所示。



图 4.20: Word 中的不坑盒子 UI 界面

目前，不坑盒子插件集成了多项实用功能，主要涵盖内容自动排版、模板库调用、批量处理工具、格式转换、OCR 识别、写作辅助等核心模块。用户可以根据自己的需求选择相关功能，这些功能大多数为一键生成式，能够快速

[3]不坑盒子官网地址：`https://www.bukenghezi.com/`

[4]目前，不坑盒子仅支持 Windows 操作系统。

简化文档处理过程，提升文档处理效率。例如，点击自动排版功能后，用户无需手动调整段落间距、字体大小、行距等排版细节。插件会根据文档的类型和结构自动进行排版，确保文档的整洁与美观。

接下来，主要介绍模板库、智能写作和人工智能功能。关于其他功能，用户可以根据自己的需求进一步探索，在此不做赘述。

（1）模板库调用

不坑盒子的模板库是其核心功能之一，它为用户提供了便捷的文档模板调用与管理服务。模板库涵盖了公文、合同、简历、报告等常见办公文档类型，用户可以根据自己的文档需求，快速选择合适的模板，自动生成符合标准格式的文档，如图4.21所示。

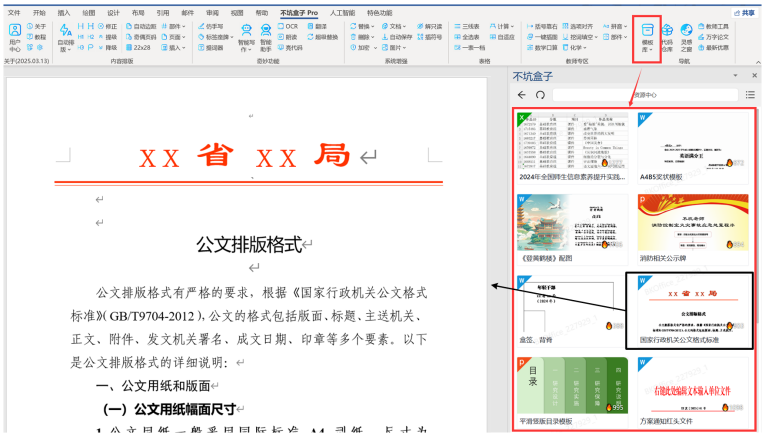


图 4.21: 模板库调用

通过这一功能，用户不仅能够节省大量的排版和格式设置时间，还能确保文档风格的一致性和专业性。此外，用户还可以自定义模板，方便将常用格式保存并进行重复调用，进一步提高工作效率。

（2）智能写作

智能写作是不坑盒子的另一个核心功能，主要为用户提供智能化的写作辅助。该功能通过集成大模型技术，帮助用户生成高质量的文本内容。用户可以根据需要设置写作风格、内容的创意度以及具体的写作要求，极大地提高写作效率。用户先在空白处输入意图创作的内容的指令，如图4.22所示选中后点击智能写作，模型会根据输入的指令进行分析并生成用户所需的文本内容。

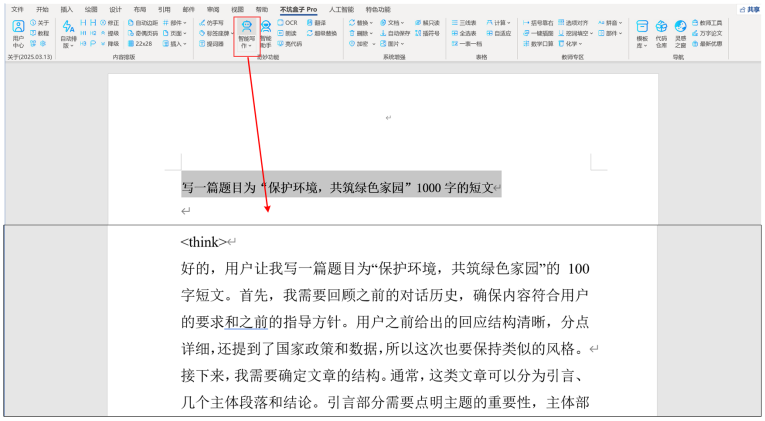


图 4.22: 智能写作

除了直接生成内容外，不坑盒子的智能写作功能还允许用户根据个人偏好创建写手，调整生成文本的风格和创意度。如图4.23所示，用户可以点击“智能写作设置”选择创建一个写手。文档右侧会出现一个如图4.24所示的写手设置的画面，用户可根据自身喜好修改该功能所使用的大模型，通过调整“temperature”和“top_p ”参数来控制生成内容的多样性和创意程度。

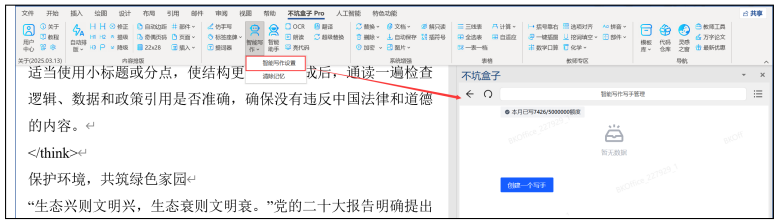


图 4.23: 智能写作设置

temperature 参数用来控制生成文本的随机性，数值较低时（如 0.2）会让模型生成更为保守、精确的内容，而较高时（如 0.8）则会增加文本的多样性和创造性。top_p 参数通过限定生成词汇的概率分布来控制输出的多样性，数值较低时（如 0.3）会限制模型仅从高概率的词汇中选择，较高时（如 0.9）则会增加选择的自由度，允许模型选择更多样化的词汇。需要注意的是，在使用时建议只修改 temperature 或 top_p 其中之一，避免同时修改两个参数对生成结果造成过多干扰。

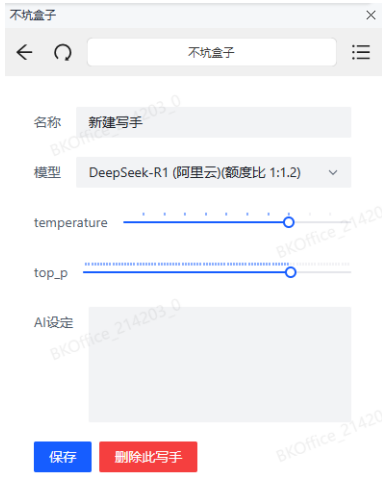


图 4.24: 写手参数设置

（3）人工智能

人工智能模块主要用来优化文本内容，比如内容润色和内容改写，其主要功能如图4.25所示。用户可以选中需要处理的文本内容，并点击相应的 AI 功能，例如精确措辞、调整风格或语法检查等，即可直接调用相关功能，一键优化文本内容。



图 4.25: 人工智能模块

4.3.2 PDF

大模型在 PDF 处理中的应用主要包括内容摘要、信息提取和阅读理解等方面。相比传统的 PDF 阅读工具，大模型不仅能理解文档内容，还能通过交互式对话帮助用户快速定位关键信息，减少手动查找的时间成本。目前，专门针对 PDF 的智能工具较少，本小节将重点介绍 ChatPDF[5]，探讨其核心功能及在 PDF 阅读与分析中的应用。

ChatPDF 是一款基于 OpenAI Chat API 的智能文档阅读工具，特别适用于研究者、大学生和职场人士等需要频繁阅读和分析 PDF 文档的用户。它能够让用户与 PDF 文档进行交互式对话，提出问题、获取摘要信息，并直接

[5]ChatPDF 网址:https://www.chatpdf.com/

从内容中提取特定信息。相比传统的 PDF 阅读方式，ChatPDF 能够更高效地帮助用户理解文档内容，尤其适用于学术论文、书籍、报告等多种文档类型。进入 ChatPDF 官网后，界面如图4.26所示，用户只需上传 PDF 文档，系统便会自动解析内容，随后即可进行交互查询。例如，如果用户正在阅读一篇冗长的研究论文并需要查找特定数据，可以直接询问：“这项研究的主要发现是什么?”或“总结结论”。ChatPDF 将快速分析文档并提供精准回答，从而帮助用户高效获取关键信息，节省时间和精力。



图 4.26: ChatPDF 对话框

ChatPDF 提供多种功能，包括智能摘要、信息提取、深度思考与探索等。这些功能通过对话形式实现，用户可以与文档直接互动，提出问题并快速获取所需的信息。

（1）智能摘要

例如，用户上传了一篇名为《A Comprehensive Examination of the Potential Application of Chat GPT in Higher Education Institutions》的文章，并在对话框中输入“这项研究的主要发现是什么?”，ChatPDF 会自动提取文档的关键内容并提供简洁明了的总结，帮助用户快速了解文档的核心信息。返回结果如图4.27所示。

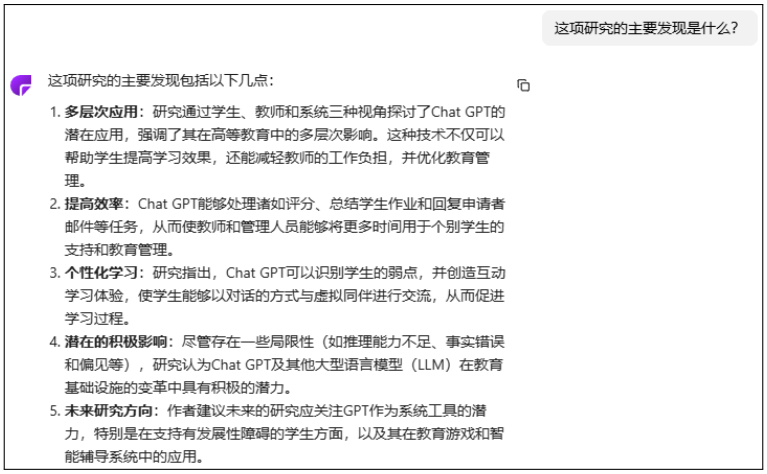


图 4.27: 智能摘要

（2）信息提取

如果用户想查找某一章节中引用了哪些论文，并把它们整理出来，可以在对话框中输入“第一节 introduction 中引用了哪些论文，请帮我整理出来。”，ChatPDF 会根据文档中的内容自动识别引用的文献，并列出相关信息，帮助用户迅速找到文献出处。返回结果如图4.28所示。

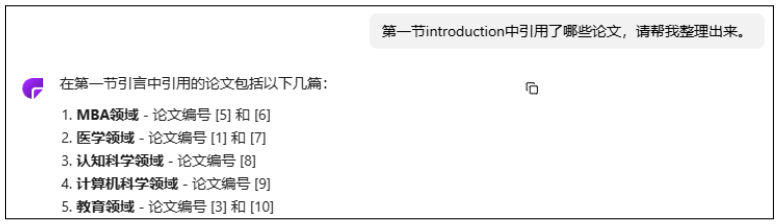


图 4.28: 细节信息提取

（3）深度思考与探索

如果用户希望在文章的基础上进行更深入的研究，但缺乏明确的思路，可以在对话框中输入“基于文章内容，还有哪些值得进一步研究的问题?”或“文章中提到的研究还有哪些可以扩展的方向?”等问题，ChatPDF 会根据文档的核心内容提出相关的延伸思路和研究方向。返回结果如图4.29所示。

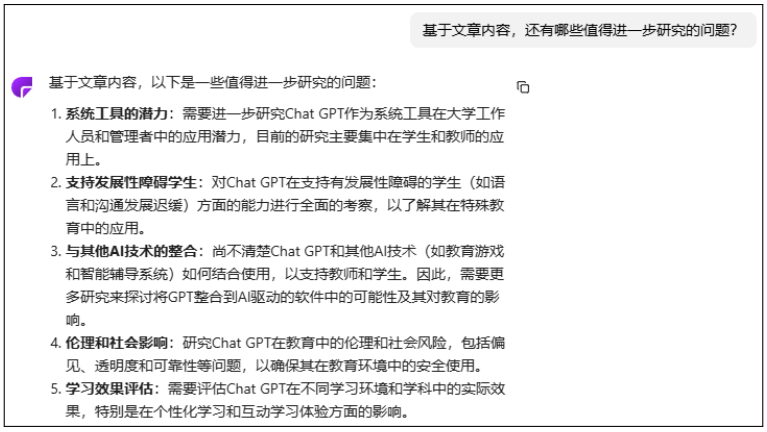


图 4.29: 进一步研究

此外，用户还可以探讨该文章的实际应用和具体实施方法。例如，输入“如何将文章中的理论应用到实际工作中?”ChatPDF 会帮助用户将理论与实践结合，提供相关的应用建议和实施步骤，返回结果如图4.30所示。

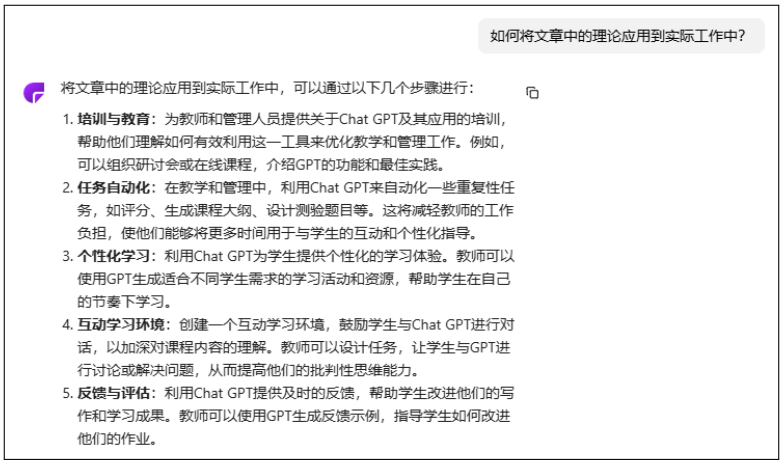


图 4.30: 应用与实践

如果文章中包含多个表格或图片，用户可以让 ChatPDF 分析表格中的数据内容，或者整理每张图片的具体信息。用户只需在对话框中提问关于表格或图片的问题，ChatPDF 便会根据文档中的信息进行分析并返回答案。由于示例文章中未包含表格或图片，在此不再赘述。更多功能读者可自行探索。

4.3.3 Excel

大模型与 Excel 的结合主要用于提高数据处理的效率、准确性和智能化水平，例如数据分析、自动化公式生成、表格处理和可视化展示等。通过大模型的辅助，Excel 用户能够更高效地进行数据操作和分析，同时提高数据处理的智能化程度。本小节将主要介绍两个与 Excel 相关的智能工具：OfficeAI[6]和 ChatExcel[7]，重点探讨它们的核心功能及应用方式。

1.OfficeAI

OfficeAI 助手是一款免费的智能办公工具，专为 Microsoft Office 和 WPS 用户设计。它帮助用户快速解决各种办公问题，比如自动完成公式计算、选择合适的函数等。其 UI 界面如图4.31所示。



图 4.31: OfficeAI 界面

OfficeAI 最为亮眼的地方是用户可以通过与助手进行对话，直接完成所需操作。用户可以在图4.32所示的可交互聊天框内输入想要执行的指令，助手会自动解析并执行。

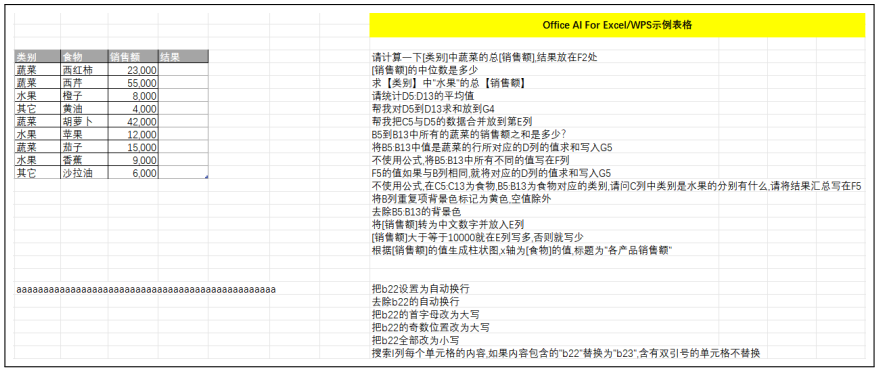
[6]OfiiceAI 网址：`https://www.office-ai.cn/`

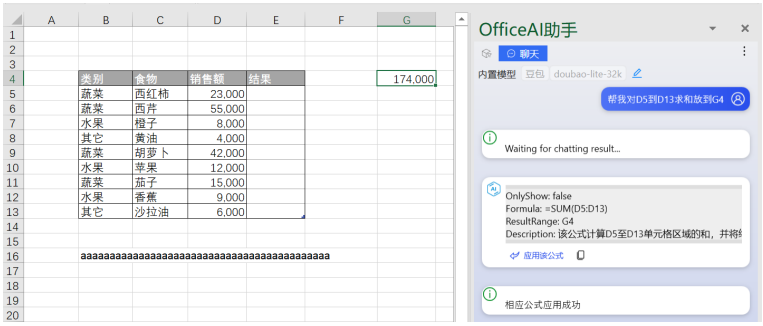
[7]ChatExcel 网址:`https://chatexcel.com/`



图 4.32: OfficeAI 助手聊天框

OfficeAI 的 Excel 插件为用户提供了多种实用功能。其官方示例文档中罗列了许多指令输入提示词（见图4.33），涵盖了数据分析、数据处理、智能替换和数据可视化等多种操作，帮助用户高效地处理和展示数据。


Office AI For Excel/WPS示例表格
类别 食物 销售额 结果
蔬菜 西红柿 23,000
蔬菜 西芹 55,000
水果 橙子 8,000
其它 黄油 4,000
蔬菜 胡萝卜 42,000
水果 苹果 12,000
蔬菜 茄子 15,000
水果 香蕉 9,000
其它 沙拉油 6,000
请计算一下[类别]中蔬菜的总[销售额],结果放在F2处
[销售额]的中位数是多少
求【类别】中"水果"的总【销售额】
请统计D5:D13的平均值
帮我对D5到D13求和放到G4
帮我把C5与D5的数据合并放到第E列
B5到B13中所有的蔬菜的销售额之和是多少?
将B5:B13中值是蔬菜的行所对应的D列的值求和写入G5
不使用公式,将B5:B13中所有不同的值写在F列
F5的值如果与B列相同,就将对应的D列的值求和写入G5
不使用公式,在C5:C13为食物,B5:B13为食物对应的类别,请问C列中类别是水果的分别有什么,请将结果汇总写在F5
将B列重复项背景色标记为黄色,空值除外
去除B5:B13的背景色
将[销售额]转为中文数字并放入E列
[销售额]大于等于10000就在E列写多,否则就写少
根据[销售额]的值生成柱状图,x轴为[食物]的值,标题为"各产品销售额"
aaaa

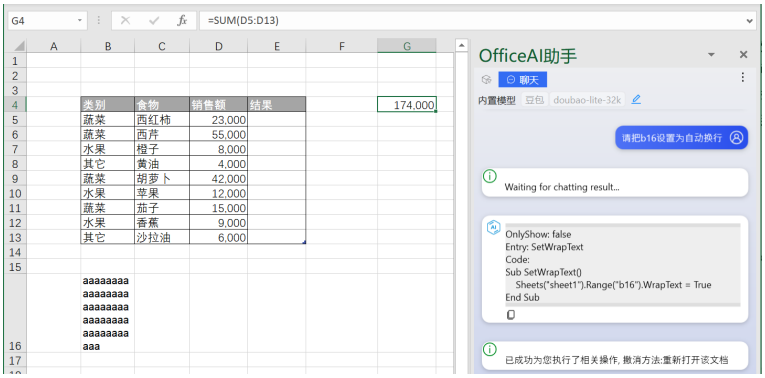


类别	食物	销售额	结果
蔬菜	西红柿	23,000	
蔬菜	西芹	55,000	
水果	橙子	8,000	
其它	黄油	4,000	
蔬菜	胡萝卜	42,000	
水果	苹果	12,000	
蔬菜	茄子	15,000	
水果	香蕉	9,000	
其它	沙拉油	6,000	

图 4.34: 单元格求和

（2）数据处理

用户可以与 AI 对话进行单元格格式设置。例如，图4.34中单元格 b16 处有一个很长的字符，通过与 AI 对话，用户可以让这一串很长的字符实现自动换行，与上一行单元格对齐。用户在对话框中输入“请把 b16 设置为自动换行”，执行结果如图4.35所示。



类别	食物	销售额	结果
蔬菜	西红柿	23,000	
蔬菜	西芹	55,000	
水果	橙子	8,000	
其它	黄油	4,000	
蔬菜	胡萝卜	42,000	
水果	苹果	12,000	
蔬菜	茄子	15,000	
水果	香蕉	9,000	
其它	沙拉油	6,000	

图 4.35: 设置单元格格式

（3）智能替换

用户可以通过 AI 对话实现复杂的替换功能，如将特定字符替换成用户需要的字符，或者将特定位置的字母进行替换。例如，用户在对话框中输入“请把 b16 的奇数位置替换成汉字川，将偶数位置替换成大写的 A”，执行结果如图4.36所示。

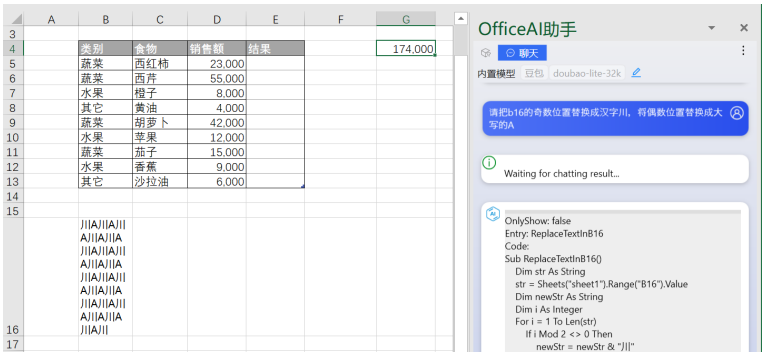


图 4.36: 智能替换

（4）数据可视化

用户可以通过 AI 对话将表格数据进行可视化，绘制成折线图、柱状图、饼图等。例如，用户在对话框中输入“根据 D5:D13 的值生成柱状图，x 轴为 C5:C13 的值，标题为‘各产品销售额’，执行成功后会直接弹出一个如图4.37所示的柱状图。

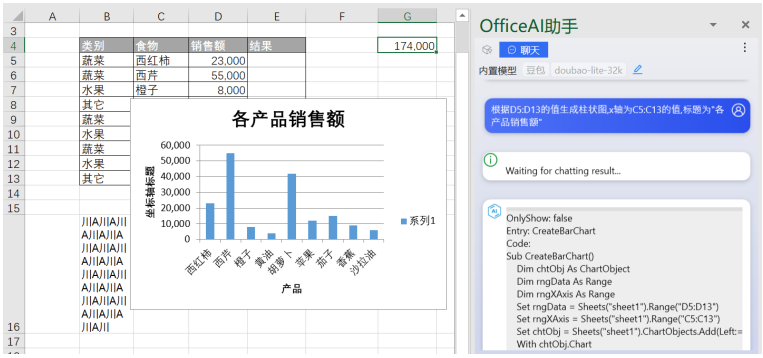


图 4.37: 数据可视化

2.ChatExcel

ChatExcel 是由北大团队开发的 AI 工具，主要用于处理 Excel 表格，其工具栏如图4.38所示。用户通过聊天的方式描述需求，ChatExcel 自动执行相应的 Excel 操作。



图 4.38: ChatExcel 工具栏

ChatExcel 提供了多种实用功能，在此选取了几种常用的功能进行介绍，包括数据清洗、逻辑运算、相关性分析和图表生成等。

（1）数据清洗

用户可以通过数据清洗功能进行处理缺失值、去除重复数据、纠正错误数据等。如图4.39所示，Excel 表格里面存在用户不想要的空白单元格，用户可以输入“删除空白单元格: 删除所有有空白单元格的行，处理好的文件，下载给到我”来使用数据清洗功能删除空白单元格。



图 4.39: 请求数据清洗功能

数据清洗前后结果对比如图4.40所示：



世界排名	学校名称	地区	综合得分	学术声誉	雇主声誉	每位教员引用率	师生比
1	麻省理工学院	美国	100	100	100	100	100
2	剑桥大学	英国	98.8	100	100	92.3	100
3	斯坦福大学	美国	98.5	100	100	99.9	100
4	牛津大学	英国	98.4	100	100	90	100
4	某某大学1		80	90		100	
5	哈佛大学	美国	97.6	100	100	100	99.4
6	加州理工大学	美国	97	96.5	87.1	100	100
6	帝国理工学院	英国	97	98.3	99.7	86.5	99.5
8	伦敦大学学院	英国	95	99.4	98.6	77	97.6
9	苏黎世联邦理工大	瑞士	93.6	98.6	91.3	99.2	74.2
		未知	0	99	0		
10	芝加哥大学	美国	93.2	99.2	92.2	86.8	92.9
11	新加坡国立大学	新加坡	92.7	99.5	94.1	91.8	79.8
12	北京大学	中国大陆	91.3	99.3	96.5	96.7	87.3
13	宾夕法尼亚大学	美国	90.6	96.5	92.5	70.9	99.9
14	清华大学	中国大陆	90.1	98.9	97.7	98.1	92.8
15	爱丁堡大学	英国	89.5	98	97.2	70.1	81.1
16	洛桑联邦理工学院	瑞士	89.2	83.7	71.8	99.4	91.1
16	普林斯顿大学	美国	89.2	99.9	98.9	100	72.6
18	耶鲁大学	美国	89	99.9	100	51.5	100
18		意大利	83	92		48	

图 4.40: 删除空白单元格前后对比

（2）逻辑运算

用户可以通过逻辑运算功能进行条件判断与分类、数据筛选与过滤、条件计算等。如4.41所示，有一个年级的学生成绩表共 6 门科目，每个科目 100 分，满分 600 分，用户想根据每个学生的总分成绩批注 ABCDEF 等不同等级，可以在对话框里输入对应指令“满足单一条件: 填充‘等级’列，对全年级学生总成绩进行等级划分: 500-600 为 A，450-499 为 B，420-449 为 C，400-419 为 D，0-399 为 F。填充到表格里，下载给到我”。逻辑运算前后结果对比如图4.42所示。

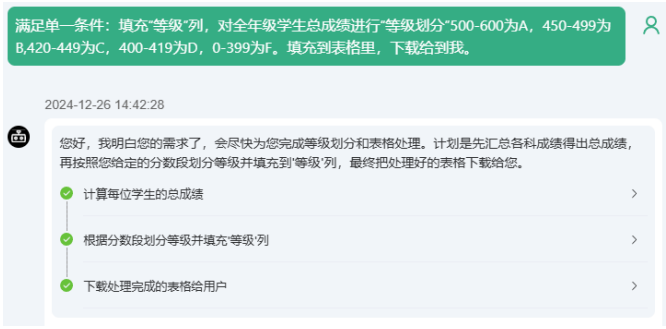


图 4.41: 请求逻辑运算功能

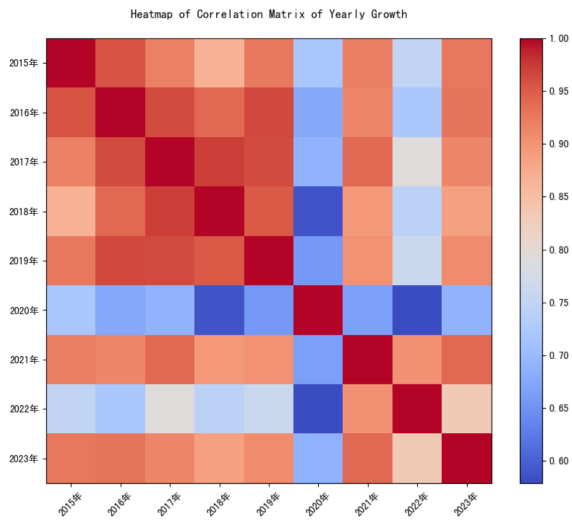
	A	B	C	D	E	F	G	H	I	J	K	L
1	班级	学号	姓名	性别	数学	语文	英语	物理	生物	地理	等级	
2	三(一)班	101	曹玉凤	女	51	79	83	95	51	84		
3	三(一)班	102	陈虹	男	52	73	53	76	90	98		
4	三(一)班	103	严嘉	男	99	85	66	62	98	60		
5	三(一)班	104	李大芬	女	88	80	61	89	95	93		
6	三(一)班	105	康蓓	女	72	69	47	58	56	66		
7	三(一)班	106	孔香秀	男	68	71	76	73	50	96		
8	三(一)班	107	周南莲	男	87	49	55	66	84	53		
9	三(一)班	108	金明露	女	92	91	51	94	62	84		
10	三(一)班	109	孙翠花	男	69	54	94	93	65	89		
11	三(一)班	110	秦灿灿	女	71	47	72	93	67	56		

	A	B	C	D	E	F	G	H	I	J	K	L
1	班级	学号	姓名	性别	数学	语文	英语	物理	生物	地理	等级	总成绩
2	三(一)班	101	曹玉凤	女	51	79	83	95	51	84	C	443
3	三(一)班	102	陈虹	男	52	73	53	76	90	98	C	442
4	三(一)班	103	严嘉	男	99	85	66	62	98	60	B	470
5	三(一)班	104	李大芬	女	88	80	61	89	95	93	A	506
6	三(一)班	105	康蓓	女	72	69	47	58	56	66	F	368
7	三(一)班	106	孔香秀	男	68	71	76	73	50	96	C	434
8	三(一)班	107	周南莲	男	87	49	55	66	84	53	F	394
9	三(一)班	108	金明露	女	92	91	51	94	62	84	B	474
10	三(一)班	109	孙翠花	男	69	54	94	93	65	89	B	464
11	三(一)班	110	秦灿灿	女	71	47	72	93	67	56	D	406

图 4.42: 逻辑运算结果前后对比

（3）相关性分析

用户可以通过相关性分析功能计算变量间相关系数、可视化关联关系（如相关性热力图）等。例如，用户想知道某种数据的增长是否与时间存在相关性，可以输入指令“2014 年至 2023 年，各地区的数据增长与时间之间是否存在相关性”使用 ChatExcel 的相关性分析功能。相关性分析结果如图4.43所示。



数据分析与结论

数据增长趋势

通过提取并计算2014年至2023年各地区的数据，我们发现不同地区在这十年间的数据变化呈现出不同的增长模式。例如，北京市在2014年至2023年间的数据从22926.0增长到43760.7，而山西省在同一时间段内的增长相对较小。

相关性分析

为了验证各地区的数据增长与时间之间是否存在相关性，我们进行了相关性矩阵的计算，并生成了热图以直观展示这些关系。相关性矩阵显示：

- 各年份之间的相关性普遍较高，大部分年份的相关系数接近于1，这表明随着时间推移，各年份的数据变化具有很强的一致性。
- 特别是相邻年份之间的相关性尤为显著，如2015年与2016年的相关系数为0.956769，2022年与2023年的相关系数为0.832108，这些高相关性进一步证实了数据增长与时间的紧密联系。

结论

根据以上分析，可以得出以下结论：2014年至2023年间，各地区的数据增长与时间之间存在显著的相关性。这种相关性不仅体现在整体趋势上，还具体表现在每年数据变化的高度一致性。因此，可以认为，在这段时间内，各地区的数据增长与时间有着明显的正相关关系。

图 4.43: 相关性分析结果

（4）图表生成

用户可以通过图表生成功能将数据进行可视化，例如生成折线图、柱状图、饼图等。例如，用户输入指令“使用 2018 年到 2023 年浙江、山东、北京的数据，横轴为‘省份’，‘纵轴为 GDP 增速’绘制折线图”，生成的折线图如图4.44所示。

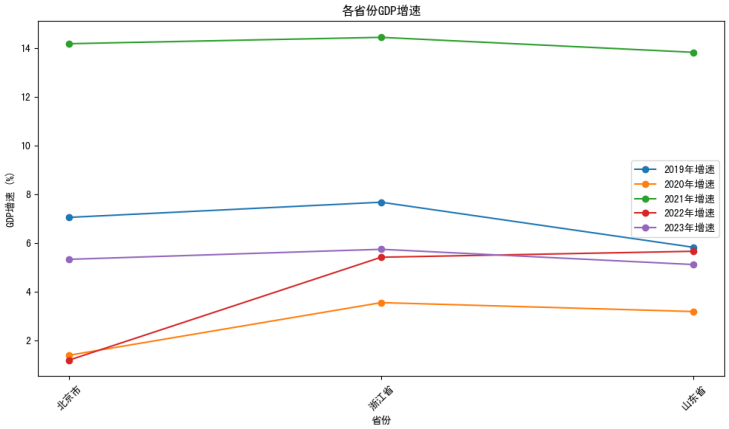


图 4.44: 折线图绘制

4.3.4 PPT

大模型在 PPT 制作中的应用主要包括内容生成、智能排版和美化优化等方面，能够帮助用户快速创建高质量的演示文稿。目前，已有多种 PPT 自动生成解决方案，本小节将主要介绍 DeepSeek+Kimi 和讯飞智文，介绍它们的核心功能及应用方式。

1.DeepSeek+Kimi

DeepSeek+Kimi 制作 PPT 的流程本质上是利用 DeepSeek 的大模型生成能力，根据用户需求生成 PPT 大纲，并将该大纲传递给 Kimi 内部集成的 PPT 助手。二者通过自然语言交互与多模态生成技术，将用户需求快速转化为结构化内容与视觉化演示文档，大幅降低传统 PPT 的制作时间。

DeepSeek+Kimi 制作 PPT 主要包括以下几个步骤。

（1）大纲生成

首先，用户需要根据自己的需求撰写提示词，通过对话描述 PPT 的主题、目标受众和主要内容。DeepSeek 模型会自动提取关键词并生成一个结构化的大纲，包括章节标题、关键要点和逻辑层次。如示例 4.1所示在 DeepSeek 的对话框内输入一段获取 PPT 大纲的提示词。

示例 4.1

提示词：请帮我生成一份详细介绍全球人工智能发展情况的 ppt 大纲，受众为大学本科生。

DeepSeek 模型输出（部分）：



（2）PPT 生成

在得到 PPT 大纲后，打开 Kimi+，找到 PPT 助手（见图4.45），在对话框里输入由 DeepSeek 输出的 PPT 大纲。



图 4.45: kimi 中的 PPT 助手

Kimi 会对输入的 PPT 大纲进行再优化，调整逻辑流畅度，随后点击如图4.46所示文段末尾的“一键生成 PPT”即可获得一份 PPT。

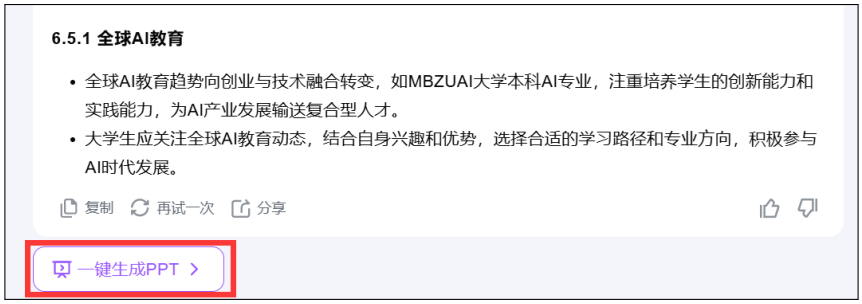


图 4.46: kimi 生成 PPT

（3）PPT 优化

在生成 PPT 的过程中，用户可以根据个人喜好优化细节。Kimi 提供了多种个性化设置（见图4.47），例如模板选择、设计风格、配色方案自定义等。



图 4.47: 模板选择

用户选择相关配置，会得到如图4.48所示的一份完整 PPT。用户可以选择直接下载，或者继续编辑优化内容。用户可以对文档中的内容进行编辑，包括增删文字、调整排版、替换图片等，如图4.49所示。



图 4.48: 完整的 PPT



图 4.49: PPT 优化选项

2. 讯飞智文

DeepSeek 生成的 PPT 大纲不仅可以和 Kimi 的 PPT 助手结合，也可以用于其他 AI 生成 PPT 的工具，如讯飞智文[8]。

讯飞智文是由科大讯飞推出的一款基于人工智能技术的文档创作平台。它利用了科大讯飞的讯飞星火认知大模型，具备强大的文本生成和处理能力。用户可以通过输入主题或内容，讯飞智文能够快速自动生成 PPT，支持一句话主题、长文本、音视频等多种输入方式。此外，它还提供了文案改写、润色、扩写、翻译等多种文本编辑功能，帮助用户优化文稿内容，从而提高创作效率和文档

[8]讯飞智文:https://zhiwen.xfyun.cn/

质量。在讯飞智文的主界面，点击左侧的“AI PPT”选项，界面如图4.50所示。该界面主要包括 4 个功能区域：一句话创建、文本创建、文档创建和高级创建。



图 4.50: 讯飞智文 AI PPT 界面

（1）一句话创建

一句话创建功能允许用户通过输入简短的主题描述，快速生成 PPT 文档。用户只需在输入框中输入一句话，讯飞智文会根据用户输入的需求生成 PPT 的主题、大纲和内容，然后用户可以根据需求进行编辑和修改，如输入主题、选择语言、是否需要配图等，甚至可以添加演讲配图和打开联网搜索功能。如图4.51所示，用户输入“帮我生成一个介绍北京故宫的 PPT”并点击确认。



图 4.51: 一句话创建

随后讯飞智文会生成该 PPT 的大纲，用户可以选择修改大纲或者保持不变，如图4.52所示。



图 4.52: PPT 的大纲

点击下一步后，用户选择需要的模板，挑选喜欢的风格和颜色，如图4.53所示。



图 4.53: 模板选择

点击开始生成，最后可以得到一个完整的 PPT，如图4.54所示。

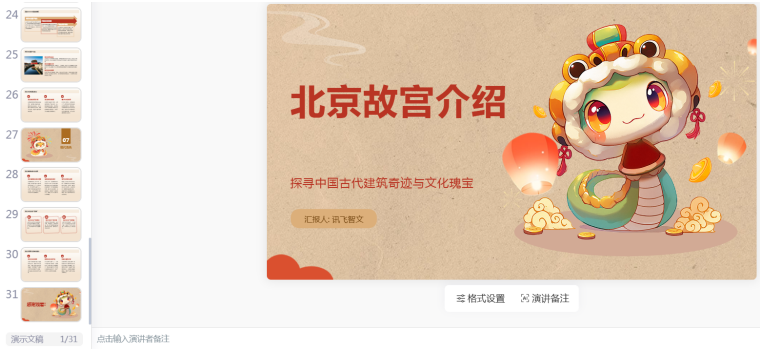


图 4.54: 完整的 PPT

（2）文本创建

文本创建功能允许用户输入或粘贴一段长文本或者整篇文章，系统将自动整理和提炼这些内容，生成相关的标题和大纲，然后创建形成结构化的 PPT。

例如，用户在输入框中输入一篇题为《美丽的小兴安岭》的文章，点击“下一步”，系统将自动分析文章内容，提取主要观点和章节结构，生成相应的 PPT 大纲（见图4.55a）。随后，用户可以选择合适的模板和配色方案（见图4.55b），系统将根据生成的大纲和所选模板，自动生成完整的 PPT 演示文稿，如图4.56所示。



(a) 提取大纲　　(b) 模版选择

图 4.55: 文本创建 PPT

对生成后的 PPT，用户可以选择进一步优化和定制演示文稿。例如，用户可直接在平台上修改文本内容，调整文字样式和排版，或者更改内容的版式和布局。



图 4.56: PPT 生成效果

（3）文档创建

文档创建功能与文本创建类似。用户可以上传各种格式的文档，如 PDF、DOC、TXT 等，系统将自动分析并提炼关键信息，生成相应的大纲和内容，最终创建结构化的 PPT，如图4.57所示。用户也可以自由编辑自动生成的大纲，选择不同的主题风格。 该功能适用于需要将现有文档快速转换为演示文稿的场景，极大地提高了文档制作的效率。



图 4.57: 文档创建

（4）高级创建

高级创建功能提供了丰富的模板支持，帮助用户高效制作专业的 PPT 演示文稿。用户可以根据自己的需求，选择适合的模板，例如论文选题、框架设计、内容填充等方向。

如图4.58所示，用户想制作年终汇报 PPT，可点击“年终汇报”模板，输入框中将自动显示模板提示词。用户只需根据个人信息填写相关内容，系统会根据输入内容自动生成 PPT 大纲和页面，极大地提高制作效率。



图 4.58: 高级创建

此外，用户可以根据自身需求调整文本设置以及配图风格等，创建符合自己风格和需求的 PPT，如图4.59所示。



图 4.59: 个性化设置

4.4 大模型辅助创作

随着人工智能技术的飞速发展，大模型在多媒体创作领域的应用日益广泛。无论是图像、视频还是音乐，大模型都展现出了强大的生成能力和创造力，为艺术创作、内容生产和技术创新带来了新的可能性。本节将深入探讨如何利用大模型生成图片、视频和音乐，并介绍相关的工具和技术，帮助读者掌握这一前沿领域的核心技能。

4.4.1 图片生成

图片生成是将文字描述转化为图像的技术。其原理是人工智能模型解析输入文本的语义，并通过算法生成符合描述的图片。图片生成技术应用广泛，不仅可用于生成普通图片，还可创建 Xmind 格式的思维导图和各类可视化图表，为设计、教育、办公等众多领域提供高效可视化工具。本小节将以普通图片、思维导图和流程图的生成为例，介绍大模型在图片生成方面的能力。

1. 普通图片

在文生图技术中，普通图片是指根据用户输入的文字描述生成的一般性图像，这些图片可以是风景、人物、动物、静物等各种主题和风格的图像。这类图片的生成依赖于大模型对文本的理解，并通过深度学习技术合成符合描述的视觉内容。例如，用户输入“一只在花丛中飞舞的蝴蝶”，模型可以生成一幅色彩鲜艳、栩栩如生的蝴蝶与花朵的画面。文生图技术支持多种风格，从写实到抽象艺术，广泛应用于设计、广告、游戏等领域。表4.2总结了一些较为主流的图片生成大模型。

表 4.2: 主流的图片生成大模型

名称	特点	网址
可灵 AI	支持不同风格高清画作和不同画幅选择	https://app.klingai.com
即梦 AI	生成高分辨率、细节丰富的图像	https://jimeng.jianying.com
海螺 AI	能生成表情丰富的人物或背景	https://hailuoai.com
无界 AI	生成中国式多种风格的图片	https://www.wujieai.com
Midjourney	生成的图像具有独特的艺术风格	https://www.midjourney.com
DALL.E2	生成多样化，多种风格的图像	https://openai.com/dall-e-2
Ideogram	支持多种图像风格和高质量的图片	https://ideogram.ai

下面将以即梦 AI 为例介绍图片生成模型的使用。即梦 AI 是字节跳动推出的一站式 AI 创作平台，集成了 AI 绘画、智能画布、视频生成和故事创作等功能。由于依托字节跳动的短视频平台积累了海量高质量数据，即梦 AI 在生成高质量内容方面具有显著优势。即梦 AI 的界面如图4.60所示，用户登录后可选择“图片生成”进入 AI 作图界面。

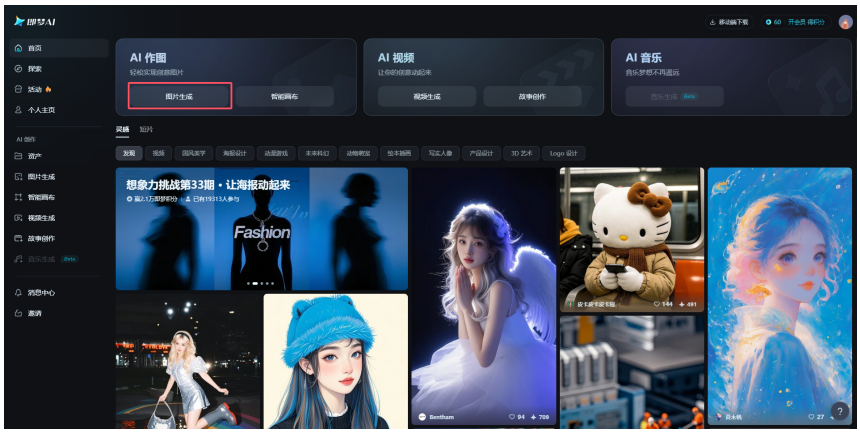


图 4.60: 即梦 AI 主界面

（1）文生图提示词

作图界面左侧设有提示词输入框，用户可以通过输入文字描述引导模型生成符合期望的图片。提示词的选择对最终图像质量和风格有重要影响，用户的输入最好包括主体内容和风格氛围。例如用户想要生成一幅男人和狗的图片，可以输入示例 4.2所示的提示词，提示词中明确了主体（男人和狗），还描述了图片风格（动漫）以及一些画面细节。

示例 4.2

提示词：图片风格为动漫，白衬衫的男人和狗站在草坪上望着云层，从远处拍摄。

即梦 AI 会根据输入的提示词，生成 4 幅相关图片供用户选择，用户可以浏览这些生成的图像，挑选最符合自己需求的图像。生成结果如4.61所示。



(a) 生成的图片 1



(b) 生成的图片 2



(c) 生成的图片 3



(d) 生成的图片 4

图 4.61: 图片生成结果

（2）细节调整

即梦 AI 提供了多种细节调整选项（见图4.62），帮助用户精细化控制生成的图片效果。用户可以导入参考图像，作为风格或内容的参考，还可借助 DeepSeek 改进提示词，提升生成效果。此外，用户也可以选择不同的图片生成模型，调整图片精细度，更改生成图片的比例尺寸等。这些功能使用户能够根据具体需求，对生成的图片进行个性化定制，提升创作的自由度和满意度。



图 4.62: 即梦 AI 图片生成设置

例如，示例 4.2中的提示词经过 DeepSeek 模型优化，可以得到示例 4.3所示的提示词。将图片生成模型调整为图片 2.1，并设置精细度为 7，图片生成结果如图4.63所示。

示例 4.3

提示词：动漫风格，穿白衬衫的男人与金毛犬并肩站立在翠绿草坪，仰头凝视天空中的层积云，远景构图展现开阔空间感，阳光穿透云隙形成丁达尔光效，柔和的青草色与纯白衣衫形成色彩对比



(a) 生成的图片 1



(b) 生成的图片 2



(c) 生成的图片 3



(d) 生成的图片 4

图 4.63: 调整后的图片生成结果

2. 思维导图

思维导图是一种图形化的思维工具，用于组织和呈现信息，促进创造性思维和记忆。它从一个中心主题出发，通过分支向外发散，形成层次分明的结构，每个分支上标有关键词或短句。

目前生成思维导图的方式主要有两种:

- 利用大模型生成提示词 + 思维导图软件: 通过输入主题或已有文段, 利用大模型生成思维导图的标题大纲或内容，再将其转换为思维导图可用的格式, 如 Markdown 或 opml, 最后在思维导图软件（如 Xmind、ProcessOn 等）中手动构建导图。

- AI 自动生成思维导图：目前，一些思维导图网站或工具（如 Xmind、boardmix 等）支持 AI 一键生成思维导图，用户只需输入问题、文章或数据，就能快速得到可视化的思维导图。

本节以 DeepSeek 生成思维导图为例来介绍利用大模型去生成思维导图。

（1）生成思维导图大纲

首先用户需要在 DeepSeek 中提供一个主题或相关文段。若分析文档，可以点“附件上传”上传文件，然后在对话框输入“帮我对以上内容生成思维导图，以 Markdown 格式输出”；若依主题生成，在输入框输入主题，比如“帮我生成一份电商运营思维导图大纲，以 Markdown 格式输出”，如图4.64所示。



图 4.64: DeepSeek 模型生成思维导图大纲（部分输出）

（2）绘制思维导图

将大模型输出的 Markdown 格式的内容保存到“名称.md”的文件中，名称可以根据需要设定，确保文件后缀名一定是“.md”。打开制作思维导图的软件（如 Xmind），选择导入文件（界面如图4.65所示），然后导入 Markdown 格式文件。



图 4.65: 导入 Markdown 格式文件

导入成功后即可生成思维导图，如图4.66所示。

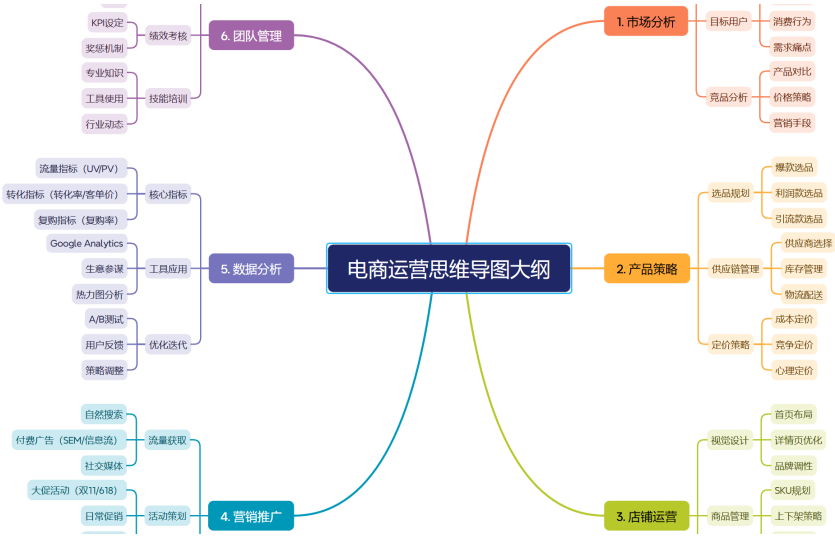


图 4.66: 生成的思维导图

3. 可视化图表

可视化图表是一种通过图形和图像的形式呈现数据的方法，帮助用户直观地理解和分析数据。常见的可视化图表包括柱状图、饼图、折线图、散点图、流程图等，能够将复杂的数据和信息转化为易于理解的视觉内容，广泛应用于数据分析、报告展示以及决策支持等领域。

传统的可视化图表通常需要借助 Excel、Python 等工具，通过手动编程或

者调整参数来创建图表。虽然这些方法灵活，但也需要较高的技术背景和较多的时间投入。相比之下，当前通过大模型生成数据分析的提示词，用户可以只通过简单的文字描述或指令，快速获得所需的图表分析提示和结构，并依托图表生成工具（如 Mermaid[9]等）来实现图表的自动化生成。

本小节以 DeepSeek+Mermaid 为例，介绍如何通过大模型生成数据分析的提示词，并利用 Mermaid 生成可视化图表。

（1）甘特图

甘特图是一种常用于项目管理的图表，主要用于展示项目任务的时间安排及进度。用户通过甘特图能够直观地看到各任务的开始和结束时间，以及任务之间的依赖关系，如图4.67所示。



图 4.67: 甘特图示例

使用 DeepSeek 辅助生成甘特图时，用户首先提供项目的任务列表、各任务时间及依赖关系。如示例 4.4所示，用户可以通过指令“生成一个包含 5 个任务的电商运营项目甘特图，开始时间为 2025 年 4 月 1 日，每个任务项目周期为 1 个月，包含 3 个子任务，以 Mermaid 语法输出”来请求生成相应的甘特图代码。

[9]Mermaid 是一个基于 JavaScript 的图表绘制工具，允许用户通过简单的标记语言生成流程图、序列图、甘特图等多种图表类型。

示例 4.4

提示词：生成一个包含 5 个任务的电商运营项目甘特图，开始时间为 2025 年 4 月 1 日，每个任务项目周期为 1 个月，包含 3 个子任务，以 Mermaid 语法输出。

DeepSeek 模型输出（部分）：

图表 代码 复制 下载

```
gantt
    title 电商运营项目甘特图（2025年4月-8月）
    dateFormat  YYYY-MM-DD
    axisFormat  %m-%d

    section 市场调研
    行业趋势分析         :a1, 2025-04-01, 30d
    竞品数据收集         :a2, after a1, 15d
    用户问卷投放         :a3, after a2, 15d

    section 店铺搭建
    平台注册认证         :b1, 2025-05-01, 10d
    首页视觉设计         :b2, after b1, 15d
    商品上架测试         :b3, after b2, 5d
```

将该段代码复制到 Mermaid 网站[10]的代码框内，Mermaid 将自动渲染成甘特图（见4.68），用户能够直观看到各个任务的时间安排以及任务之间的重叠和依赖关系。在生成甘特图的过程中，用户也可以根据需要在代码中进一步调整细节，满足特定的需求，例如每个任务的开始时间、持续时间和依赖关系。

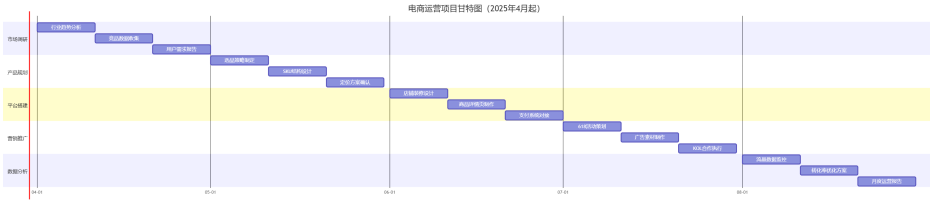


图 4.68: 电商运营项目甘特图

（2）流程图

流程图是一种用于表示工作过程、系统流程或决策过程的图形化工具。它通过不同形状的框和箭头，可以清晰展示操作步骤、决策点及其之间的逻辑关系。流程图广泛应用于工程设计、流程优化、业务分析等领域。

使用 DeepSeek 辅助生成流程图时，用户需要确定流程图的主题，列出流程中的主要步骤以及各个步骤之间的顺序和逻辑关系。此外，如果含有决策步

[10]Mermaid 在线编辑器：`https://mermaid.live`。

骤，决策步骤还需要提供决策条件。如示例 4.5所示，用户可以通过指令“我想生成一个处理订单的流程图，步骤包括接收订单、确认支付、准备货物、发货和订单完成。确认支付需要确认支付是否成功，支付成功后，再准备货物。以 Mermaid 语法输出”来请求生成相应的代码。

示例 4.5

提示词：我想生成一个处理订单的流程图，步骤包括接收订单、确认支付、准备货物、发货和订单完成。确认支付需要确认支付是否成功，支付成功后，再准备货物。以 Mermaid 语法输出。

DeepSeek 模型输出（部分）：

图表 代码 复制 下载

```
flowchart TD
    A[接收订单] --> B{确认支付}
    B -- 支付成功 --> C[准备货物]
    B -- 支付失败 --> E[通知客户支付失败]
    C --> D[发货]
    D --> F[订单完成]
    E -->|客户重新支付| B
```

将该段代码复制到 Mermaid 网站的代码框内，生成的流程图如4.69所示。

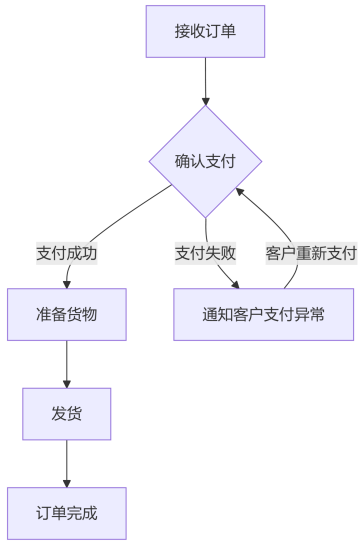


图 4.69: 处理订单流程图

除了甘特图和流程图外，Mermaid 还支持多种其他类型的图表，如序列图、类图、状态图、用户旅程图以及饼图和柱状图等。大模型辅助生成图表的

过程类似，读者可以根据需要自行探索和应用这些图表类型，满足不同的可视化需求。

4.4.2 视频生成

视频生成是利用人工智能技术，根据用户输入生成动态视频内容的过程。通过结合文本、图像和时序数据，大模型能够生成连贯、生动的视频内容。这种技术在影视制作、广告宣传、教育培训等领域具有广泛的应用前景。表4.3总结了一些较为主流的视频生成模型。

表 4.3: 主流的视频生成大模型

名称	特点	网址
可灵 AI	生成高品质的视频，短视频效果好	https://app.klingai.com
即梦 AI	在短视频制作和广告创意方面表现出色	https://jimeng.jianying.com
海螺 AI	将文本信息有效地转化为生动的视频内容	https://hailuoai.com
无界 AI	在广告、创意设计和社交媒体内容表现好	https://www.wujieai.com
Runway ML	支持视频生成、编辑和特效处理	https://runwayml.com
sora	支持高清分辨率视频，视频时长可达 1 分钟	https://openai.com/sora
Synthesia	擅长通过文本生成带有虚拟人物的视频	https://www.synthesia.io

可灵 AI 是由快手团队开发的 AI 创作工具，其视频生成功能强大且灵活，支持文生视频和图生视频。在文生视频方面，用户只需输入文字描述，即可生成符合描述的动态视频；图生视频则支持首位帧和多图生成，用户上传图片后，AI 能根据图片信息生成连贯的视频内容，同时支持添加提示词控制图像运动，生成的视频可调整参数如分辨率、帧率和宽高比等，满足多样创作需求。下面将以可灵 AI 为例介绍视频生成模型的应用。

可灵 AI 主界面如图4.70所示，用户登陆成功后，可以点击“AI 视频”即可打开视频创作界面。AI 视频生成分为两种：文生视频和图生视频。

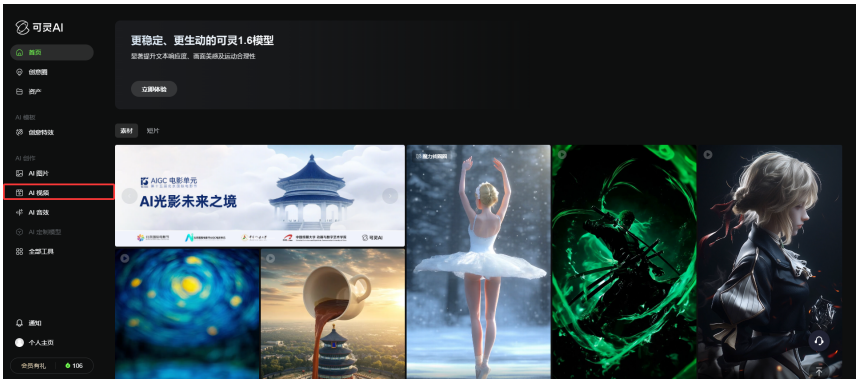


图 4.70: 可灵 AI 主界面

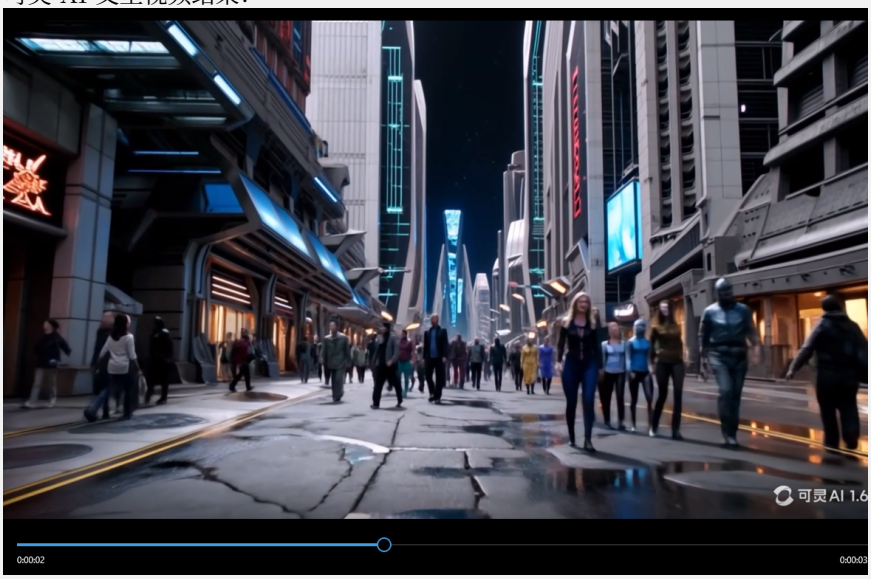
1. 文生视频

可灵 AI 的文生视频功能允许用户通过输入文字描述来生成视频，用户只需要提供相关的提示词，包括主题、物体运动、场景描述等。示例 4.6展示了如何利用可灵 AI 文生视频功能生成内容。

示例 4.6

提示词：外星球上，赛博朋克的城市街景，建筑具有未来感，镜头缓慢向前推进，街道上有行人。

可灵 AI 文生视频结果：



除此之外，可灵 AI 对文生视频也提供了很多细节上的调整。例如，用户可选择不同的视频生成模型、提示词可以通过 DeepSeek 优化，也提供了一些场景和镜头语言的描述，控制模型对不期望内容的控制和创意性的设置等。具体设置见图4.71



图 4.71: 可灵 AI 文生视频生成设置

例如，将示例 4.6中的提示词经过 DeepSeek 模型优化可得到示例 4.7的提示词。调整图片生成模型为可灵 1.5，设置创新性为 0.3，不希望呈现模糊、低质量的内容，可灵 AI 文生视频结果如图4.72所示。

示例 4.7

提示词：未来感十足的赛博朋克城市街景矗立在外星球表面，高耸的金属建筑布满霓虹灯管与全息广告，街道上行人穿着发光服饰缓慢行走，镜头推进展现建筑表面流动的数据流与机械结构，霓虹光效笼罩整个城市，充满科技感的背景音隐约可闻，丰富细节展现异星文明的独特美学。

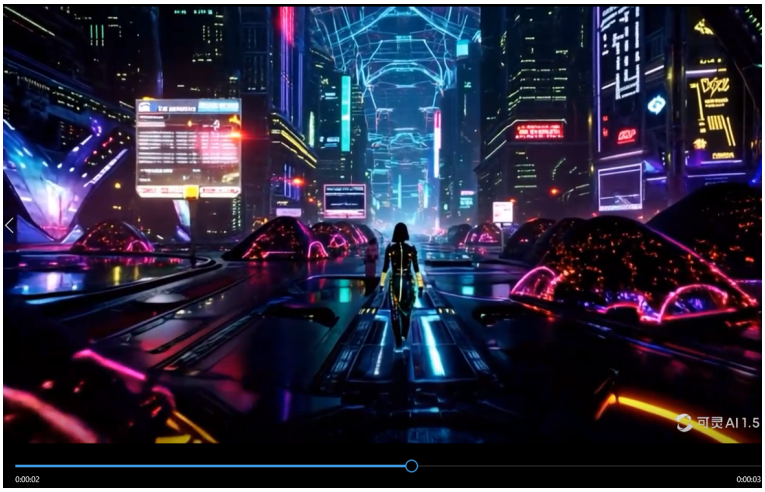


图 4.72: 调整设置后文生视频

2. 图生视频

图生视频功能允许用户通过上传图片来引导模型生成相关的视频内容。用户可以上传一张或多张图片，并提供描述，模型将根据这些图像生成动态的视频序列。图生成视频的提示词最好包括主题运动、背景运动和镜头语言。例如，用户选择图4.73作为首帧图上传，并填写图片创意描述，模型输出效果如示例4.8所示。



图 4.73: 图生视频的首帧图

示例 4.8

图片创意描述（可不填）：一只猫在厨房揉面团。

可灵 AI 图生视频结果：



除此之外，可灵 AI 对于图生视频也提供了很多细节调整选项，有些与文生视频类似，有些则不同。例如，用户可选择不同的视频生成模型、提示词可以选择图片出现在首帧还是尾帧，也提供了视频生成结果的创意性描述等等。具体设置见图4.74

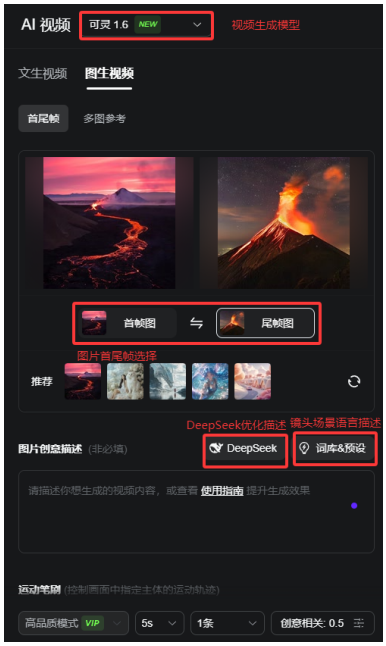


图 4.74: 可灵 AI 图生视频设置

例如，示例 4.9展示了将图4.75作为首尾帧两张图片，并采用镜头场景描述，采用高质量模式，设置创新性为 0.3（不希望呈现模糊、低质量的内容）的结果。



(a) 首帧图



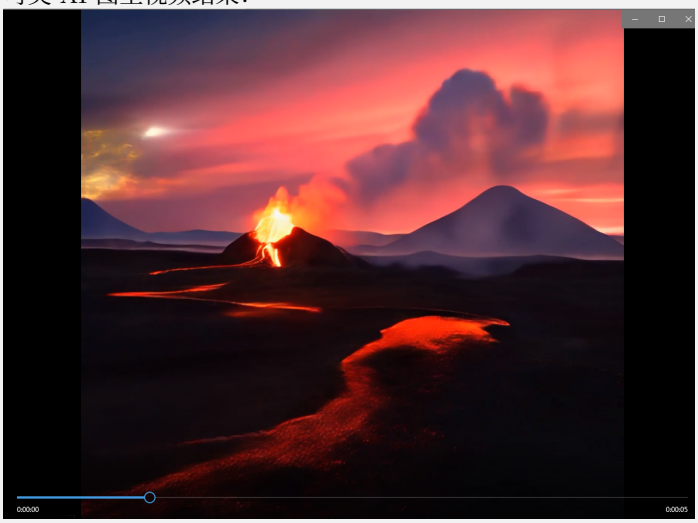
(b) 尾帧图

图 4.75: 图生视频的首帧和尾帧图

示例 4.9

图片创意描述（可不填）：远处的火山下流出一条岩浆，冒着浓烟，蔓延到镜头前，镜头推进，转眼间，火山正在喷发，冒出熊熊烈火。

可灵 AI 图生视频结果：



可灵 AI 对于图生视频还提供了多图参考生成视频的功能。用户可以上传最多四张参考图片，这些图片可以涵盖人物、动物、场景、服饰等各种主题或背景。用户需要提供一段文字描述，阐明这些图片之间的互动或相互关系，系统将根据图片素材生成更加协调和创意的视频内容。例如，以图4.76 为参考，可灵 AI 多图参考生成视频的效果如示例 4.10所示。



(a) 参考图片 1



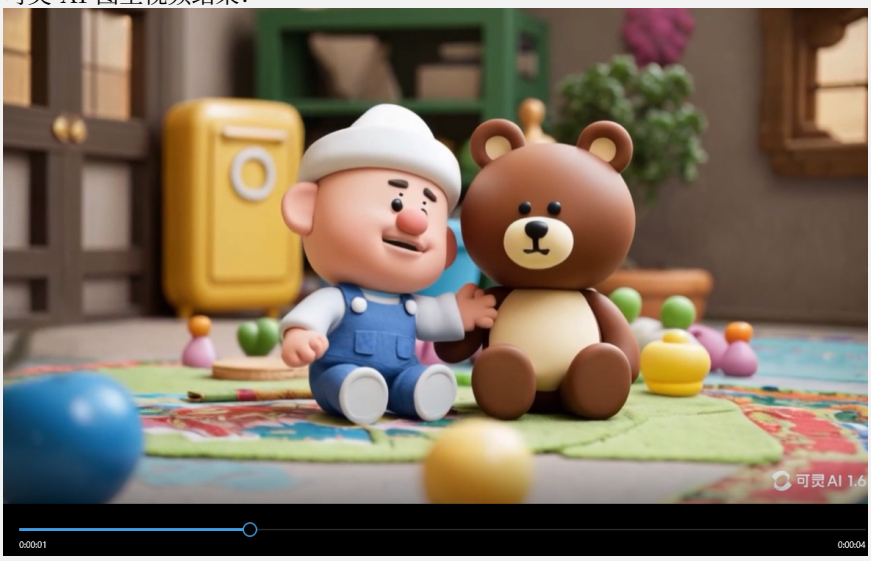
(b) 参考图片 2

图 4.76: 多图参考

示例 4.10

图片创意描述（必填）：带着白色帽子的卡通小人，和棕色小熊，并排坐在一起向镜头打招呼。

可灵 AI 图生视频结果：



4.4.3 音乐生成

音乐生成是利用人工智能技术，根据用户输入的不同形式的内容生成音乐作品的过程。通过结合文本、图片和视频等多模态数据，大模型能够生成符合需求的音乐内容。这种技术在影视配乐、游戏音效等领域具有广泛的应用前景。表4.4总结了一些较为主流的音乐生成大模型。

表 4.4: 主流的音乐生成大模型

名称	特点	网址
天谱乐	支持实时生成音乐，支持多种乐器的模拟	https://www.tianpuyue.cn
天工 SkyMusic	支持多种音乐风格，界面友好	https://www.tiangong.cn/music
海绵音乐	支持情感驱动，自动调整音乐的节奏和风格	https://www.haimian.com
Suno	支持个性化音乐生成，语音合成与音乐结合	https://suno.com
AIVA	擅长生成古典音乐，支持情感驱动	https://www.aiva.ai

天谱乐是趣丸科技唱鸭团队推出的全球首个多模态音乐生成大模型，支持

文本、图片和视频等多种输入方式生成音乐，其主界面如4.77所示。在文生音乐方面，用户输入文字描述后，即可生成最高长达 3.5 分钟的纯音乐作品；图片和视频生音乐则支持上传图片或不超过 60 秒的视频，天谱乐能根据输入素材生成与之高度匹配的带人声唱词的完整歌曲。这种技术在音乐创作、短视频配乐、影视和游戏配乐等领域具有广泛的应用前景。下面将以天谱乐为例介绍音乐生成模型的应用。

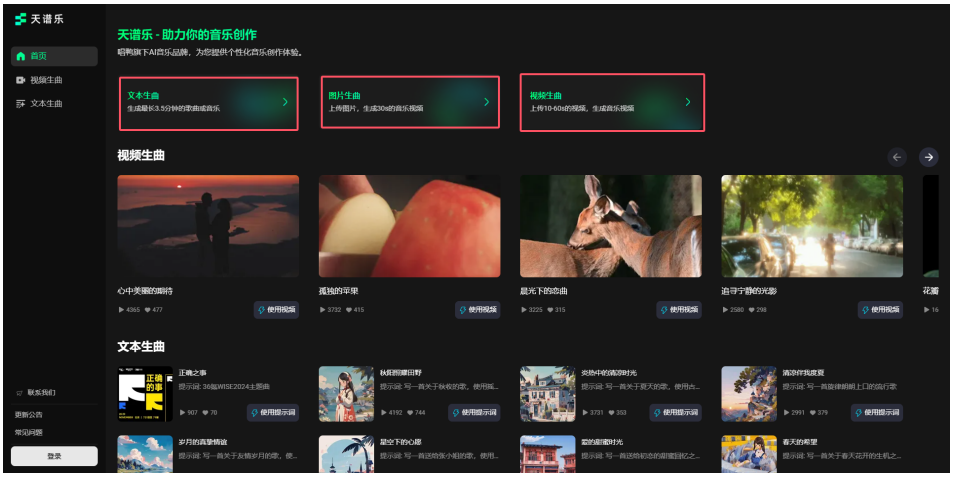


图 4.77: 天谱乐主界面

1. 文生音乐

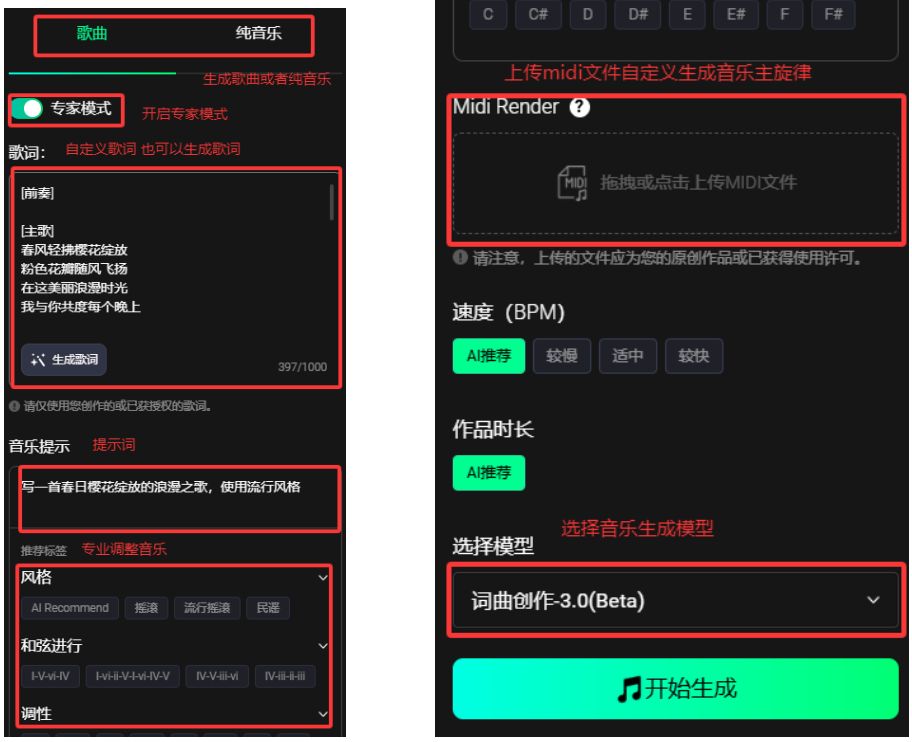
文生音乐允许用户通过输入文本描述来生成音乐。用户进入官网主界面后选择“文本生曲”后可直接输入简洁的描述来引导模型生成期望的音乐。文本生成音乐的提示词最好包括音乐风格和情感氛围。例如，用户想要生成一首樱花主题的歌曲，可以输入提示词，天谱乐的文本生成音乐的生成效果如示例4.11所示（在此只展示生成音乐的界面）。

示例 4.11

提示词：写一首春日樱花绽放的浪漫之歌，使用流行风格。
天谱乐输出：



除此之外，天谱乐对于有专业需求的创作者也提供了专家模式。例如，用户自定义歌词内容、调性、风格，也可以自己上传主旋律文件等等。具体设置见图4.78。



(a) 文本生曲设置 1 (b) 文本生曲设置 2

图 4.78: 文本生曲设置

2. 基于视觉内容的音乐生成

基于视觉内容的音乐生成是指使用图片或视频内容自动生成相应背景音乐。用户进入官网主界面后选择“图片生曲”或者“视频生曲”，无须提供提示词，可直接通过图片或者视频并选择音乐模板来引导模型生成期望的音乐。在天谱乐中，使用图片生成音乐与视频生音乐使用方法一致，这里以图片生成音乐为例进行介绍。图4.80展示了以图4.79为参考的音乐生成效果。



图 4.79: 音乐生成参考图片



图 4.80: 天谱乐图片生曲

4.5 大模型辅助学习

在学习过程中，学生和自学者常常面临信息检索困难、知识梳理不清晰以及作业解答缺乏指导等挑战。传统学习方式往往难以满足个性化需求，而大模型的引入为学习提供了智能化支持，使知识获取更加高效。无论是智能问答、知识梳理，还是作业辅导和个性化推荐，大模型都能提供精准、快速的学习支持，帮助用户快速理解复杂概念，提高学习效率。接下来，本节以豆包爱学为例，介绍大模型在学习场景中的应用及使用方式。

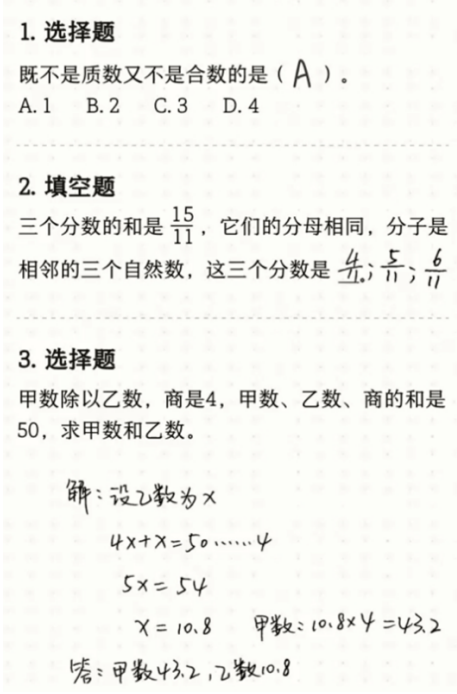
豆包爱学是字节跳动基于“豆包”大模型推出的 AI 教育应用，专为学生提供个性化学习支持。它具备作业批改、拍照答疑、作文写作指导和口语练习等多项智能化功能，帮助用户高效提升学习能力。

（1）作业批改

作业批改功能支持对口算题、练习册和作文等多种作业形式进行快速、准确的批改。如图4.81所示，用户可以通过拍照提交待批改的作业。



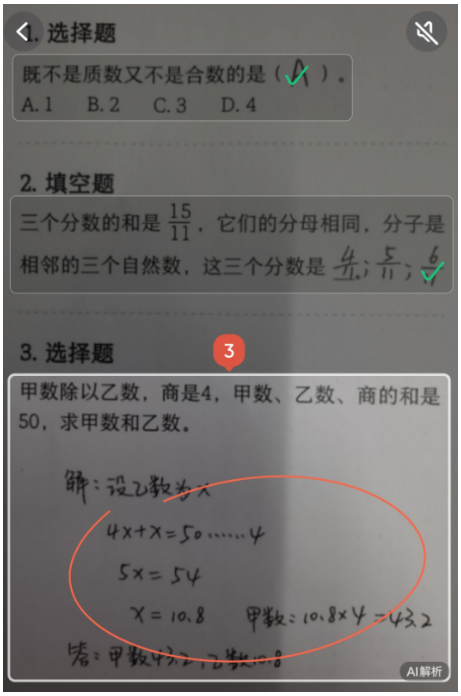
(a) 豆包爱学



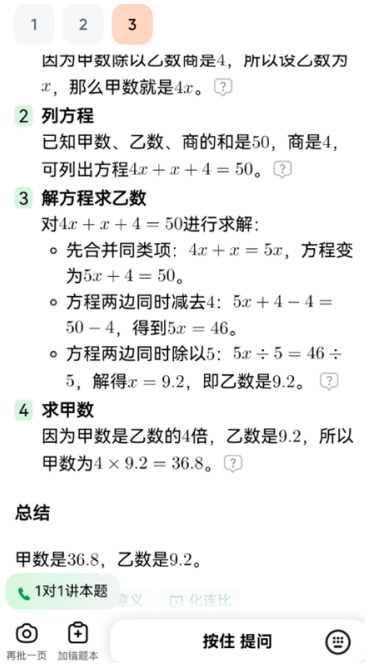
(b) 数学作业

图 4.81: 豆包爱学作业拍照

豆包爱学通过光学字符识别（OCR）技术和 AI 算法，对提交的作业进行智能批改。它不仅能判断答案的正确与否，还会提供详细的解题步骤和解析，帮助学生理解解题思路，如图4.82所示。此外，豆包爱学还会自动收录错题，生成个性化的错题本，方便学生针对性复习。通过精准的错因分析，还能推荐相应的必练题，助力学生巩固知识，提升学习效果。



(a) 作业批改

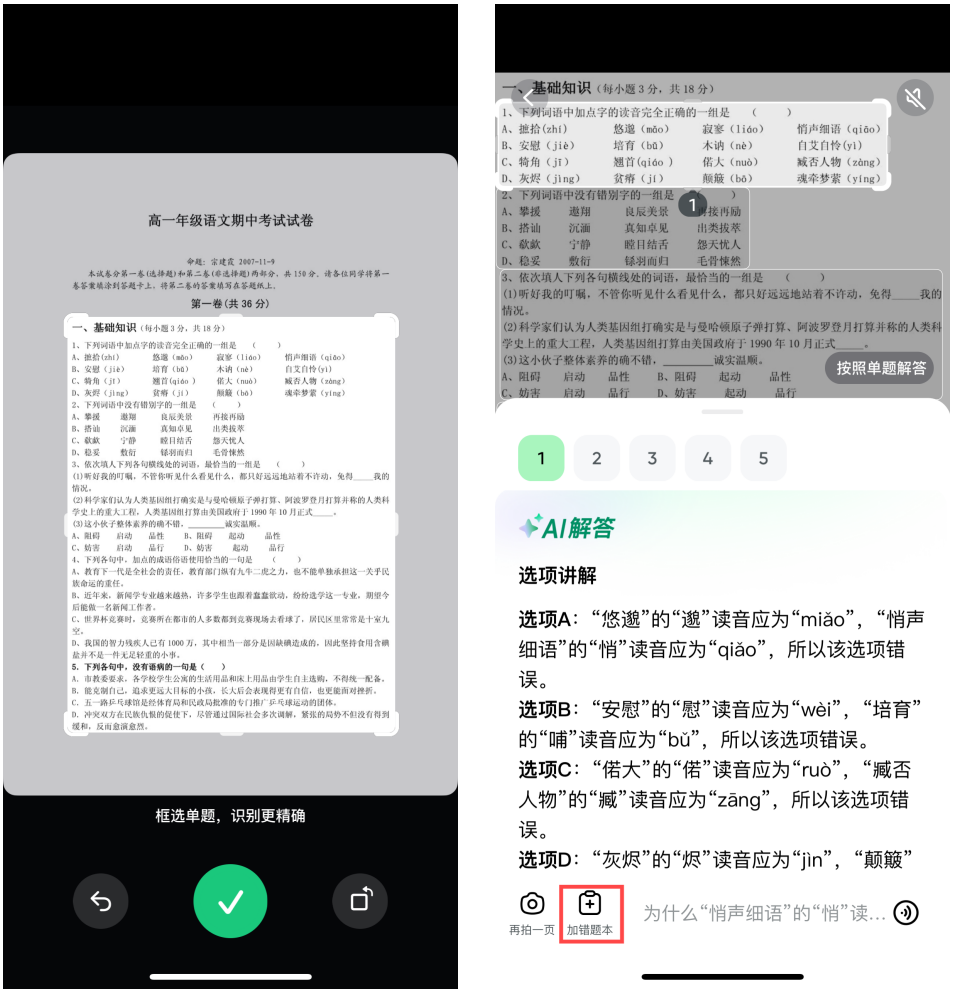


(b) 作业解答

图 4.82: 豆包爱学作业批改

（2）拍照答疑

拍照答疑功能与作业批改类似。用户通过拍照或图片识别输入问题，系统会提供分步骤解析和知识点讲解。如图4.83b所示，用户还可以将易错的题加入错题本。



(a) 拍照答疑 (b) 答案解析

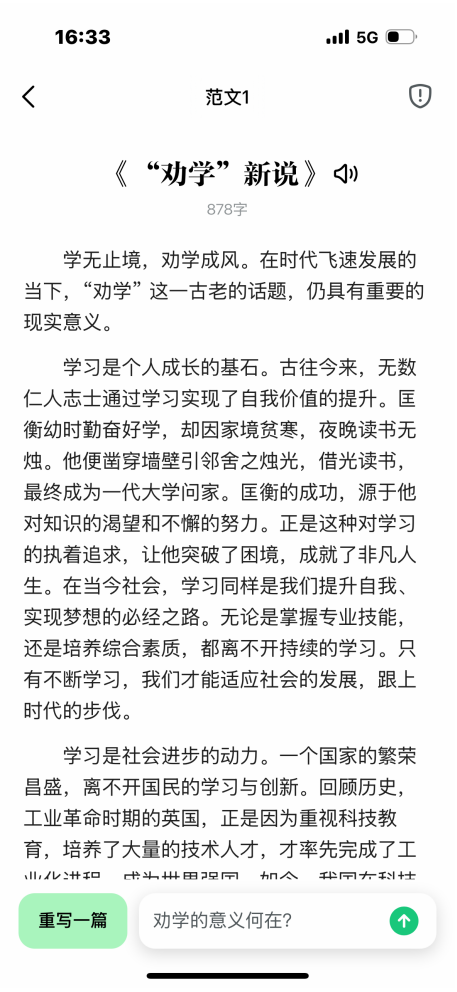
图 4.83: 豆包爱学拍照答疑

（3）AI 作文

除了习题解答外，豆包爱学在作文辅导上也提供了支持。AI 作文功能可以帮助用户提升写作能力。如图4.84所示，根据用户提供的作文题目和要求，它可以提供个性化的写作思路建议，协助生成范文，进行全文润色。此外，该功能支持中英文写作。



(a) 中文作文



(b) 文章生成

图 4.84: 豆包爱学 AI 作文

（4）口语练习

口语练习功能主要针对英语发音、对话和表达训练，帮助用户提升语言应用能力。如图4.85所示，用户可通过朗读单词、句子或短文，AI 即时分析发音准确度（如语调、连读、重音）并打分，并标注错误点，使得用户能够在练习过程中获得发音准确度、语调和流利度等方面的即时反馈。此外，口语练习功能还能进行情景对话模拟，用户可与 AI 角色互动，提升表达流畅度。这一功能尤其适用于需要提升英语口语水平的用户。



(a) 口语练习 (b) 发音练习

图 4.85: 豆包爱学口语练习

4.6 大模型辅助编程

大模型辅助编程是指利用大模型提供代码补全、代码生成、代码优化等功能，帮助开发者提高编程效率，减少编码错误，加速软件开发过程。目前，主流的大模型辅助编程工具主要包括 GitHub Copilot、通义灵码、Tabnine、Cursor 以及文心快码等。这些工具各具特色，开发者可根据自身需求选择合适的 AI 编程助手，提升编程效率和代码质量。表4.5总结了相关模型及其特点。

表 4.5: 主流的大模型辅助编程工具

名称	特点
GitHub Copilot	支持多种主流编程语言，可以根据代码注释自动生成相应的代码实现，还支持自动测试生成
通义灵码	深度适配阿里云生态，方便在阿里云环境下进行开发和部署，支持多种常见的编程语言
Trae	具有全中文界面和交互设计，支持自然语言描述需求生成代码，集成了多个大模型，具有强大的代码理解和生成能力
JetBrains AI Assistant	作为内联操作在特定 IDE 工具窗口中使用，或者通过内置聊天进行交互，通过理解代码的语义和上下文
Tabnine	支持本地模型部署，可避免敏感代码外泄，能够根据开发者的代码风格和习惯进行个性化学习，对硬件资源要求相对较低
CodeWhisperer	深度集成 AWS 生态，为使用 AWS 服务的开发者提供优化的代码建议
Cursor	具备智能代码补全、多文件编辑等功能，支持复杂的代码结构和逻辑关系理解

通义灵码是阿里云推出的一款智能代码生成和辅助工具，具备强大的自然语言处理能力，能够根据用户描述自动生成高质量的代码片段，支持多种编程语言，并能高效解决代码、指令和工程问题，帮助开发者快速完成复杂任务。此外，它是一款可免费使用的大模型编程辅助工具，支持集成至各种开发环境和 IDE 插件，提供便捷的使用体验。因此，本节以通义灵码为例，介绍大模型辅助编程工具的主要应用与优势。

4.6.1 安装与注册

通义灵码的安装和注册流程相对简单，用户可以通过官方网站[11]或 VS Code 扩展商店下载并安装插件。在 Visual Studio Code 中找到左下角的管理，选择“扩展”，在搜索栏输入“TONGYI Lingma”就可以找到通义灵码的插件，选择安装即可完成通义灵码插件的安装, 如图4.86所示。

[11]通义灵码官方网址：`https://lingma.aliyun.com/`

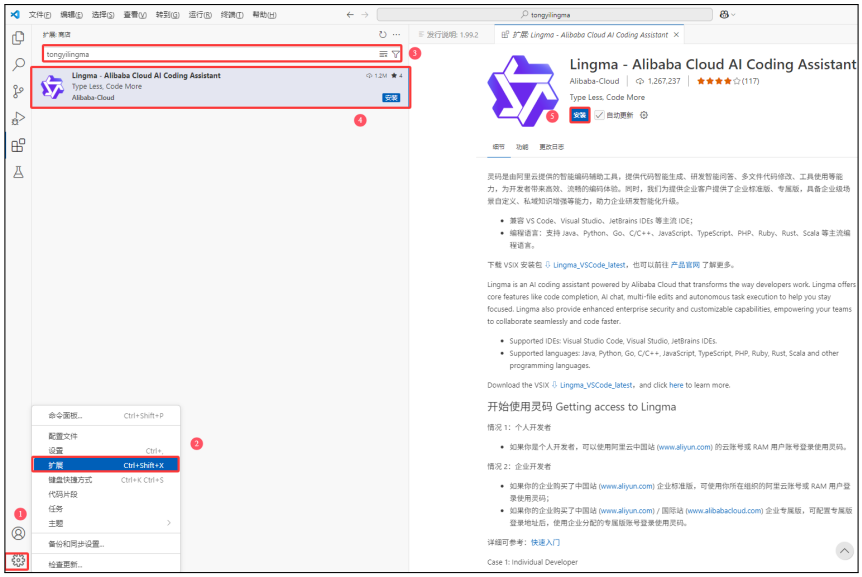


图 4.86: 通义灵码的下载

安装成功后，在右下角会有通义灵码的图标，未登录会提示登录后才能使用，点击登录会跳转到登录界面，如图4.87所示。

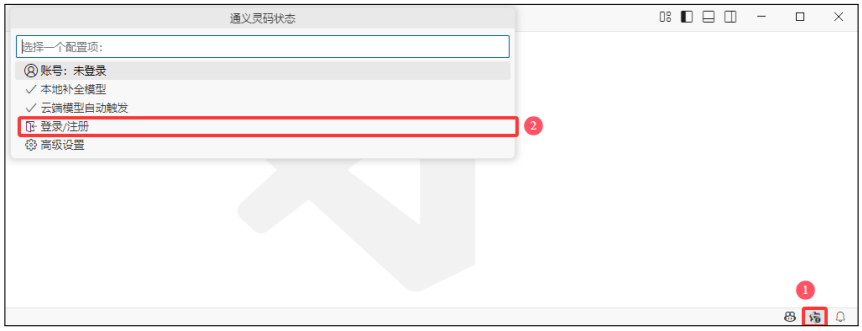


图 4.87: 通义灵码的登录

4.6.2 功能介绍

通义灵码具备多项智能辅助功能，能够显著提升开发效率和代码质量，其中包括代码智能补全、AI 智能问答、智能生成指令和问题修复建议等。

1. 代码智能补全

当在编辑器区进行代码编写时，在开启自动云端生成的模式下，通义灵码会根据当前代码文件及相关代码文件的上下文，自动生成行级/函数级的代码建议，此时可以使用快捷键采纳、废弃，或查看不同的代码建议，其中代码建议相关的快捷键如下表4.6所示。

表 4.6: 通义灵码中的快捷键

操作	macOS	Windows
接受行间代码建议	Tab	Tab
废弃行间代码建议	esc	esc
查看上一个行间推荐结果	\(option)+[	Alt+[
查看下一个行间推荐结果	\(option)+]	Alt+]
手动触发行间代码建议	\(option)+P	Alt+P

（1）自然语言生成代码

通过自然语言描述需要实现的功能，自动生成相应的代码片段，如图4.88所示。



图 4.88: 使用自然语言生成代码

（2）行级/函数级实时补全

根据当前代码文件及相关代码文件的上下文，自动生成行级/函数级的代码建议，如图4.89所示。

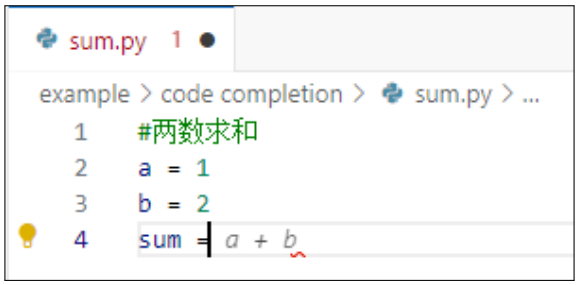


图 4.89: 实时补全

2.AI 智能问答

通义灵码作为一个智能编码助手，当编码遇到困难或者问题的时候可以向它询问，与其互动，通义灵码提供了多种互动方式，包括自由问答和代码问答。

（1）自由问答

通过询问通义灵码智能问答助手，即可快速获得答案和解决思路，可通过Ctrl+Shift+L 快捷键唤起通义灵码助手，也可以通过单击左侧工具导航栏唤起，如图4.90所示。

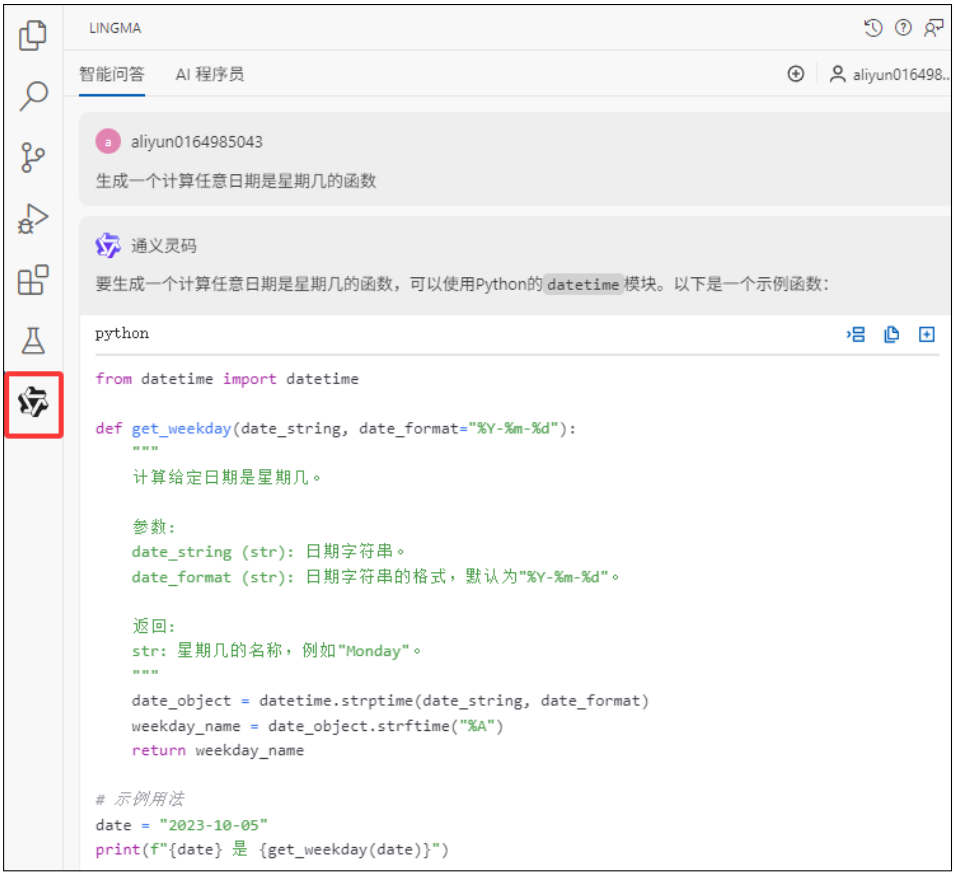


图 4.90: 自由问答

（2）代码问答

选中代码后可在智能问答窗口的输入框中输入问题，通义灵码将围绕选中代码开展对话，如图4.91所示。



图 4.91: 根据选中代码问答

@workspace: 结合当前工程生成简单需求或缺陷的整体修改建议和相关建议代码，如图4.92所示。



图 4.92: 关于工程的问答

@terminal：主要用于解答对于终端指令的问题，通过自然语言描述终端指令诉求，通义灵码将可以生成需要的终端命令，也可以输入终端指令让通义灵码解释，如图4.93所示。



图 4.93: 关于终端的问答

3. 智能生成指令

通义灵码中内置了多个智能生成指令，这里主要介绍解释代码、生成注释和代码优化这些使用较多的功能。

（1）指令一：解释代码

覆盖各种编程语言，选中代码后可自动识别编程语言并生成代码解释，如图4.94所示

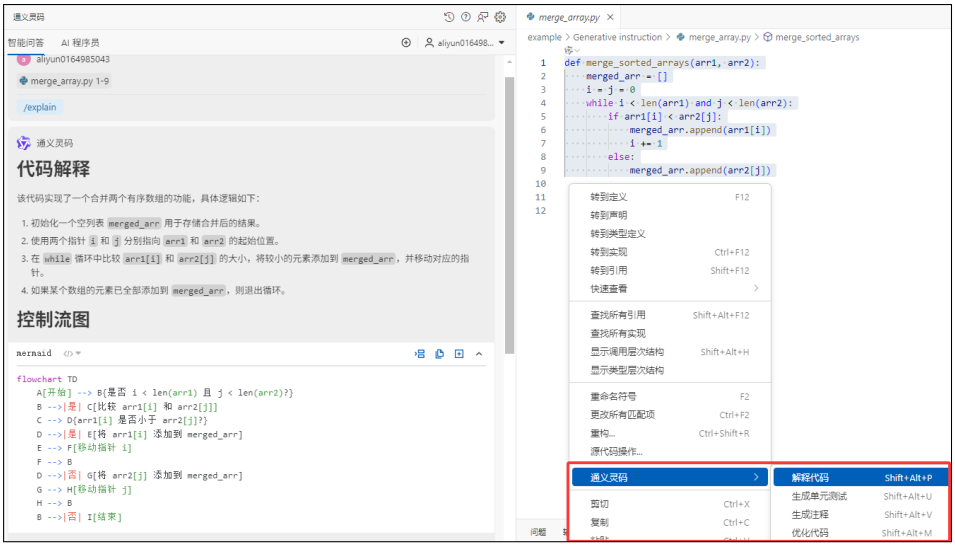


图 4.94: 根据选中代码进行解释

（2）指令二：生成注释

一键生成方法注释及行间注释，提升代码可读性，如图4.95所示。

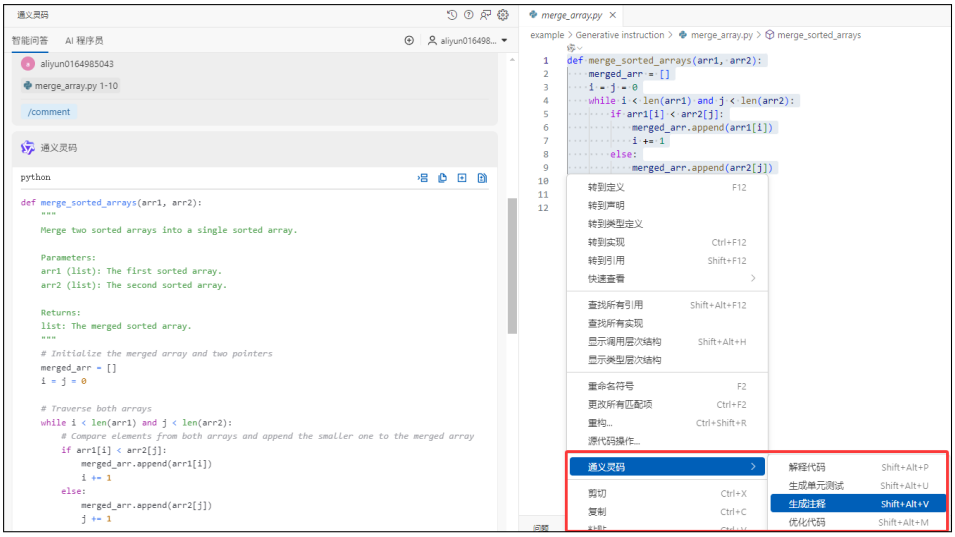


图 4.95: 生成方法注释及行间注释

（3）指令三：代码优化

分析代码及其上下文，识别潜在的编码问题，如图4.96所示。

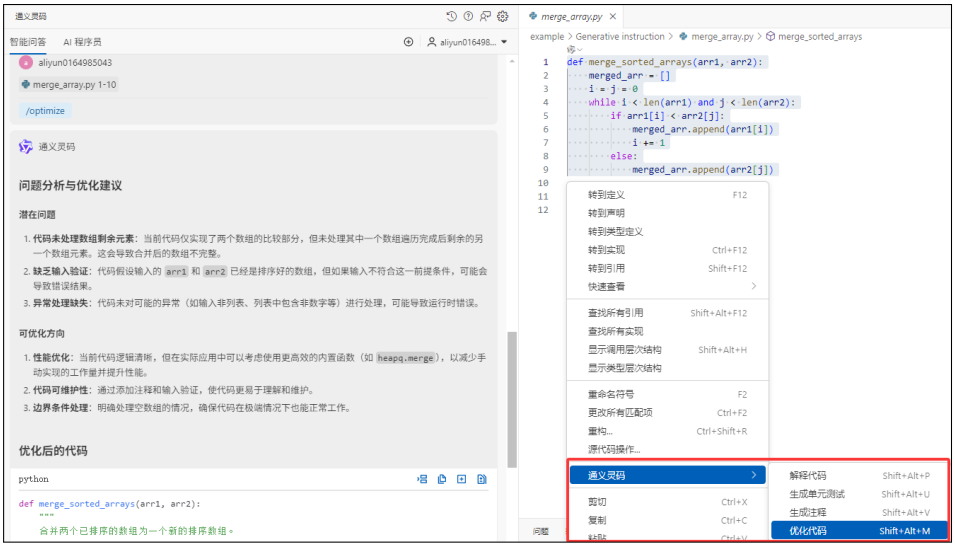


图 4.96: 分析代码问题并进行优化

4. 问题修复建议

对于代码中出现的代码问题，通义灵码会给出相应的修复建议；对于运行出现的报错异常，通义灵码也会给出排查思路或修复建议。

（1）代码问题修复

当在编码过程中出现代码问题需要进行修复时，通义灵码可以给出修复建议，快速修复代码问题，如图4.97所示。

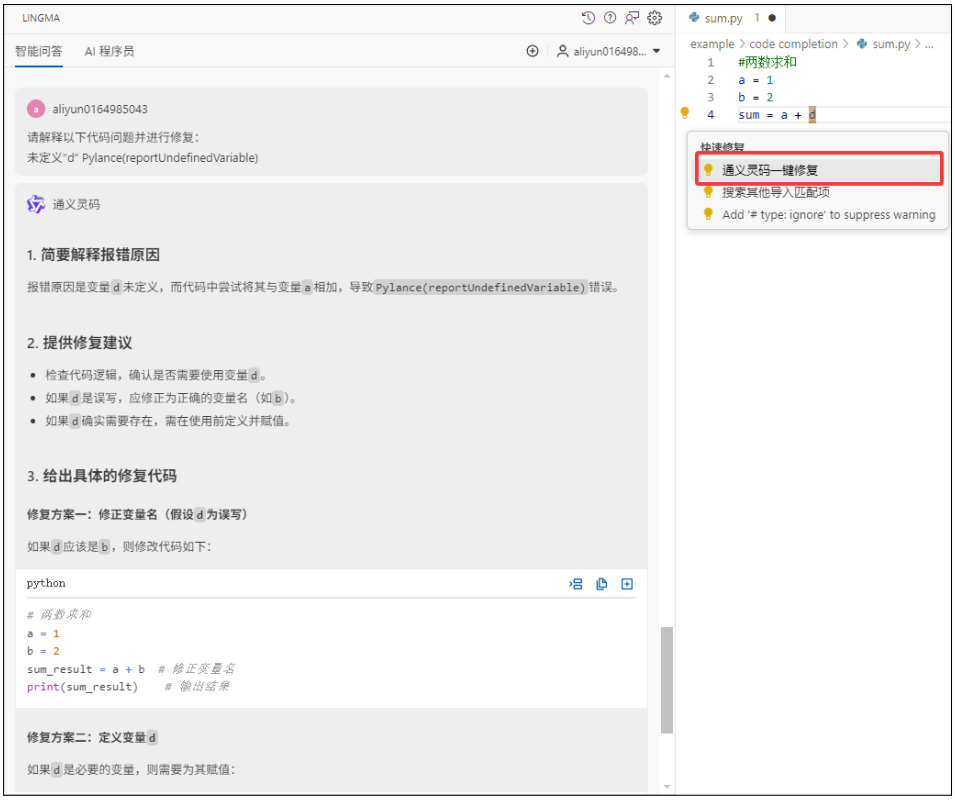


图 4.97: 代码问题修复

（2）异常报错排查 (IntelliJ IDEA) 在 IntelliJ IDEA 中，当运行出现异常报错时，在运行标准输出窗口中，即可看到通义灵码的快捷操作按钮，单击后，通义灵码将结合运行代码、异常堆栈等报错上下文，快速给出排查思路或修复建议，如图4.98和4.99所示。

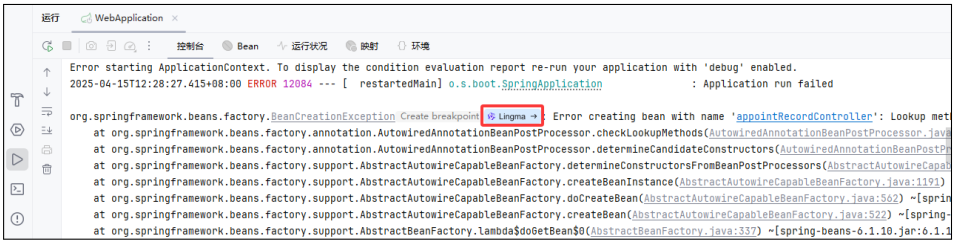


图 4.98: 运行报错

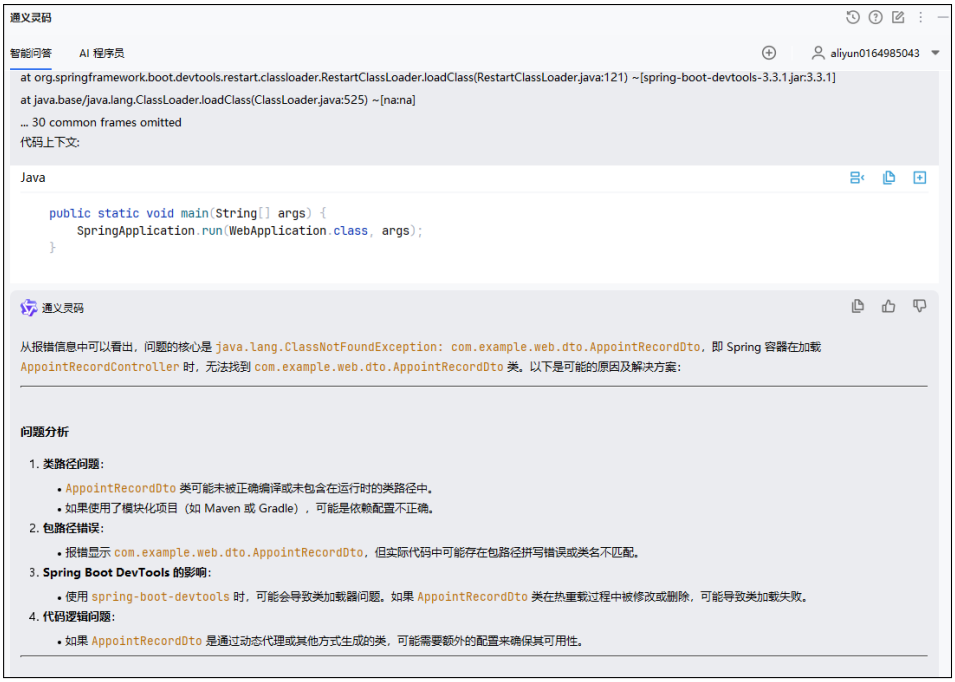


图 4.99: 通义灵码报错排查

4.7 综合案例

前几节分别探讨了大模型在信息检索、文本办公、内容创作、学习与编程等典型任务中的应用。但在实际工作中，用户往往面临更加复杂、多步骤、跨领域的综合性任务，这类任务对模型的理解、推理与规划能力提出了更高要求。以短视频制作为例，一个完整的流程可能包括选题策划、脚本撰写、视觉元素设计、配音生成、镜头分镜规划及文案发布等多个子任务。本节以短视频制作为案例，展示大模型如何在多环节任务中实现智能协同，助力用户高效完成复杂工作目标。

短视频是指时长较短、内容精炼的视频形式，通常在几秒到几分钟之间，广泛应用于社交平台、内容营销和日常表达等多个场景。它以轻量、直观、传播快的特点，成为当前最受欢迎的内容类型之一。过去，制作一条质量较高的短视频往往需要多个创作环节协同完成，包括画面设计、视频剪辑、配乐与字幕生成等，门槛较高、周期较长。而如今，生成式 AI 的发展正不断重塑短视频的创作流程。借助如即梦 AI、可灵 AI、天谱乐等大模型工具，创作者可以快速生成高质量的图片、动态画面与配套音乐，显著提升内容生产效率，降低

创作门槛，同时增强短视频的个性化表达与视觉创意。

接下来将以《英雄联盟》中的经典角色“金克丝”为主角，展示一个完整的短视频创作流程。该案例将详细呈现如何借助多个 AI 生成模型，依次完成从场景构思、素材生成、动态构建、音效匹配到最终视频合成的全过程。

1. 场景构思

短视频创作的第一步是确定清晰的主题与场景背景。这是决定整体叙事节奏、视觉风格和氛围基调的关键步骤。好的构思不仅要围绕主角展开，更要在短时间内展现出明确的情境和情感张力。

《英雄联盟》中的经典角色金克丝具有叛逆、疯狂且极富动感的性格特征，因此在场景设定上，应突出其混乱美学与强烈视觉冲击。例如，金克丝漫步在一个繁华且美丽的街道上、金克丝在烈火中拿着武器或者金克丝以一个屠龙者的姿态面对巨龙等等场景。除此之外，创作者也可以借助大语言模型进行思维发散，通过提示词生成多样化的创意设定与场景描述。例如，输入如下提示词：

示例 4.12

提示词：请帮我设计一个金克丝为主角的短视频剧情，她出现在一个赛博朋克风格的城市中，内容应包含她引发一场爆炸并在混乱中逃脱的场景，强调视觉冲击与动感。

2. 素材生成

在利用可灵 AI 生成高质量视频之前，需要获取高质量的场景图片和人物图片作为基础素材。对于本案例中的主角金克丝，其人物素材在网络上资源丰富，获取相对容易。相比之下，场景图需要根据具体设定进行定制化生成，确保整体风格一致、氛围契合。因此，创作者可以借助即梦 AI，根据前一节中构思的赛博朋克风格和剧情设定，生成与金克丝角色形象相匹配的高质量场景图。这些图片将作为后续视频合成的视觉基础，为内容创作提供统一的美术风格支撑。示例 4.13展示了利用即梦 AI 生成一张背景匹配的高质量图片（见图4.100）。画面风格与金克丝的叛逆个性和混乱氛围较为契合。

示例 4.13

提示词：可爱卡通，C4D，设荧光，朦胧光，热闹的商店街，暗色暖色调，光影对比，高光，轮廓清晰，红衣女孩红色高达，特写近景照。



图 4.100: 即梦 AI 生成高质量图片

一个视频效果可以参考多个场景图片，一个场景图片也可以生成多个视频效果。利用图片生成模型，创作者根据创作需要生成一定数量的场景图片。创作者也可以利用即梦 AI 其他图片生成功能生成多样化的高质量图片。

3. 动态构建

在获取即梦 AI 生成的高质量场景图和金克丝人物图后，可以利用可灵 AI 的多图参考功能来生成视频。通过上传这些图片，并添加适当的创意描述，可以让金克丝在不同场景中与环境进行互动。以图4.101 为参考，可灵 AI 多图参考生成视频的效果如示例 4.14所示。



(a) 参考图片 1



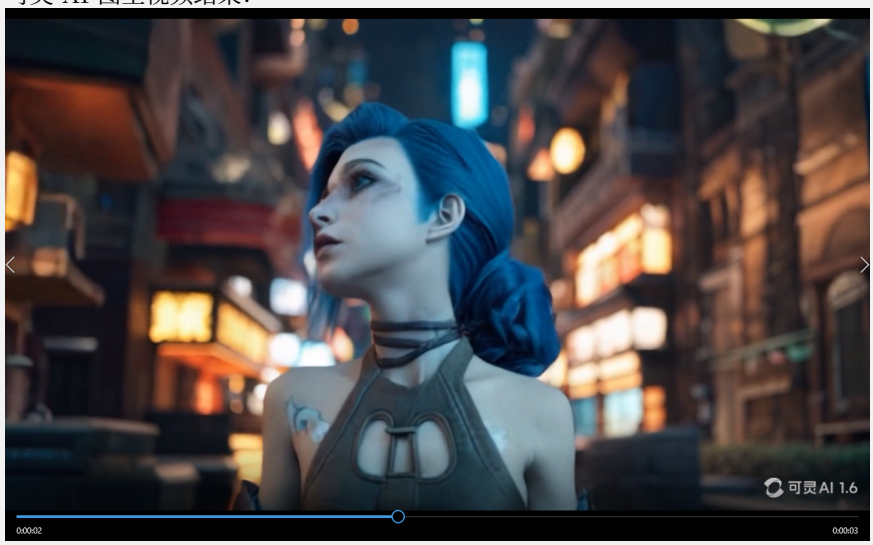
(b) 参考图片 2

图 4.101: 多图参考

示例 4.14

图片创意描述（必填）：第一张中的人物代替第二张图片的中主人公，行走在第二张图片的背景上，她一边往前走一边向左看，过一会又向右看，再过一会看向镜头。镜头设置跟随。

可灵 AI 图生视频结果:



尝试不同的场景图, 通过多次使用可灵 AI 的视频生成功能，可以得到多个视频片段。创作者也可尝试使用可灵 AI 的其他视频生成的功能来创造多样化的视频。

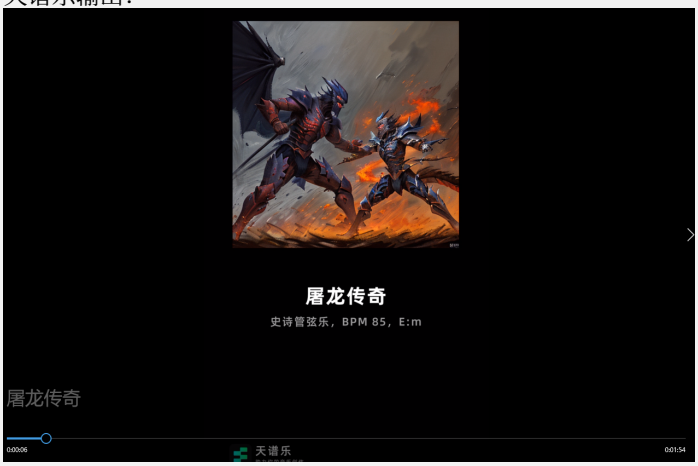
4. 音效匹配

在短视频创作中，音效与音乐不仅起到烘托氛围的作用，还能增强观众的沉浸感与情绪共鸣。根据视频的节奏、画面风格以及金克丝这一角色本身的调性（如叛逆、狂野、节奏感强），创作者可以选择匹配的音乐风格，例如电子摇滚、朋克或快节奏电子乐等。

创作者可以借助天谱乐等文本生成音乐工具，只需输入简要的情绪描述或场景提示词，即可自动生成节奏与风格高度匹配的背景音乐。示例 4.15展示了利用天谱乐的文本生成音乐的效果。

示例 4.15

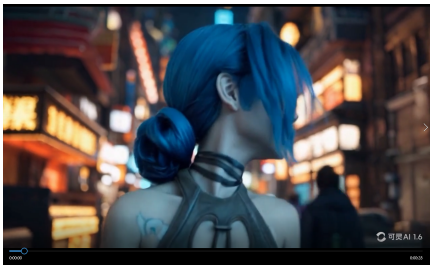
提示词：写一首背景宏大，音乐深沉，激昂关于屠龙勇士的纯音乐。
天谱乐输出：



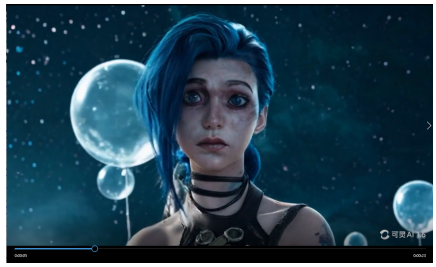
创作者可以生成多个音乐风格选择最合适的音乐，有能力的创作者的也可以将多个音乐融入到一起。创作者也可以直接使用天谱乐中的其他音乐生成的功能来创造多样化的音乐。

5. 最终合成

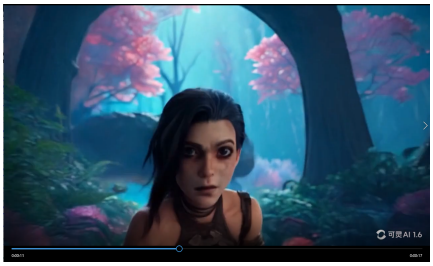
在完成场景图、人物素材、动态视频片段与背景音乐的准备后，最后一步是将所有内容整合为完整的视频作品。创作者可以使用专业的剪辑软件将这些视频片段进行合并，并在剪辑过程中删减掉低质量或不符合预期的内容，同时，将生成的音乐添加到视频上，经过细致的音画同步调整和特效处理，最终可以得到一个流畅、连贯且富有创意的短视频。受限于图文展示，图4.102展示了本案例短视频的关键画面效果。



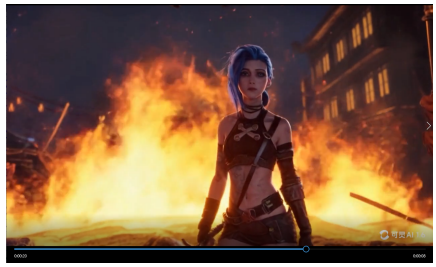
(a) 短视频效果 1



(b) 短视频效果 2



(c) 短视频效果 3



(d) 短视频效果 4

图 4.102: 短视频效果

4.8 讨论

讨论 4.1. 请讨论大模型辅助工作与就业的关系。

讨论 4.2. 请讨论由生成式模型创作的内容的版权，应该归属用户还是模型供应商?

4.9 习题

习题 4.1. 请比较大模型与传统搜索引擎在信息检索方面的异同，讨论其各自的优势和局限性。

习题 4.2. 请阐述 OfficeAI 工具在处理 Word 文档时的功能，并选取 3 个不同功能介绍。

习题 4.3. 请阐述 ChatPDF 工具在处理 PDF 文档时的功能，并选取 3 个不同功能介绍。

习题 4.4. 请阐述 ChatExcel 工具在处理 Excel 文档时的功能，并选取 3 个不同功能介绍。

习题 4.5. 请阐释 DeepSeek+Kimi 一键生成 PPT 的步骤。

习题 4.6. 请讨论在集成大模型的办公工具中可能遇到的挑战，如数据隐私和安全性问题。

习题 4.7. 请使用不同的图片生成模型，比较它们生成图片的质量和效果。

习题 4.8. 请使用不同的视频生成模型，比较它们生成视频的质量和效果。

习题 4.9. 请使用不同的音乐生成模型，比较它们生成音乐的质量和效果。

第二部分

大模型高阶应用：关键技术

第五章　大模型检索增强生成

5.1　概述

第四章介绍了 LLM 在实际应用中的表现，其通过语义理解与知识推理能力，在辅助办公、内容创作、个性化学习支持等方面展现出显著优势。然而，当前大模型仍面临因知识更新滞后带来的时效性局限、垂直领域知识密度不足导致专业性缺失，以及生成结果的事实一致性难以校验等问题。

检索增强生成（Retrieval Augmented Generation, RAG）技术通过 “先检索相关知识再生成回答” 的架构设计，为应对大模型的时效性、专业性等局限提供了可行的解决方案。同时，RAG 通过明确的溯源机制，将生成内容与检索到的参考信息直接关联，增强了模型输出的透明性。该机制不仅提高了用户对生成结果的信任度，还进一步提升了 LLM 在法律、金融和医疗等风险敏感场景中应用的可靠性。

RAG 的概念最早于 2020 年在人工智能与机器学习领域国际顶会 NeurIPS 收录的一篇名为 “Retrieval-augmented generation for knowledge-intensive NLP tasks”（知识密集型自然语言处理任务的检索增强生成）的论文中被提出[29]。在该文中，RAG 的流程主要包括三个阶段：1）首先对非结构化[1]的维基百科数据进行预处理；2）其次根据用户问题检索得到语义相关的背景知识；3）最后将其作为提示词上下文用于增强模型的输出。在本书中，我们将这一阶段的 RAG 称为经典检索增强生成，即主要通过构建非结构化文档的向量数据库来实现的 RAG。除非结构化文档外，大模型还可以依赖包括搜索引擎、知识图谱和关系数据库等外部知识。鉴于此，本书将不局限于传统非结构化文档的 RAG 称为广义检索增强生成，即 RAG 作为一种人工智能框架将传统信息检索系统（如搜索引擎和数据库）的优势与生成式 LLM 的能力相结合，该定义

[1]非结构化数据缺乏定义，没有预定义字段或格式，常见的文档、图像、音频和视频等均是非结构化数据。与之相较，结构化数据则采用预定义的预期格式，具有明确的字段和数据类型，比如关系数据库中的表。

和谷歌云对 RAG 的定义类似[2]。

本章将首先介绍经典检索增强生成，并围绕向量数据库重点讨论数据分块和向量搜索等关键技术。然后，本章将进一步介绍广义检索增强生成，阐述如何在 RAG 应用中有机地结合搜索引擎、知识图谱和关系数据库。通过对本章的学习，读者将能够系统掌握 RAG 的基本原理与关键实现技术，理解其在提升 LLM 可靠性、专业性和可控性方面的重要作用。同时，读者还将具备基于不同外部知识源设计与构建 RAG 系统的能力，为后续在实际业务场景中部署高效、可信的生成式人工智能系统奠定坚实基础。

5.2 经典检索增强生成

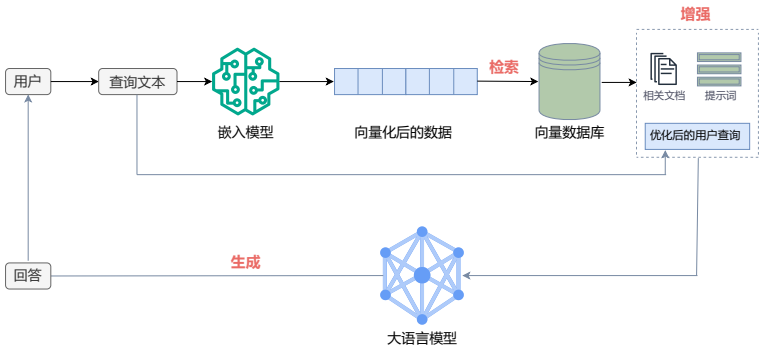


图 5.1: 经典检索增强生成示意图

图5.1展示了经典检索增强生成的架构，即用户的查询文本经过嵌入模型（Embedding Model）[3]得到向量化的数据后，在向量数据库中进行相似性查询并返回若干相关信息（即“检索”）[4]；接着，综合利用知识库中得到的相关数据并注入提示词上下文（即“增强”）；最后，大模型通过上述提示词产生回复（即“生成”）。不难发现，RAG 的“增强”和“生成”步骤与常规大模型应用并无差异，s 因此，本节将聚焦于“检索”过程，并重点回答下面三个关键问题：1）查询文本如何被编码成向量？（第5.2.1节）；2）原始文档在向量数据库中如何存储？（第5.2.2节）；3）如何在向量数据库中进行相似性检索？（第5.2.3节）

[2]参见https://cloud.google.com/use-cases/retrieval-augmented-generation。

[3]自 Word2Vec 技术[30]提出以来，基于深度神经网络的语义表示方法通过将离散文本映射为稠密向量空间中的连续表示，实现了语义关系的几何化度量。

[4]由于大模型上下文的限制，一般还需要对搜索结果进行后处理，包括精化（Refiner）和重排序（Reranker）等；此外，输入文本也往往需要经过查询重写（Rewriter）。为了便于读者理解，图中并未体现上述过程。

5.2.1 文本嵌入表示

传统的检索系统往往依赖于关键词匹配，无法有效地利用其语义信息。例如，当用户输入“如何让手机电池更耐用?”的检索信息时，传统检索可能仅匹配“手机”“电池”“耐用”等关键词，从而返回广告或技术科普等无关内容。相较于传统检索，语义检索能够识别检索信息的真正需求，即“延长电池寿命”，从而关联“续航时间”“充电习惯”等隐性需求，排除“更换电池教程”等歧义内容。语义检索实现了检索范式的转变，使检索系统从基于关键词的匹配跃迁至需求理解，实现了用户场景适配的智能匹配。

在语义检索中，如何表示文本的语义信息至关重要，其不仅是语义检索的基础技术，也是实现精准匹配的前提条件。文本嵌入表示可将查询文本转换为稠密向量空间中的数值表示，这种向量化编码能够有效保留文本的语义特征。

上述将数据（如文本、图像等）转化为连续数值向量的编码过程通常被称为嵌入。根据输入数据的格式不同，可将嵌入分为文本嵌入和图片嵌入等，用于生成嵌入表示的模型被称为嵌入模型，本节聚焦于讨论文本嵌入模型。表5.1给出了文本嵌入表示的直观表达示例。例如，在 OpenAI 的 `text-embedding-3-small` 模型中，“计算机”、“电脑”与“教师”的嵌入表示即可表达为如表5.1所示的连续数值向量。

表 5.1: 文本及其嵌入表示

文本	向量
计算机	[0.01357127912342548, 0.010388375259935856, ...]
电脑	[0.00580993015319109, 0.007956036366522312, ...]
教师	[0.00580993015319109, 0.007956036366522312, ...]

表5.2展示了部分常用文本嵌入模型，其中 `text-embedding-3-small` 和 `voyage-3` 分别是 OpenAI 和 Claude 的闭源模型，其语义表达能力相较其他模型而言更强。不过，不同模型在不同语言、不同任务等方面表现均有差别，用户应根据实际任务特性并综合考虑成本等因素后，选择更为合适的模型。

表 5.2: 常用文本嵌入模型

模型名称	维度	类型
GloVe	100	开源
FastText	300	开源
all-MiniLM-L6-v2	384	开源
text-embedding-3-small	1536	闭源
voyage-3	1024	闭源

当文本被编码为数百维甚至上千维的高维向量后，每个维度的数值通常不具备可解释的语义含义。然而，这些文本的语义关系及语言学规律可通过向量运算得到。早在 2003 年，研究人员就发现词类比现象[30]，比如“国王”的嵌入减去“王后”的嵌入大致等于“男人”的嵌入减去“女人”的嵌入。一般而言，可通过向量的距离表示其语义相似性，其中，最常见的距离度量方法是余弦距离。两个向量的余弦距离即是两个向量的夹角度数，夹角度数越小则两个向量的相似度越大。图5.2是二维向量的余弦距离示意图。例如，若以“计算机”作为查询向量，那么它到“电脑”和“教师”的余弦相似度分别约为 0.74 和 0.30，由此可以认为“计算机”与“电脑”更加相似，该结果也符合其真实语义关系。

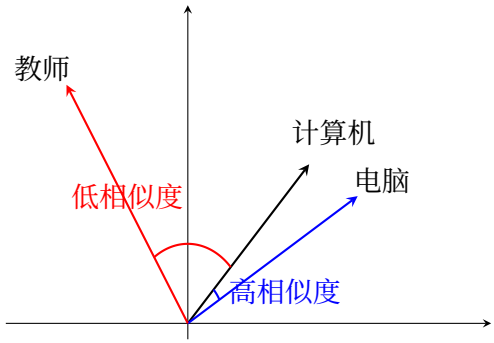


图 5.2: 二维向量下余弦距离的示意图

5.2.2 文档分块存储

无论是查询文本还是原始文档，均可通过文本嵌入模型生成对应的向量编码[5]。但目前主流的文本嵌入模型对输入上下文长度均有限制，一般无法将很大的文档直接进行嵌入操作。事实上，即使不考虑上下文长度的限制，若检索步骤直接返回整个文档，文档中的大量冗余信息也会干扰大模型的分析和推理。为应对上述问题，需将原始文档以分块形式存储在向量数据库中。合适的数据分块技术能够以不同粒度对原始文档进行有效拆分，这是构建向量数据库的重要前提，也是确保数据检索效率的基础。下面介绍几种常见的数据分块（Chunk）技术。

固定大小分块 固定大小分块是最基础的分块策略，其核心原理是按照预设的字符长度或分隔符对文本进行机械切割，无视文本内部的语义结构。关键参数设置涵盖块长度、重叠区长度以及分隔符。该策略适用于新闻稿件、百科条目等结构松散的通用文本，例如将一篇 1500 字符的新闻报道分隔为两个块（0-1000 字符与 900-1500 字符），通过 100 字符的重叠区域保留上下文关联性。

递归分块 递归分块通过分层迭代的分隔符优先级机制改进固定分块的缺陷，其核心流程是按照分隔符优先级序列逐级切割文本[6] 若初始分隔后的块仍超过设定长度，则递归调用下一级分隔符继续分块。例如，处理包含段落和子章节的技术手册时，系统会优先按段落分块，对超长段落进一步按句子分块，最终形成兼顾结构完整性与长度约束的分块结果。

基于文档结构的分块 基于文档结构的分块是指针对特定文件类型设计专用分隔逻辑。比如 Markdown 文档可以按照标题层级分隔，而 HTML、XML 和 JSON 文档则按照标签分隔。对于代码文档，常见的策略是先将代码转成抽象语法树 (Abstract Syntax Tree, AST)[7]，再按类和函数边界进行分隔，从而确保每个代码单元的结构完整性。

语义分块 语义分块则是通过构建句群并比较嵌入相似度来判断主题边界。具体而言，首先使用句号、问号等标点符号对原始文档进行语法层面的切分，得

[5]需要强调的是，原始文档和查询文本所使用的编码器可以是相同的，也可以是不同的，但必须确保它们在同一个语义空间中对齐。

[6]常见策略是按照段落分隔符、换行符和空格符的优先级切割。

[7]抽象语法树是一种用于理解和处理源代码的数据结构，被广泛应用于编译器设计、静态分析和代码转换等领域。

到句子序列；再针对每个句子，以它为锚点，选取其前后若干句（如前 2 句、后 2 句）组成一个句群，并使用嵌入模型将每个句群转换为一个嵌入表示；然后顺序计算相邻句群嵌入表示之间的距离（如余弦相似度），如果某个距离超过预设阈值，即发生了语义突变，则需在此处划分新的块。

选择合适的分块策略是数据预处理的关键。不同的文本类型对应着不同的分块方法。同时，分块参数的优化也至关重要，块尺寸决定了信息的粒度，重叠区保障了信息的连贯性，语义相似度阈值则确保了内容的相关性。此外，结合具体领域的特性定制规则，有助于使分块更加贴合实际需求，从而显著提升分块效果，实现更有效的内容检索和生成。

5.2.3 高维向量索引

第5.2.1节介绍了如何基于向量相似性找到与查询文本最相关的语义（常称为 k 近邻查询），实现精准语义匹配，然而，这种暴力检索的方式难以有效扩展，特别是面对 10 亿级别的大规模向量数据库时。因此，需构建高维向量索引结构，通过近似 k 近邻查询算法平衡效率与精度。

索引的工作原理类似于图书的目录，是一种典型的“以空间换时间”的方案。例如，如果用户希望查询书中某个主题，可以从头开始一页一页地查找（类似“暴力检索”），也可以通过目录直接定位到页码（类似“索引检索”）。传统的数据库索引一般是通过 B^+ 树和哈希实现，考虑到嵌入向量一般有几百甚至上千维度，相关的索引技术故而常被称为高维向量索引。目前，主流的高维向量索引一般有基于哈希、树和图三种形式，其中基于图的 HNSW 索引[31]是最主流的实现之一。对于高维向量索引的设计与实现的详细阐释已经超出了本书的范围，读者可以参考 Faiss 维基[8]或查阅其他资料以了解更多相关内容。本节将通过最简单的 `IVFFlat` 索引，帮助读者了解向量数据库索引是如何加速检索过程的。

如图5.3所示，`IVFFlat` 是典型的倒排索引应用。首先通过聚类算法将整个向量空间划分为 $nlist$ 个簇（Cluster），在示意图中 $nlist = 7$；对于查询向量 q，探测指定的最近的 $nprobe$ 个簇，此时，$nprobe = 1$（对应图中蓝色区域），即仅需要检索蓝色区域的向量。这种方式可以将搜索空间减少 $nlist/nprobe$ 倍，通过调节这两个参数可以平衡向量检索的效率和质量。

[8]参见`https://github.com/facebookresearch/faiss/wiki`。

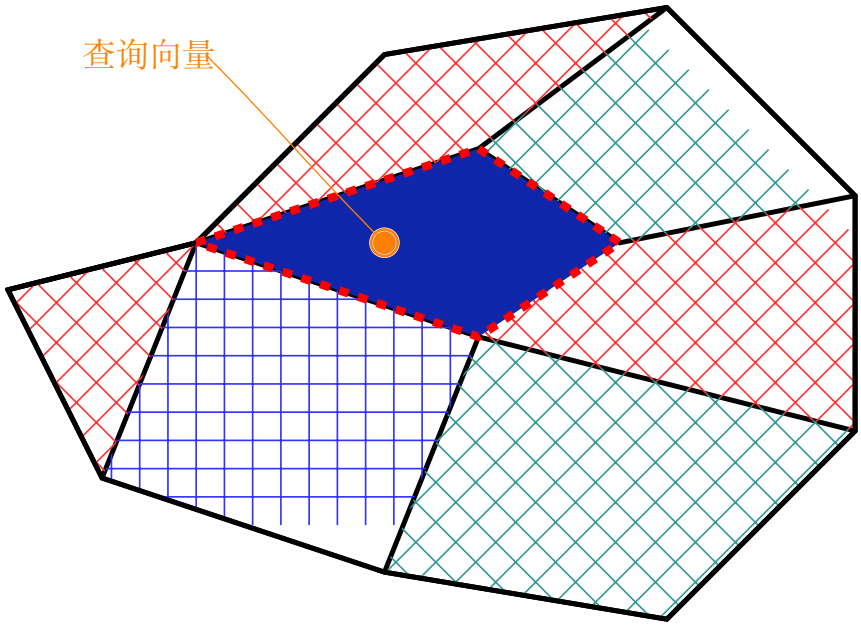


图 5.3: IVFFlat 搜索示意图

5.3 广义检索增强生成

第5.2节讨论的 RAG 主要以向量数据库作为外部知识库，来增强大模型的生成能力。本节将进一步讨论广义的检索增强生成，其基本结构与图5.1类似，两者的主要区别在于检索外部知识库的方式不同。需要强调的是，本节讨论的广义检索增强生成不仅涵盖基于向量数据库的 RAG，还包括基于联网、关系数据库或知识图谱的 RAG，基于联网、关系数据库或知识图谱的 RAG 也可与向量数据库技术结合使用。

5.3.1 基于联网检索的 RAG

1. 基本概念

互联网作为数字时代的基础性信息媒介，其持续生成的大规模实时数据流，为人工智能系统的算法训练与模型优化提供了肥沃土壤。据 Statista 等权威媒体统计[9]，互联网每天生成约 4.0274 亿 TB 的数据。到 2025 年，互联网数据总量预计将达到 182 ZB。面对如此庞杂的动态信息环境，基于联网检索的 RAG 应运而生。

[9]参见https://www.statista.com/topics/1145/internet-usage-worldwide/

基于联网检索的 RAG 通过实时抓取互联网上的网页、学术期刊、行业报告等开放数据，动态扩展了大语言模型的知识边界，有效弥补了传统 RAG 依赖静态知识库的局限。时效性是网络检索增强的一大优势。目前主流的大模型应用，包括 DeepSeek、豆包、Kimi、ChatGPT 和 Grok 等提供的联网搜索功能均可以看作是此类 RAG 技术的应用。比如，当用户查询“最新的 AI 技术趋势”时，基于联网检索的 RAG 能够从互联网上获取最新的研究、新闻和报告。然而，与经典检索增强生成相比，基于联网检索的 RAG 也存在明确不足，即如果网络上的信息质量良莠不齐，检索到的内容可能包含不准确、过时甚至有害的信息。

总体而言，联网检索的工作方式类似搜索引擎，即根据用户输入匹配相关网页[10]。例如，用户可以使用必应[11]等搜索引擎提供的开放平台快速实现网页检索服务。除了传统的搜索引擎，目前也出现了专为 LLM 设计的搜索引擎，即旨在帮助开发者实现具备高效、精准检索互联网信息能力的智能体，其中代表性的服务是 Tavily[12]。此外，Perplexity AI[13] 等平台也专注于为 LLM 提供语义化搜索能力，支持自然语言查询与上下文理解，进一步优化了知识获取效率。除了通用的搜索引擎，对指定网页的检索也可以看作是基于联网检索的简易 RAG 实现。例如，如果大模型应用需要关注实时股票价格，则可以通过设计网络爬虫访问新浪财经[14]等网页，通过解析 HTML 的内容得到相关信息。

2. DeepSearch 和 DeepResearch

在标准 RAG 架构中，检索质量直接制约着最终生成效果。面对互联网海量异构数据源，传统检索方式常面临语义漂移、知识碎片化等挑战。为此，研究者提出了深度检索增强方案：DeepSearch 和 DeepResearch。DeepSearch 作为搜索范式的系统性重构方案，其核心机制建立在“搜索-阅读-推理”三阶段的闭环迭代框架之上，如图5.4所示。该框架通过分布式搜索引擎对互联网进行查询，并利用大模型专用解析工具（如 Jina Reader[15]）将 HTML 网页内容转换为适合大模型处理的纯文本格式，并根据提取的信息和推理结果动态优化模型的决策，包括问题拆解、策略调整或应答终止等。从技术架构视角，DeepSearch 实质是一个集成网络工具的大模型智能体，其决策引擎可通过提

[10]在实践中，这个过程需要对用户的查询问题进行重写，使其对搜索引擎算法更加友好。

[11]必应平台参见网址：https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

[12]Tavily 官网介绍：https://tavily.com/

[13]Perplexity AI 介绍参见：https://www.perplexity.ai/

[14]新浪财经参见：https://vip.stock.finance.sina.com.cn/mkt/

[15]Jina Reader 介绍参见https://jina.ai/reader/

示工程或专用推理模型（如 Deepseek-R1）驱动决策调整，核心设计原则始终围绕搜索、阅读与推理的循环迭代以实现渐进式优化。

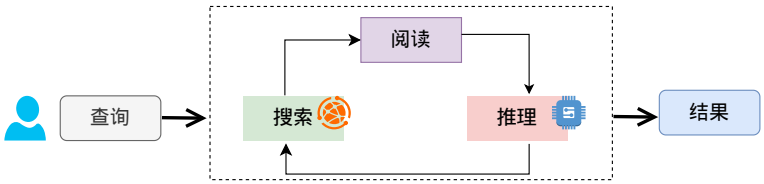


图 5.4: DeepSearch 架构图

在此基础之上，DeepResearch 进一步扩展为结构化研究框架的自动化实现，专注于端到端的长篇报告生成。以生成学术报告为例，DeepResearch 首先构建符合学术规范的目录模板，并将研究目标分解为章节子问题；随后，通过 DeepSearch 对各部分进行证据检索与内容合成；最终，利用全局一致性校验模块整合输出，确保术语统一性与逻辑连贯性。二者的协同体现为开放式搜索能力与结构化写作范式的深度融合，既保障事实准确性，又满足专业文档的格式与逻辑要求，为自动化研究工具的设计提供了参考范式。表5.3梳理了 2025 年 2 月发布的相关产品，由此可知基于 DeepSearch 或 DeepResearch 的 RAG 是目前最火热的人工智能应用之一。

表 5.3: 基于 DeepSearch 或 DeepResearch 的 RAG 产品

发布时间	**公司**	**产品**	**类型**
2025-02-01	LangChain	Ollama+OpenSearch	开源
2025-02-02	OpenAI	ChatGPT+DeepResearch	私有
2025-02-04	Hugging Face	Open DeepResearch	开源
2025-02-15	Perplexity	DeepResearch	私有
2025-02-17	X AI	Grok3+DeepSearch	私有
2025-02-22	百度	百度搜索 +DeepSeek R1	私有
2025-02-23	腾讯	微信搜索 +DeepSeek R1	私有

5.3.2 基于关系数据库的 RAG

1. 基本概念

数据库管理系统是当今信息系统的基石，被广泛应用于金融、航空和电信等领域。尽管历史上出现了多种数据模型（如层次模型、网状模型和面向对象模型等），关系模型凭借其易用性在 20 世纪 80 年代占据了统治地位。时至今日，关系模型依然是最主流的数据模型[32]。人们常说的“数据库”一般是指基于关系模型的数据库。在关系模型中，数据以类似二维表格的形式存储[16]。表 5.4展示了一个教师的关系，其关系模型可以表示成“教师 (ID，姓名，学院)”。数据库系统主要由存储管理和查询处理两个模块组成，其主要目标是提供一种方便且高效的数据库信息存取方式[33]。

表 5.4: 教师关系

ID	姓名	学院
101	张三	计算机
102	李四	金融
103	王五	统计

常见的关系型数据库有 MySQL、PostgreSQL、Oracle 和 SQL Server 等。尽管它们各有特色，但是均提供了相似的数据定义语言（用于定义数据库模型）和数据操纵语言（用于表达数据库的查询和更新），即后文将介绍的结构化查询语言 (Structured Query Language, SQL)[17]。

2. SQL

SQL 是关系数据库的标准语言，最早由 IBM 开发。目前，最新的 SQL 标准是由美国国家标准化组织和国际标准化组织联合发布的 SQL-2023。根据知名编程问答网站 StackOverFlow 最新的问卷结果[18]，SQL 在开发人员最想掌握的编程语言中排名第三。SQL 是典型的声明式语言，能够以接近英语自然语言的方式操作数据库[19]。不同于其他编程语言，用户在使用 SQL 时，往往只需表达其查询目标，而无须关心其查询步骤。例如，当用户想查询工号为 101 的教师所在的学院时，对应的 SQL 语句可以写成：

[16]术语“关系”来自集合论，在关系数据库中它表示未一张表。

[17]尽管 SQL 从字面意义上是查询语言，但它也可以执行数据定义、授权等其他操作。

[18]来源：`https://survey.stackoverflow.co/2024/technology`。

[19]需要说明的是，尽管由 SQL 标准的存在，各大数据库系统所支持的 SQL 语法略有差异。

```
SELECT 学院
FROM 教师
WHERE ID = 101;
```

从上面的例子可以看出，用户仅需要掌握 SELECT、FROM 和 WHERE 等少量关键词即可使用 SQL 高效地完成数据库查询。然而，SQL 和真正的自然语言仍有差异，对于普通用户而言仍有一定的难度。因此，研究人员提出了多种算法和模型用于将自然语言转换成 SQL，如下面将介绍的 Text2SQL（也称 NL2SQL）。

3. Text2SQL

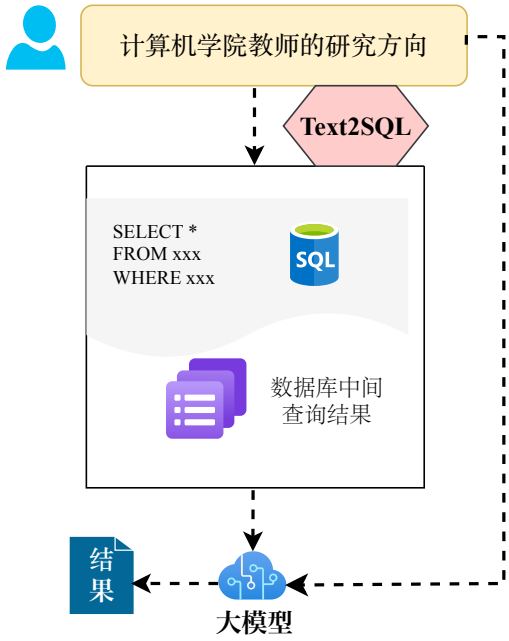


图 5.5: 关系数据库检索增强示意图

Text2SQL 技术发展主要经历四个阶段，分别基于规则/模板方法、传统 Seq2Seq 模型、预训练模型和大语言模型实现，其中前两种方法的准确率通常不到 20%，目前已经被淘汰。Text2SQL 分为模式关联、编码和解码三个步骤，为了取得更好的效果，一般在编码步骤采用预训练模型（如 BERT、T5），并使用图神经网络辅助执行模式关联，部分工作[34]还研究了 Text2SQL 的中间表示技术。近两年，最先进的 Text2SQL 方法都是围绕着大模型技术，主要通过

思维链 (Chain-of-Thought, CoT) 和情景学习等提示工程技术, 其中 MiniSeek 模型在 Spider 评测基准[35]上准确率超过 90%, 其表现媲美人类专家的能力[20]。

在使用 RAG 构建大模型应用的过程中, 可以首先将用户请求转换成 SQL 并在数据库中执行得到结果, 再将相关结果作为提示词的上下文用于增强大模型的输出。图5.5展示了当用户查询“计算机学院教师的研究方向”时, 可以通过 Text2SQL 技术将其转换成一条或多条 SQL 语句, 并在数据库中执行得到相关中间结果 (即数据库中多行记录), 接着大语言模型整合从数据库中所得中间结果和查询文本以生成最终查询结果, 并将生成结果以报告等方式呈现。此外, 最新的研究还提出了表格增强生成 (Table-Augmented Generation, TAG)[36]的技术, 研究如何综合应用大模型的推理能力从而挖掘关系数据库中复杂的表格关联语义。

5.3.3 基于知识图谱的 RAG

1. 基本概念

知识图谱 (Knowledge Graph) 的概念最早由谷歌 2012 年提出, 其将知识图谱定义为用于增强搜索引擎功能的辅助知识库[37]。知识图谱是一种结构化知识的表示形式, 它将现实世界中的实体及其关系以图的形式进行建模。简单来说, 知识图谱是三元组形式的事实集合, 每个三元组通常表示为“实体-关系-实体”。上述三元组集合可以直观地表示为图结构, 其中节点代表实体, 边表示关系, 边的方向则表明实体之间的逻辑关系。图5.6展示了一个简单的知识图谱, 表示“爱因斯坦于 1864 年出生在德国”、“柏林是德国首都”、“爱因斯坦获得了诺贝尔奖”等信息。

在大模型技术爆发式发展之前, 知识图谱凭借结构化认知优势被广泛认为是实现认知智能的核心基础架构。其强大的语言认知能力体现在对复杂语义关系的精准建模, 良好的可解释性则源于其显式的三元组表示形式, 这使得知识推理过程具有透明性和可追溯性。即便在当今大模型主导的技术范式下, 知识图谱仍在 RAG 体系中发挥着不可替代的作用: 其结构化知识库为语言模型提供事实性校验基准, 显著缓解生成内容的幻觉问题; 实体关系网络可增强检索系统的语义关联挖掘能力, 提升复杂查询的解析精度; 与向量数据库融合后, 更形成” 符号推理 + 语义匹配” 的双引擎架构, 兼顾精确术语检索与模糊语义扩展。这种符号主义与连接主义的深度融合, 正推动人工智能系统突破静态知识边界, 向具备逻辑严谨性与动态适应性的新一代认知架构演进。

[20]Spider 是 Text2SQL 领域一个知名的评测基准, 于 2018 年由耶鲁大学发布。

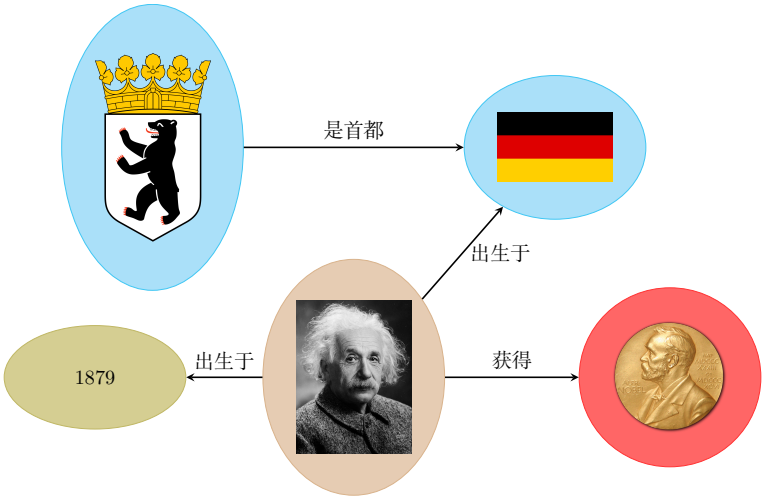


图 5.6: 知识图谱示意图

2. SPARQL

构建知识图谱的一种常见方式是基于资源描述框架 (Resource Description Framework, RDF)。RDF本质上是一个数据模型，它提供了一个统一的标准来描述网络上的资源（包括类、属性和实例等）。RDF也是采用三元组的形式（主语-谓语-宾语）来表达实体及其关系，从而构建一个语义网络。这种模型为知识表示和管理提供了统一的语言和框架，广泛应用于语义网和人工智能领域。

SPARQL 是标准的用于操作 RDF资源的语言，与 SQL 类似，SPARQL 能够灵活地处理 RDF 三元组数据的模式匹配、图遍历和复杂推理。下面展示了如何使用 SPARQL 查询爱因斯坦出生地，其中 `ex` 表示一个命名空间前缀，`WHERE` 子句表示一个三元组。本节不对 SPARQL 的详细语法进行阐述，但读者可以从下面的示例发现其与 SQL 的相似之处。在 RAG 系统中，SPARQL 可用于精准检索知识图谱中的结构化事实，弥补纯向量检索的不确定性，从而提升生成答案的准确性和可解释性。

```
PREFIX ex: <http://example.org/>

SELECT ?birthPlace
WHERE {
  ex:爱因斯坦 ex:birthPlace ?birthPlace .
}
```

3. GraphRAG

与 Text2SQL 技术类似，研究人员也提出了 Text2SPARQL 方法[38]，该方法致力于将自然语言直接转换为图谱查询语句。然而，知识图谱的构建逻辑与关系数据库存在本质差异，其知识体系的建立往往需要实体消歧、关系推理等复杂认知过程。因此，完整的业务流程通常需要分两个阶段实现：首先利用大语言模型辅助知识图谱的构建，随后在查询阶段通过知识图谱补全结构化事实。微软研究院于 2024 年推出的 GraphRAG 框架[21]正是这一技术路线的典型代表。大量的实验表明，GraphRAG 能够提升生成模型的准确性、逻辑性和上下文关联性[39]。

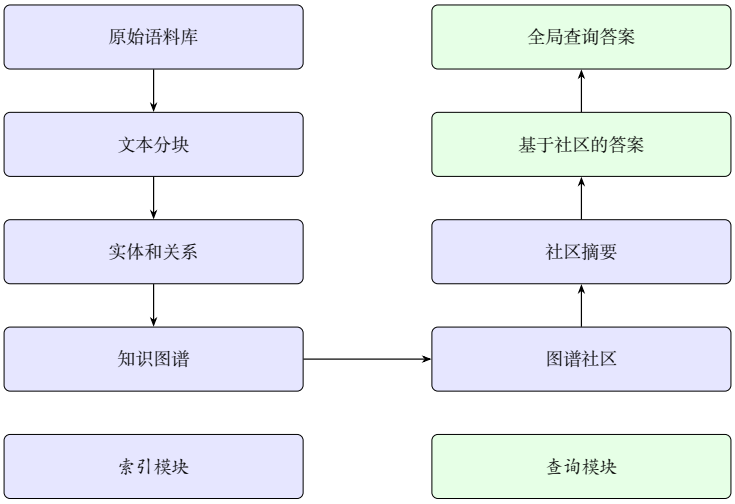


图 5.7: GraphRAG 索引和查询流水线

GraphRAG 项目是微软在图机器学习领域深耕多年的结果，其核心组件包括索引和查询两个模块。对于索引模块，它首先将原始文本语料库切分成一系列文本单元（类似 5.2.2 中提及的分块），然后，从这些文本单元中使用大模型抽取实体和关系等并构建知识图谱；接下来，对知识图谱进行层次聚类，一组紧密相连的节点被称为社群（Comminity）；最后，自底向上生成每个社群及其组成元素的摘要，这种方法有助于从整体上理解数据集。

对于检索模块，GraphRAG 主要支持两种查询模式，分别是 1）全局查询，表示通过调用社群摘要数据，支持对全量语料库进行整体性推理分析（如宏观趋势、跨领域关联等问题）；2）局部查询，表示通过展开特定实体的相邻节点

[21] 参见项目主页：`https://github.com/microsoft/graphrag`。

及相关概念网络，实现细粒度实体级推理（如人物关系链、事件细节溯源等场景）。图5.7展示了索引阶段和全局查询的过程[22]，系统采用社群检测算法将图索引划分为多组，使得大模型能够在索引构建和查询响应阶段并行地执行摘要生成。通过对所有相关社群摘要进行多轮迭代，每轮迭代基于社群间的关联强度和用户查询的相关性进行更新，并经过重排名（根据用户查询的语义相关性和社群的重要性进行排序）和过滤（去除与用户查询不相关的社群摘要）等操作，最终得到针对用户查询的“全局答案”。多轮迭代的具体规则包括：在每轮迭代中，根据社群摘要与用户查询的语义相似度进行打分，保留得分较高的社群摘要进入下一轮迭代，直到达到预设的迭代次数或满足收敛条件。

5.4 讨论

讨论 5.1. 本章讨论了基于嵌入向量的语义检索，然而传统的 BM25 等算法在精确匹配关键词方面仍依然发挥着重要的作用。因此，研究人员提出了混合向量检索和 BM25 的方案。请讨论其优缺点。

讨论 5.2. 主流的嵌入向量是稠密的，而研究人员也设计了基于神经网络的稀疏嵌入向量（如 SPLADE）。请分析 SPLADE 等稀疏嵌入向量的优缺点。

5.5 习题

习题 5.1. 非结构化数据与结构化数据有何区别？请举例说明。

习题 5.2. 简述向量数据库在大模型应用中的角色及优势。与传统关系型数据库相比，它适用于哪些典型场景？

习题 5.3. 在经典检索增强生成架构中，文本分块技术的作用是什么？常见的分块策略有哪些？

习题 5.4. 除了余弦距离，哪些常用的度量嵌入向量距离的方法？

习题 5.5. 请解释 `IVFFlat` 索引的基本思路，并分析其如何平衡向量检索的效率和质量。

习题 5.6. 联网功能是否为大模型的内置功能？请进行分析。

习题 5.7. 如果联网搜索的多条结果相互冲突、存在矛盾，请提出若干可行的解决思路。

[22] 示意图参考了文献[39]。

习题 5.8. 如何理解 SQL 是声明式语言？这有什么优势？

习题 5.9. 知识图谱的基本元素是什么？

习题 5.10. 请分析图5.7中的流水线哪些步骤一般需要通过大模型完成。

第六章　大模型认知框架

6.1　概述

认知框架 (Cognitive Architecture)是智能系统用来组织和理解信息的基本结构和模式。认知框架借鉴人类思维模式，通过提示工程技术构建具有人类水平的通用人工智能系统，是目前通用人工智能研究的重要方向[40]。同时，认知框架的运用也改变了用户和大模型的交互方式，是目前提升大模型性能的可行方法之一。

大模型认知框架的常见方法包括情景学习 (In-context Learning)[41]、思维链 (Chain-of-Thought, CoT)[42]、自我一致性 (Self-consistency Prompt)[43]、计划与执行[44]、ReAct[45]和反思 (Reflexion)[46]。情景学习认知框架利用大模型的语言理解和泛化能力，从具体提示词示例总结经验与规律从而更好的完成用户问答；CoT认知架构则通过将复杂问题分解成多个子问题，并通过连续的推理步骤，使大模型回答更加准确可控；自我一致性框架借鉴了人类从多个视角对问题进行求证的研究方法论，通过多维度验证机制提高大模型输出的准确性和可靠性；计划与执行框架模仿人类分解复杂任务的认知特征，通过系统化的任务规划和执行策略，增强了大模型完成复杂任务的能力；ReAct 框架结合了 “边思考边行动” 的认知模式，将推理过程与具体行动有机结合，让大模型具备了 “思考-执行-反馈” 的循环工作模式；反思框架借鉴了人类自我反省的能力，建立自我评估和优化机制，使模型能够审查并改进自身输出质量。

本章将介绍当前主流的六种认知框架，从原理概念、框架设计到应用案例进行梳理。同时，介绍蚁群智能、认知循环理论等前沿多智能体认知框架，并探讨其在自动驾驶、机器人协同和智能医疗诊断等领域的应用前景。

6.2 情景学习

6.2.1 基本概念

情景学习是大型语言模型涌现行为的具体表现之一，即大模型能够借助提示中的示例理解任务需求，并完成训练数据中未出现的特定任务[47]。在情景学习框架下，大模型会基于提示词中的上下文信息，深入理解用户意图，从而生成更精准且稳定的回答。这些上下文信息可以包括用户的先前提问、相关背景知识或特定的任务演示示例，为模型提供更全面的参考。其中，基于少量示例进行引导的方式被称为少样本提示 (Few-shot Prompting)，是情景学习中的一种常见形式。

少样本提示通过在提示中提供少量输入-输出示例，引导模型在无需参数更新的情况下完成特定任务。这一过程类似于人类的 “观察与模仿” 学习方式，即用户通过精心设计的示例，让模型理解并复现特定问题的回答模式。值得注意的是，不同于需要调整模型参数的传统少样本学习 (Few-shot Learning)方法，少样本提示不进行模型参数更新，能够在少样本情形下高效调整大模型输出。

少样本提示之所以能够实现少量示例条件下的任务适应，根本原因在于大规模语言模型在预训练阶段已从海量数据中学习到丰富的语言模式和推理能力, 即模式泛化能力。该能力在遇到适当的上下文提示时可以被激活，从而支持模型在新任务上的推理与生成。通过在提示中提供精心设计的输入输出示例，用户能够引导模型更好地调用内部知识，复现特定问题的解答方式。相关研究[48]还表明，即使在没有示例的零样本提示 (Zero-shot Prompting)情形下，大模型依然能够在一定程度上展现出情景学习能力。因此，通过一定的示例设计，用户可以有效引导模型的输出方向，使其更契合实际需求。

6.2.2 框架设计

当前，情景学习通常会引入示例仓库作为示例池，并从中根据任务选取代表性示例引导大模型。如图6.1所示，整体认知框架主要由三个关键模块组成：示例仓库、示例选择和反馈优化，三者共同协作来提升模型的提示效果。

- 示例仓库模块：示例仓库是情景学习的基础模块，存储了大量用于指导模型回答问题的高质量示例，为后续示例选择提供了丰富的案例库。
- 示例选择模块：示例选择模块通过分析用户问题，从示例仓库中选择最合适的一条或多条案例作为模型回答的上下文信息，从而确保所选示例

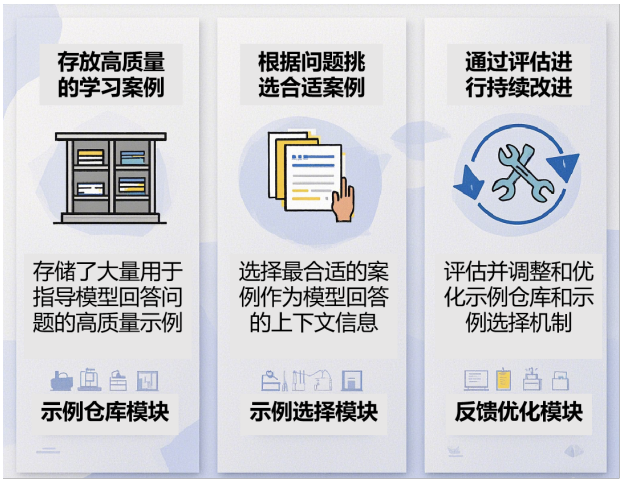


图 6.1: 情景学习框架图

能够有效引导模型输出更加准确的回答。

- 反馈优化模块：反馈优化模块的主要功能是通过模型回答，评估示例选择模块和示例仓库中案例质量，并用于调整和优化示例仓库和示例选择机制。

在上述三个模块中，示例选择是情景学习的核心模块，需要遵循代表性、简洁性、多样性以及相关性四个基本原则。

- 代表性：代表性是指选择的示例能够充分反映任务的核心特征或典型场景。一个好的示例应当涵盖任务中的关键要素，使模型能够从中学习到普遍适用的模式。例如，在情感分类任务中，选择一个包含正面、负面和中性情感的示例集，能够帮助模型理解情感分类的基本逻辑。如果示例不具代表性（如缺少中性情感），模型可能无法正确识别某些关键类别。

- 简明性：简明性强调示例的表达应当明确、易懂，避免冗长或模糊的语言。简明的示例能够帮助模型快速抓住重点，减少误解的可能性。例如，在机器翻译任务中，一个简洁且语法正确的句子对（如 “Hello, how are you?” → “你好，你好吗？”）比一个复杂且含有歧义的句子对更有利于模型学习。过于复杂的示例可能会让模型难以捕捉核心信息。

- 多样性：多样性是指示例应当覆盖任务中的不同形式、场景或边缘情况，避免单一化。例如，在文本分类任务中，选择来自不同领域（如科技、文学、新闻）、不同风格（正式与非正式）和不同长度（短句与长篇）的文

本作为示例，能够使模型适应更广泛的输入场景。如果示例过于单一（如仅包含科技领域的短文本），模型可能无法处理其他领域的复杂文本。

- 相关性：相关性强调示例必须与目标任务高度相关，能够直接支持模型的训练或推理。不相关的示例可能会误导模型，降低其性能。例如，在问答系统中，选择一个与问题主题（如“地球的自转周期是多久？”）紧密相关的示例（如“地球上一天是多少个小时？”→“24 小时”）比选择一个无关的示例（如“太阳系有八大行星”）更有助于模型学习如何生成正确的回答。相关性的缺失会导致模型生成偏离主题的结果。

6.2.3 应用案例

1. 提升大模型回复中的情感表达

在实际应用中，仅生成正确的答案不足以满足需求，许多场景要求大模型以自然、亲切的方式进行交流，使得交互体验更接近于人际沟通。例如，在客服系统中，模型需要学会礼貌提问、耐心解释，并根据用户反馈调整回答。合理设计提示词有助于模型理解背景并生成富有情感的回应。下面以耳机故障问题为例，展示情境提示对大模型输出风格的影响。

在未使用情境提示时，发送给大模型的原始输入如下：

示例 6.1

用户提问：“我的耳机有一边没有声音了，我该怎么办”

Kimi 模型在接收该输入后生成的回答如图6.2所示。



图 6.2: 无情景提示的客服回复

可以看到，在没有情景提示的情况下，大模型会产生机械化的回复内容。为了引导模型生成更具情感的回答，可以通过情境学习的三个模块来引导模型以更贴近人类交流的方式进行回应。

（1）示例仓库模块

示例仓库模块中存放了各类典型交流场景的示例，作为情境学习的基础素材。假设现在示例仓库中的部分内容如示例 6.2所示。

示例 6.2

客服类示例片段 1（目标类型）
用户："我买的手机有点问题，电池不充电了。"
客服："您好，抱歉听到这个问题。请问您尝试过重新启动手机吗？"

客服类示例片段 2（目标类型）
用户："我试过了，还是不行……"
客服："非常抱歉给您带来不便，请问手机型号……"

非相关示例片段（背景噪音）：
问题："请写一首关于秋天的诗。"
回复："枫叶如火，秋风如诗，黄昏掠过心头的思绪。"
问题："Python 如何连接数据库？"
回复："您可以使用 `sqlite3` 模块，示例如下……"

这些示例质量很高，但它们风格与场景完全不同，因此需要示例选择模块对示例仓库中的示例进行筛选。

（2）示例选择模块

示例选择模块将从示例仓库中筛选出与用户问题最匹配的示例，确保所选示例不仅在内容上相关，还在风格上与目标任务一致。例如，当用户提出问题："我的耳机有一边没有声音了，我该怎么办?"时，系统从示例库中智能筛选结构相似、场景匹配的客服类对话示例。系统会优先选择与"设备故障"或"用户困扰"相关的客服对话示例，从而避免混杂如技术说明类或与情感无关的示例。

根据筛选出的示例，系统将生成提示词，如示例 6.3所示，帮助模型模仿示例中的回复模式和语言风格，确保生成的回答具有一致的风格和语气。

示例 6.3

提示词:
下面我将发送给你一个示例和一个问题，任务是模仿示例中的回复模式和语言习惯来对我发送的问题进行回复。
示例:
用户:“我买的手机有点问题，电池不充电了。”
客服:“您好，抱歉听到这个问题。请问您尝试过重新启动手机吗？”
用户:“我试过了，还是不行。电池显示充电中，但一直不增加电量。”
客服:“非常抱歉给您带来不便。请问您能告诉我手机的具体型号以及您使用的充电器和数据线是否是原装的吗？”
用户提问:“我的耳机有一边没有声音了，我该怎么办？”

模型回复:
“客服:您好，很抱歉听到您的耳机出现了问题。请问您尝试过重新连接耳机或者重启设备吗？另外，您能告诉我耳机的具体型号以及您使用的设备是什么吗？”

(3) 反馈优化模块

在模型生成回复之后，反馈优化模块将对结果进行用户反馈的收集和质量评价。用户的反馈信息将被用来优化示例库中的示例排序。

系统通过以下两个主要任务进行反馈优化:

a) **示例排序与淘汰**

i) 用户评分高的示例(如回复中包含理解用户情绪、提供实用建议)将在后续被优先调用。

ii) 用户评分差的示例将被自动下调优先级或剔除出提示词候选库。

b) **判断是否将新生成的示例加入到示例仓库中**
当模型针生成了优质回复时（见示例 6.3)，系统将执行以下步骤:

i) 内容结构检测:语气自然、礼貌;回复包含合理步骤建议。

ii) 用户正反馈判断:如用户点赞、评论“很有帮助”、点击“是否解决问题”选项为“是”。

iii) 示例纳入机制:将该生成结果纳入客服类示例库，打上标签“耳机问题”、“多步排查”、“高情感评分”。

通过这些优化，系统实现了一个动态进化的示例库，能够随着用户反馈不断提升模型的回应质量。

2. 激发大模型的创意回答

除了提升自然交流风格，大模型还可以通过情境提示学会更具创意的回答方式。比如，当用户提问“如何让一只猫学会使用马桶”时，模型不仅能给出可行的步骤，还能让回答充满趣味和想象力。这种创意并非天生具备，而是通过少样本提示引导模型学习而来的。通过提供一些生动、有趣的示例，模型可以掌握如何在回答中融入幽默和想象力，不再只是机械地列出步骤。

在引入少样本提示之前，大模型的回答通常较为规整而缺乏创意。例如，当用户提问：“如何让一只猫学会使用马桶?”时，模型的回复往往简短而机械，缺少幽默和生动的表达，如图6.3所示。

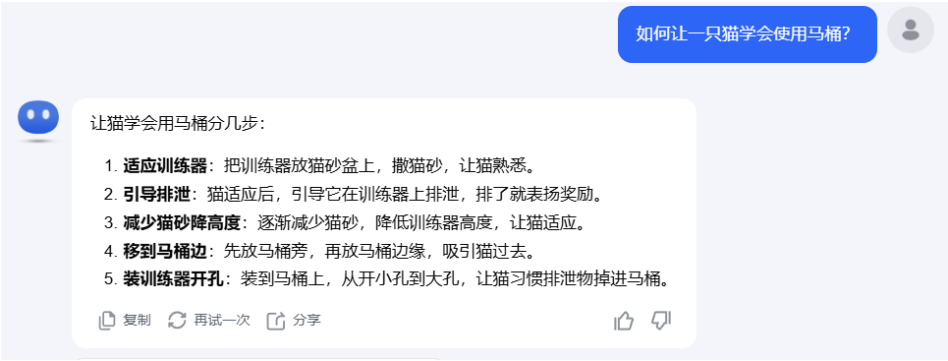


图 6.3: 无情景提示的客服回复

在引入少样本提示后，大模型能够展现出截然不同的回答风格。如图6.4所示，用户提供了“教鹦鹉说人话”和“教金鱼跳舞”的示例，模型根据这些示例不仅给出了实际的建议，还融入了幽默和创意，回答更加生动有趣。

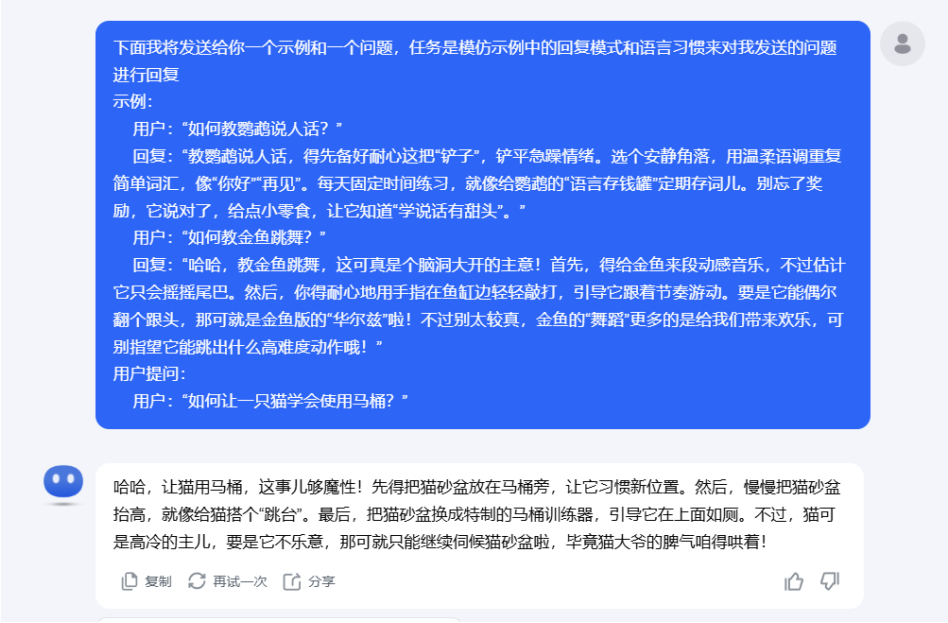


图 6.4: 有情景提示的客服回复

6.3 思维链

6.3.1 基本概念

CoT[49]是一种通过将复杂问题拆解为一系列子问题，逐步推理得出最终答案的方法。该框架的核心思想是通过连贯的推理步骤将问题分解，每个中间结果为最终结论提供支持，从而帮助模型更精确地解决需要多步推理的任务，如数学题、推理题和常识推理等。DeepSeek-R1 论文提出长思维链奖励，通过奖励模型鼓励多步推理，从而提高模型在解决复杂问题的能力[15]。思维链不仅使得模型的思维过程更加清晰易懂，还能够提高答案的准确性，因为每一步推理都经过独立验证，确保最终结论的可靠性。

6.3.2 执行框架

CoT的框架结构主要包括问题分解、逐步推理和总结回答三个模块。

- 问题分解模块：该模块将复杂、高难度的问题拆解为多个简单、易解的子问题。在问题分解阶段，模型需要根据问题的特性识别出不同的层次

和关键点。例如，对于一个包含多个计算步骤的数学问题，可以将其按步骤分解为多个子计算任务。

- 逐步推理模块：当问题被分解后，模型进入逐步推理阶段。在此阶段，模型会根据每个子问题的具体情况，依次进行详细的推理和解答。推理的过程可以按照已有的知识、规则或模式执行。在每一步推理中，模型需要明确每个步骤的逻辑依据，并结合前一步的结果，为下一个推理步骤提供支持。

- 总结回答模块：当所有子问题都被逐步推理并解决之后，模型将综合所有的中间结果，形成最终的结论。此时，模型的思维链已经完成了从问题到答案的推理过程，并且每一步都可以被单独检查和调整。

6.3.3 应用案例

思维链应用涵盖多种类型的问题。接下来以算术推理、常识推理、逻辑推理和复杂问题求解为例，分别从问题分解、逐步推理和总结回答三个模块进行讲解，展示模型推理过程中的关键环节与方法。

1. 算术推理

算术推理是思维链应用中的一个经典范例，它通过对数学问题的逐步分解和推理，帮助模型更清晰地理解并解决问题。示例 6.4展示了如何运用思维链的方法对一个简单的算术问题进行问题分解、推理过程和最终答案的总结。

示例 6.4 算术推理

问题：如果约翰有 5 个苹果，而玛丽的苹果数量是约翰的 3 倍，那么玛丽有多少个苹果?

问题分解

- 确定约翰的苹果数量。
- 确定玛丽的苹果数量与约翰的关系，即是约翰的 3 倍。
- 计算玛丽实际拥有的苹果数量。

逐步推理

- 首先，题目中明确给出约翰有 5 个苹果。
- 然后，玛丽的苹果数量是约翰的 3 倍，也就是说需要用约翰的数量乘以 3。
- 将 5 乘以 3，得到 $5 \times 3 = 15$。

总结回答

约翰有 5 个苹果。玛丽的苹果数量是约翰的 3 倍。要找出玛丽的苹果数量，需要将约翰的苹果乘以 3。5 个苹果 $\times 3 = 15$ 个苹果。玛丽有 15 个苹果。

2. 常识推理

常识推理是从日常经验和普遍知识中推导出合乎逻辑的结论。通过分解问题并逐步推理，模型能够更好地理解常见场景中的潜在关系。示例 6.5展示了思维链如何帮助模型从多个步骤中推导出合理的结论。

示例 6.5 常识推理

问题：如果某人对狗过敏，而他的朋友邀请他到家里玩，但是他朋友家里有一只宠物狗，他该怎么办？

问题分解

- 明确这个人对狗过敏。
- 朋友家里养有一只狗。
- 与狗相处可能引起过敏反应。
- 需要决定是否接受邀请，以及如何处理这类情况。

逐步推理

- 由于这个人对狗过敏，如果去朋友家，有很大可能会出现过敏反应。
- 为了健康和安全，应该避免前往有狗的环境。
- 拒绝邀请是对自己身体负责的做法。
- 可以提出其他见面方式，比如约在外面或朋友家以外的地方见面。

总结回答

这个人对狗过敏。朋友家有一只狗。与狗相处可能会引发这个人的过敏。为了避免过敏反应，这个人应该拒绝邀请。这个人可以建议在另一个地方与朋友见面。

3. 逻辑推理

逻辑推理是指利用规则进行推导的过程，常用于逻辑学、数学、计算机科学等领域。它通过符号的转换和运算，帮助模型解决抽象的问题。示例 6.6展示了思维链如何在逻辑推理中提供系统的步骤，帮助模型逐步得出准确的结论。

示例 6.6 逻辑推理

问题：如果 $A \Rightarrow B$，且 $B \Rightarrow C$，那么是否意味着 $A \Rightarrow C$?

问题分解

- 分析第一个条件：$A \Rightarrow B$，即如果 A 为真，则 B 为真。
- 分析第二个条件：$B \Rightarrow C$，即如果 B 为真，则 C 为真。
- 推导：在 $A \Rightarrow B$ 和 $B \Rightarrow C$ 同时成立的前提下，是否可以推出 $A \Rightarrow C$。

逐步推理

- 根据第一个条件，若 A 为真，则 B 为真，即 $A \Rightarrow B$。
- 根据第二个条件，若 B 为真，则 C 为真，即 $B \Rightarrow C$。
- 将上述两个条件连接起来：若 A 为真，则 B 为真；若 B 为真，则 C 为真。
- 因此可以推出：若 A 为真，则 C 为真，即 $A \Rightarrow C$。

总结回答

$A \Rightarrow B$ 意味着如果 A 为真，则 B 也必定为真。$B \Rightarrow C$ 意味着如果 B 为真，则 C 也必定为真。如果 A 为真，则 B 为真（来自步骤 1）；如果 B 为真，则 C 为真（来自步骤 2）。因此，如果 A 为真，则 C 也必定为真。

4. 解决复杂问题

复杂问题通常是指涉及多个推理步骤或需要对多个信息进行综合分析的问题。例如，问题“如果昨天下雨，而今天气温骤降，明天可能会发生什么?”涉及对多个条件的分析和推理，或者问题“周五的昨天的后天是周几?”需要通过多次推理和时间推算来得出正确答案。示例 6.7展示了如何运用思维链将复杂问题分解为多个简单步骤，逐步推导出最终的结论。

示例 6.7 解决复杂问题

问题：周五的昨天的后天是周几?

问题分解

- 确定当前日期。
- 确定昨天的日期。
- 确定后天的日期。

逐步推理

- 首先，我们需要知道当天是周五，那么我们可以开始推理“周五昨天的后天”是什么日期。
- 既然今天是周五，那么昨天的日期就是周四。我们需要找到周四的后天，即周四之后的两天。
- 既然我们已经确定今天是周五，昨天的日期是周四，那么我们需要找到周四的后天，即周四之后的两天。周四之后的两天是周六。

总结回答

当前日期是周五。昨天日期是周四。周四的后天是从周四开始，数到周五和周六。因此，周五昨天的后天是周六。答案是周六。

6.3.4 拓展方法

除了经典的思维链结构，其他变体如思维树和思维图也在多种应用中得到了广泛使用。如图6.5所示，思维树通过层次化的结构呈现问题的分解过程，而思维图则通过图形化的方式展示推理的多条路径和逻辑关系。每种方法都有其独特的优势，能够在不同的任务和应用场景中提供不同的视角和解决方案。接下来，将介绍这两种方法的基本概念和它们与思维链的关系。

1. 思维树

思维树 (Tree-of-Thought, ToT)是一种新兴的思维链拓展框架，旨在通过多条思维链来解决复杂问题[50]。与 CoT相似，思维树强调推理过程的透明性，但在结构和方法上有所不同。思维树通过维护树状结构的思维脉络，每个节点代表一个思维步骤，而边则表示思维之间的逻辑关系。该结构允许模型在从多个不同的思维链进行推理，从而实现更深入的探索。思维树框架能够让模型在生成解决方案的同时对生成的过程进行自我评估和调整。具体而言，思维树结合了搜索算法（如广度优先搜索（BFS）和深度优先搜索（DFS)），使得模型能够在探索问题答案的同时，验证每一步的有效性。

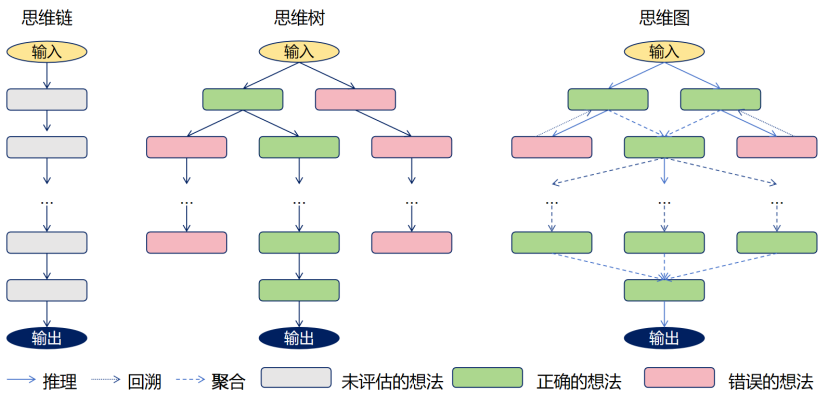


图 6.5: 思维链结构及相关变体

以“算 24 游戏”[1]为例，思维树框架首先将问题分解为多个思维步骤，每个节点代表一个操作步骤，如选择数字、应用运算符和组合结果等。如图6.6所示，模型通过逐层推理，从根节点开始，逐步探索不同的运算组合，并在每个节点生成多个候选解。通过评估这些解的有效性，模型能够逐渐缩小可能的解空间，最终找到符合要求的解决方案。具体步骤如下：

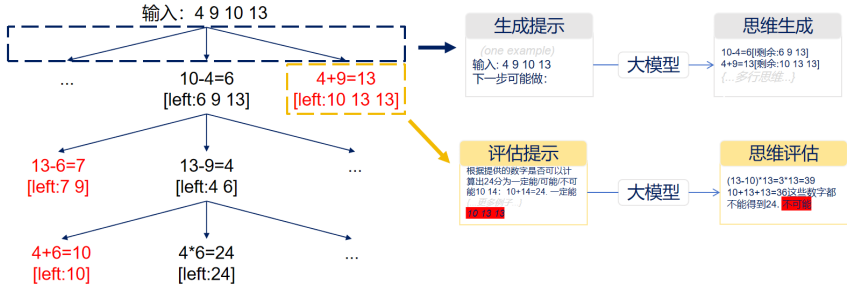


图 6.6: “算 24 游戏” 示例

- 输入问题：任务是使用 4 个数字和基本算术运算（加、减、乘、除）得到 24。例如，给定数字 4, 9, 10, 13。
- 思维分解：将问题分解为 3 个中间步骤，每个步骤是一个中间方程式。例如：
  - 第一步：选择两个数字进行运算，生成一个新的中间结果。

[1]算 24 游戏是一种数学游戏，给定四个数字（通常是 1 到 13 之间的整数），玩家需要使用加、减、乘、除和括号等运算符，通过对这四个数字进行适当组合，得到结果为 24。

  - 第二步：使用第一步的结果和剩余的两个数字进行运算，生成第二个中间结果。
  - 第三步：使用第二步的结果和最后一个数字进行运算，得到最终结果 24。

- 思维生成：对于每个中间步骤，生成多个候选项。例如，对于第一步，可以生成以下候选项：$13-9=4$、$10-4=6$，确保每个候选项都是合理的中间结果。

- 状态评估：对每个候选项进行评估，判断其是否能最终得到 24。评估结果可以是 “一定能”、“可能” 或 “不可能”。例如，对于候选项 “13 - 9 = 4”，评估结果为 “一定能”，因为 4 是一个合理的中间结果。对于候选项 “$2+3+5+7=17$”，评估结果为 “不可能”，因为 17 无法通过后续运算得到 24。

- 选择最佳路径：选择评估结果为 “一定能” 的路径，逐步构建完整的解决方案。如果某些步骤被评为 “可能”，可以进一步探索这些步骤，生成新的候选项进行验证。例如，对于候选项 “$13-9=4$”，继续探索下一步：$10-4=6$、$4\times 6=24$。最终，选择路径 “$(10-4)\times(13-9)=24$” 作为解决方案。

- 广度优先搜索（BFS）：在每一步，保留最优的 5 个候选项，继续探索下一步。例如，第一步生成 5 个候选项，评估后保留 5 个最有可能的候选项，继续生成第二步的候选项。重复此过程，直到最终生成 24。

2. 思维图

思维图 (Graph-of-Thought, GoT) 是在思维链和思维树基础上进一步拓展的认知框架，旨在通过更灵活、多样化的方式探索解决复杂问题的方法[51]。思维图的核心理念是模拟人类复杂的逻辑规划能力，允许大模型在推理过程中形成一个动态的思想网络。人类在思考时，往往会在不同的推理链之间回溯、合并和重组思想，这种灵活性是思维图所追求的目标。通过构建一个有向图，思维图能够有效地聚合和提炼不同的思维路径，从而生成更为准确和全面的解决方案。在思维图中，每个节点代表大模型生成的一个想法，而边则表示这些想法之间的逻辑关系。如图6.5所示，节点之间存在回溯、合并和重组的关系。这种结构使得思维图能够灵活地处理复杂问题，允许模型推理在多个思维链条上切换。

框架设计方面，思维图引入了“思维转换”的概念，指的是对图形结构的修改，这些修改可以是添加新的顶点或边。主要的思维转换类型包括：

- 聚合（Aggregation）：将多个想法合并为一个新的想法。
- 提炼（Refinement）：通过自我联系提炼思想中的内容。
- 生成（Generation）：基于一个想法产生多个新想法。

这些转换使得思维图能够在推理过程中不断更新和优化思想，最终得出更为准确的答案。

思维图通过引入图形结构和灵活的思维转换，显著增强了大模型解决复杂问题的能力。尽管其实现成本较高，但在处理可分解为子问题的任务时，GoT展现出了优越的性能。未来的研究可以进一步探索如何优化思维图的结构和实现，以提高其在实际应用中的效率和可行性。

6.4 自我一致性

6.4.1 基本概念

自我一致性框架[43]是一种验证模型输出正确性的方法，它模仿了人类“从不同的角度进行自我验证”的过程。比如解决复杂数学题时，人们常常会运用代数推导、几何验证或数值检验等多种方法交叉验证，当模型通过不同路径得出一致结论时，结果的可靠性会显著提升。

6.4.2 框架设计

自我一致性框架通过以下三个核心模块实现多角度的结果验证：

- 链式思维引导模块：采用链式思维提示策略，引导模型展现出清晰的推理过程。这类似于人类解决问题时的步骤分解，使推理过程更加透明和可追踪。
- 多样路径生成模块：摒弃传统“贪心搜索”的算法局限，通过采样机制从模型解码器生成多个不同的推理路径。这一步模拟了人类从不同角度思考问题的过程，确保验证的全面性。
- 一致性答案提取模块：对多个推理路径进行边际化处理，通过分析答案的一致性程度，选择最可靠的结果。这种汇总机制类似于人类在获得多个验证结果后，选择最具共识的答案。

6.4.3 应用案例

自我一致性策略通过结合链式思维引导、多样路径生成与一致性答案提取，有效提升了大语言模型在复杂推理任务中的稳定性与准确性。图 6.7展示了这一策略的整体流程。接下来结合具体问题，依次介绍链式思维引导、多样路径生成和一致性答案提取这三个关键环节的具体方法与作用。

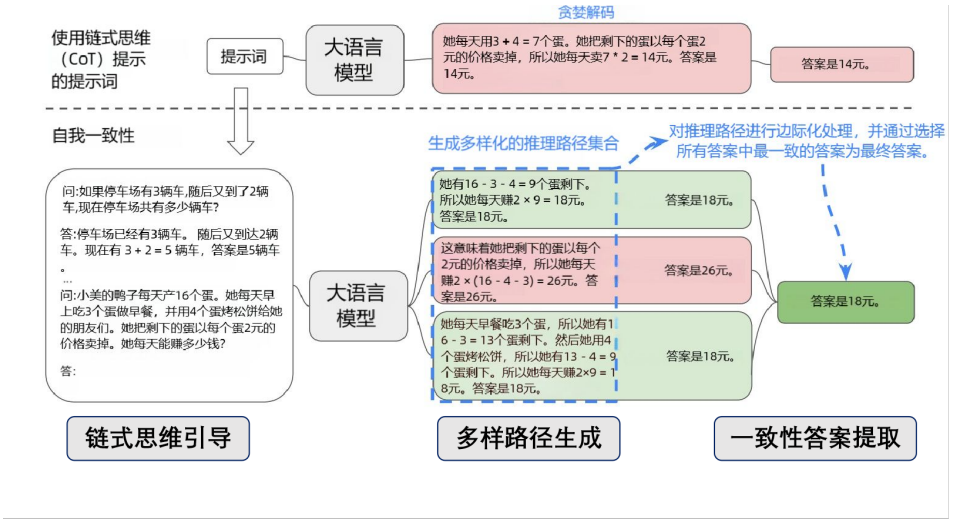


图 6.7: 自我一致性工作流程示意图

（1）链式思维引导

首先通过简单示例建立模型推理模式。以基础数学问题为例：

- 问题：停车场原有 3 辆车，新来 2 辆车，现有多少辆车?
- 答案：原有 3 辆车，新增 2 辆车，所以 $3 + 2 = 5$ 辆车。答案是 5 辆。

建立推理模式后，引入复杂问题验证策略效果：

- 问题：小美的鸭子每天产 16 个蛋。她每天早上吃 3 个蛋，用 4 个蛋烤松饼给朋友。剩余的蛋以每个 2 元价格售出。她每天能赚多少钱?

（2）多样路径生成

在自我一致性策略中，多样路径生成是核心环节之一。多样路径生成鼓励模型在回答同一问题时，沿着不同的推理路径独立展开推理。这种多样性可以通过引入随机性、不同的提示设计，或引导模型以多种角度思考来实现。通过生成多组独立的推理链，模型不仅能够覆盖更广泛的推理空间，还能够在一定程度上降低个别推理链出错的风险，为后续的一致性答案提取打下基础。以上述问题为例，多样路径生成结果如下：

示例 6.8 多样路径生成

路径 1（直接计算剩余）：16 个蛋 - 3 个（早餐）- 4 个（松饼）= 9 个剩余，最终得到 $2 \times 9 = 18$ 元；
路径 2（代数表达）：售卖数量 = 总数 - 使用数 = 16 - 4 - 3 = 9，最终收入为 $9 \times 2 = 18$ 元；
路径 3（逐步推理）：早餐后剩余：16 - 3 = 13 个，烤松饼后剩余：13 - 4 = 9 个，最终收入：$9 \times 2 = 18$ 元。

（3）一致性答案提取

通过比较多条独立推理路径的结果，模型能够识别出多个路径是否得出了相同的答案。假设在多样路径生成环节中，模型生成了若干不同的推理路径，其中路径 1、2、3 都得出了相同的答案（例如 18 元）。这种一致性表明这些路径的推理是可靠的，因此大模型可以选择将 18 元作为此问题的最终结果。通过这种方式，一致性答案提取确保了推理结果的准确性和稳定性。

如果采用的单一路径搜索策略（贪婪搜索），则模型会沿着单一推理路径逐词生成最可能的答案，得到的解答如下：

示例 6.9 单一路径搜索策略

解答：每天使用 3 + 4 = 7 个蛋，剩余蛋以每个 2 元售出，得到 $7 \times 2 = 14$ 元。

可以看到，单一路径搜索策略存在一定局限性。一旦在推理过程中出现小的偏差，最终答案（14 元）就可能错误。

自我一致性策略通过生成并验证多样化的推理路径，克服了单一路径搜索策略的局限，提升了语言模型在复杂推理任务中的表现。这种方法通过分析不同路径的结果分布，例如“18 元”在多个独立推理中反复出现，从而显著提升答案的可信度，类似于科学研究中的交叉验证。它不仅适用于数学问题，还广泛应用于自然语言处理、编程以及科学推理等领域，用于生成高质量文本、逻辑一致的代码及验证解答准确性，极大地提高了回答的准确性和用户信任度，并为高质量内容生成和逻辑一致性验证提供了有力支持。因此，该策略在自动问答系统、文本创作和程序开发等多个领域展现了重要的应用价值。

6.5 计划与执行

6.5.1 基本概念

计划与执行通常指的是如何根据当前的任务或者目标制定一个行动计划，并有效地执行这个计划。而在大语言模型中，计划与执行则是作为一种任务决策与推理的框架，该框架借鉴了人类在解决复杂任务时，会将任务进行分解规划并按步骤执行的思想。该框架让大模型能够通过系统化的任务分解和执行完成特定任务[44]。

- 任务分解：先让模型生成一个计划，明确任务的步骤，而不是直接给出答案。任务分解让模型形成清晰的推理框架，减少一步到位推理的难度。能够将复杂问题拆解成多个小问题，使模型能够更有逻辑地处理问题。
- 执行：模型会按照任务分解得到的计划一步一步进行执行，让执行具可解释性。通过计划，模型的注意力可以聚焦于当前具体的执行步骤，提高推理质量。通过这种方式，模型避免了直接从问题到答案的跨步推理，以更可控的方式完成计算。

直接让模型进行长链推理（如多步数学推理）可能会超出其短期记忆能力，导致错误。通过先思考，再执行的方式，将任务拆解，模型更容易处理复杂任务，提升了整体稳定性，减少了模型直接执行错误导致的误差，提高了对复杂任务的适应性。

6.5.2 框架设计

计划与执行强调预先规划，即在行动之前制定详细的计划，需要基于现有信息、历史数据等知识，预测未来可能发生的情况，并为这些情况设计解决方案。这种思维方式通常是线性的，先规划再执行，计划阶段和执行阶段相对独立。

- 计划阶段：当模型接收到一个任务或目标时，它先分析用户的需求，明确要达成的具体目标。随后，模型会生成并评估不同的行动方案，筛选一套具体的执行计划。
- 执行阶段：当计划完成后，模型会严格按照计划步骤逐步推进工作，每个步骤都是独立进行的。在模型执行具体步骤时，可能需要一些外部资源或输入，比如调用外部 API、工具或获取外部数据。在整个执行过程

中，除非遇到重大问题或偏差，否则不会对计划进行调整，确保整个过程稳定、有序地进行。

6.5.3 应用案例

本节以“分析一份租赁合同”作为案例，展示任务执行过程中的计划与执行两个关键阶段。

示例 6.10

以“分析一份租赁合同”作为案例进行分析，明确目标是安排会议，那么首先需要让模型生成具体的计划。

（1）计划阶段

计划阶段，主要通过分析提示词然后聚焦于任务目标的明确、任务步骤的拆解以及求解策略的制定，具体包括以下六个步骤：

1. 任务目标确认：解析租赁合同，提取关键信息包括会计信息、关键日期、条款详情等，并整理成结构化的表格，便于用户查阅。
2. 检查并解析 PDF 内容：确保文档格式正确，能够进行文本识别。
3. 识别关键信息类别：通过 OCR 技术或文本解析，识别租赁合同中的重要信息类别。
4. 创建结构化的框架：设计适合存储这些信息的结构化数据表，如 Excel 或数据库格式。
5. 提取具体信息：如租金金额、付款周期、合同起始日、租赁条款等。
6. 格式化整理信息：统一信息格式，使其更具可读性和分析价值，生成最终表格，并提供给用户。

（2）执行阶段

执行阶段围绕具体操作展开，依次完成 PDF 内容解析、关键信息提取、结构化数据处理，最终进行信息格式化与优化，然后输出结果。

1. PDF 内容解析：采用 OCR（光学字符识别）或直接解析文本层，获取合同中的原始内容，若文档存在扫描版或图片嵌入，则使用 OCR 识别文本。

2. 关键信息提取：使用 NLP（自然语言处理）方法，结合正则表达式或机器学习模型，提取关键信息，如租赁金额、合同起止日期等。

3. 结构化数据整理：依据提取的信息，按照预定义的表格格式进行整理。可能涉及数据清理，例如：纠正 OCR 识别错误字符、同一日期格式、解析金额数据，转换成标准货币格式等。

4. 信息格式化与优化：根据之前提取的信息，生成一个清晰、可读的 Excel 或数据库表，以供后续分析使用。也可能根据用户需求提供额外的分析，如预测未来支付趋势、识别即将到期的合同等。

通过预先指定计划并按照步骤执行的方式不仅提高了解决问题的系统性和可扩展性，还保证了解决方案的高效性、准确性。但由于计划阶段与执行阶段的独立性，在实际执行过程中可能出现意料之外的情况，导致原定计划失效，同时在执行过程中，前期步骤的错误可能会级联影响后续任务，导致最终答案严重偏离正确结果。为应对这一问题，ReAct 框架应运而生。ReAct 使系统能够在遇到突发情况时根据当前步骤结果自动重新评估并调整计划。通过这种方式，系统不仅能更好地应对变化的环境，还能确保会议安排的顺利进行，进一步提升用户体验和满意度。

6.6 ReAct 框架

6.6.1 基本概念

ReAct 框架与计划执行框架不同，ReAct 框架打破了“推理”与“行动”之间的严格界限，模拟了人类“边思考边行动”的动态过程[45]。在日常生活中，人们往往不会在完全想清楚后再采取行动，而是通过不断的观察、思考和调整来逐步解决问题。ReAct 框架正是借鉴了这一模式，允许系统在推理过程中动态地与环境交互，从而更灵活、高效地完成任务。这种设计理念体现在以下几个方面：

- 思维过程的动态性：类似人类在解决问题时会不断调整思路一样，ReAct 框架允许模型在执行过程中实时更新其推理策略。

- 行动反馈的即时性：模拟了人类在行动中学习的特点。每个行动的结果都会立即反馈给推理系统，让模型在尝试解决问题时不断从实践中获取经验，及时调整策略。

- 认知过程的整体性：该框架打破了计划与执行中推理和行动的割裂状态，实现了思考与行动的自然统一。

随着任务对模型动态交互能力要求的增加，ReAct 框架为大模型提供了一种与外部环境交互并解决问题的方式。

6.6.2 框架设计

ReAct 框架的核心优势在于其“推理-行动-反馈”(Reasoning-Acting-Feedback)闭环反馈机制（如图6.8所示）。该机制通过持续收集执行反馈信息，分析并优化后续决策过程。具体来说，模型先基于推理形成方案并执行，评估检查执行结果是否符合预期，若符合预期则输出，如不符合则优化方案，再次执行验证和反馈。这种设计使模型能够通过经验积累不断完善决策能力，在面对动态变化的复杂任务时展现出显著的适应性。具体来说，ReAct 框架的工作流程通常分为下列几个核心模块。

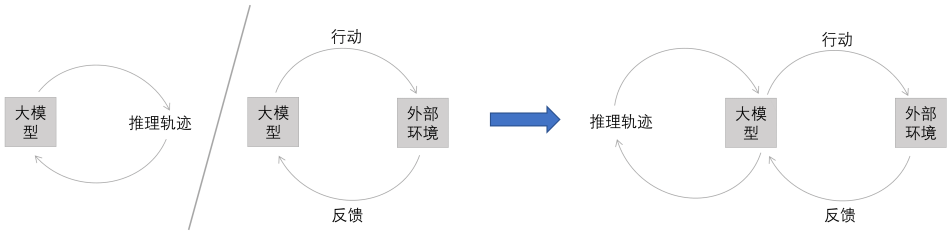


图 6.8: ReAct 框架概述图

- 推理（Reasoning）：推理是 ReAct 框架中任务处理的基础模块，其核心功能是分析任务要求并制定解决方案。在该阶段，模型会先从任务描述和上下文中提取关键信息。随后，模型基于逻辑推理和常识判断，生成具体的执行计划。该计划可能包括数据查询、数学计算或调用外部工具等步骤，为后续行动提供引导。通过推理过程，模型得以在充分理解问题本质的基础上，制定出合理的执行策略。
- 行动（Acting）：在行动阶段，模型根据前期制定的计划而行动。具体而言，模型可能需要与外部系统（如数据库、API 或其他 AI 系统）进行交互，获取额外信息或执行特定任务。这种实践性使模型能够在执行过程中获取实时反馈，并据此调整执行策略，从而确保任务的完成。
- 反馈（Feedback）：反馈与更新是 ReAct 框架的核心优化机制，遵循“推理-行动-反馈”原理。在执行行动后，模型会观察结果并获取反馈信息，

据此评估行动效果。当结果符合预期时，模型继续执行既定计划；若未达目标，则启动反馈机制，重新评估当前策略并进行必要调整。这种动态迭代的过程使模型能够持续优化决策路径，提升任务执行的准确性和效率。

ReAct 框架采用行动的循环机制。在每轮迭代中，模型基于执行反馈调整推理策略和行动计划，持续优化任务执行效果。通过不断积累经验和改进决策路径，模型逐步增强其环境适应性，确保在动态场景中实现高效可靠的任务完成。

6.6.3 应用案例

以 HotPotQA 问答系统[45] 中的一个问题为例，展示 ReAct 框架在实际推理与行动过程中的应用。假设 Agent 需要解答的问题是：“除了苹果遥控器，还有哪些设备可以控制苹果遥控器最初设计用来交互的程序?”。

在这一任务中，Agent 不仅需要基于已有知识进行推理，还需要通过与环境的动态交互来不断完善答案。为此，Agent 遵循了 ReAct 框架提出的“推理-行动-反馈”闭环机制，通过连续的推理、行动与环境反馈迭代，逐步接近正确答案。具体来说，这一过程体现为“思考—行动—观察”的循环：每一轮推理驱动一个具体行动，行动后基于观察到的反馈进行新的推理，从而不断调整和优化任务完成策略。

1. 案例流程

Agent 从问题出发，启动了“思考—行动—观察”的循环过程。具体执行轨迹如下：

示例 6.11 执行轨迹

问题：“除了苹果遥控器，还有哪些设备可以控制苹果遥控器最初设计用来交互的程序?”。

第一轮：确定初始程序

- 思考 1：我需要搜索 Apple Remote，找到它最初设计用来交互的程序。
- 行动 1：搜索 [Aplle Remote]。
- 观察 1：Apple Remote 是 Apple 于 2005 年 10 月推出的遥控器……最初设计用于控制 Front Row 媒体中心程序。

第二轮：查找控制设备

- 思考 2：Apple Remote 最初设计用于控制 Front Row 媒体中心程序。我需要接下来搜索 Front Row，找出其他可以控制它的设备。
- 行动 2：搜索 [Front Row]。
- 观察 2：找不到 [Front Row]。搜索包含有 Front Row 的其他内容，类似有 [Front Row Seat to Earth, Front Row (software), Front Row Motorsports]。

第三轮：确认详细信息

- 思考 3：找不到 Front Row。我需要搜索 Front Row (software)。
- 行动 3：搜索 [Front Row (software)]。
- 观察 3：Front Row 是一款已停产的媒体中心软件……

最终结论

- 思考 4：Front Row (software) 由 Appled Remote 或键盘功能键控制。所以答案是键盘功能键。
- 行动 4：最终结果 [键盘功能键] 。

2. 过程分析

从执行轨迹可以看到，Agent 经历了四轮“思考—行动—观察”的循环。在每轮循环中，Agent 首先确定一个明确的行动目标，随后根据目标采取行动，并根据行动反馈调整推理。接下来，对照“推理-行动-反馈”的闭环机制，逐轮分析每个阶段的执行过程。

（1）确定目标并开始搜索

- 推理：用户需要了解 Apple Remote 最初设计用于控制哪个程序。
- 行动：执行搜索，查询 “Apple Remote”。
- 反馈：根据搜索结果，得知 Apple Remote 最初设计用于控制 Front Row 媒体中心程序。

（2）进一步探索 Front Row

- 推理：既然知道了 Apple Remote 最初设计用于控制 Front Row，接下来需要了解 Front Row 是什么以及它还可以被哪些设备控制。
- 行动：执行搜索，查询 “Front Row”。
- 反馈：搜索结果不准确，返回了一些无关的条目，如 “Front Row Seat to Earth”、“Front Row Motorsports” 等。继续进行搜索，进入下一轮。

（3）精确搜索 Front Row（software）

- 推理：由于之前的搜索结果不准确，需要更精确地搜索 “Front Row (software)” 以找到正确的信息。
- 行动：执行搜索，查询 “Front Row (software)”。
- 反馈：这次搜索返回了正确的信息，确认 Front Row 是一款已停止更新的媒体中心软件。

（4）确定控制方式

- 推理：现在知道了 Front Row 是一款媒体中心软件，接下来需要确定除了 Apple Remote 之外，还有哪些设备可以控制它。根据已有信息，Front Row 可以通过 Apple Remote 或键盘功能键进行控制。
- 行动：得出结论，Front Row 不仅可以由 Apple Remote 控制，还可以通过键盘功能键控制。
- 反馈：最终答案是 “键盘功能键”。

3. 案例总结

在上面这个案例中，ReAct 框架从以下几个方面展示了其核心逻辑：

- 推理：每个思考步骤都基于前一个步骤的结果进行推理，逐步缩小问题范围，直到找到最终答案。例如，从 “Apple Remote” 到 “Front Row”，再到 “Front Row (software)”，最后确定控制方式。

- 行动：每个行动步骤都是基于推理结果采取的具体操作，通常是执行搜索以获取更多信息。例如，第一次搜索 “Apple Remote” 后，得知它用于控制 Front Row；第二次搜索 “Front Row” 时，发现需要更精确的查询；第三次搜索 “Front Row (software)” 后，找到了正确信息。

- 反馈：每次行动后，都会根据观察结果进行评估，决定下一步的行动。例如，第一次搜索 “Front Row” 时，返回了不相关的结果，因此需要进一

步精确搜索，经过不断尝试搜索，在第三次搜索 “Front Row (software)” 后，得到了正确的信息，从而结束了搜索过程。

整个过程中，推理和行动交替进行，形成了一个循环结构，直到问题得到解决。例如，从 “Apple Remote” 到 “Front Row”，再到 “Front Row”，最后得出 “键盘功能键” 的结论，整个过程不断迭代，最终得到正确答案。

6.7 反思

6.7.1 基本概念

反思框架模仿了人类 “吸取教训、总结经验” 的学习过程。就像人类在经历失败后会反思原因并记住教训一样，这个框架让大模型通过语言化的自我反思[2]来提升能力[46]。这种 “事后复盘” 机制与上一章介绍的 ReAct 框架形成鲜明对比——如果说反思是 “从错误中学习” 的导师，ReAct 则更像是 “边做边想” 的实践者。

具体而言，ReAct 框架侧重于任务执行中的动态协同，通过实时交替推理（如分析问题逻辑）与动作（如调用 API 验证信息），在过程中即时调整策略，减少错误传播和幻觉问题；而反思框架则聚焦任务后的迭代优化，通过分析失败原因生成自然语言反馈（如 “代码因变量未定义报错”），指导未来任务改进，尤其在需要多轮试错的场景（如编程调试）中，无需调整模型参数即可提升表现。两者分别解决“执行中可靠性”与“跨任务经验复用”问题，ReAct 框架强调在单次任务的高效执行与实时纠偏；反思框架擅长于从历史错误中提炼通用经验。

反思框架通过建立类似人类的自我评估和优化机制，使模型能够像人类一样对自身输出进行审视和改进，持续提升模型的认知水平和输出质量。这种方式不仅提高了模型解决问题的能力，还显著降低了模型的学习成本，推动模型通过不断总结经验来完善自身能力。

[2]语言化的自我反思通常指通过大语言模型的生成能力，自动进行推理、总结与调整。例如，模型在完成任务后，可能通过自然语言表达自己的思考过程，并分析哪些地方可以改进，进而调整其行为策略或任务执行方式。

6.7.2 框架设计

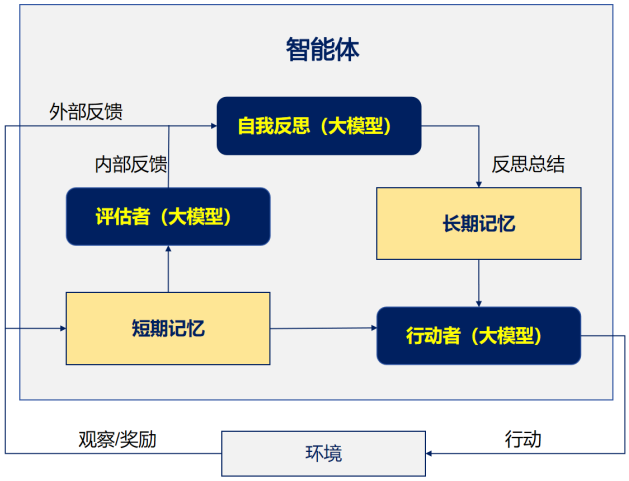


图 6.9: 反思框架图

反思框架模仿人类学习和思考的过程，如图6.9所示，该框架包含四个核心组成部分：

- 行动者 (Actor)：它建立在大语言模型基础上，负责根据当前情况做出具体的行动决策。这就像人类的 “行动系统”，在面对问题时，会根据当前状况采取相应行动一样。
- 评估者 (Evaluator)：根据 “评价指标” 评估行动结果 “好/坏”，就像人类在完成一项任务后，能够依据评价指标判断行动结果质量一样。
- 自我反思 (Self-Reflexion)：这是框架中最接近人类深度思考能力的部分。与简单的判断 “好/坏” 不同，它更像我们在失败后的深入总结。该模块要求不仅知道任务失败了，还会分析 “为什么失败”、“哪里做错了”、“下次该怎么改进”。比如在编程任务中，它不只是告诉你程序运行失败，还会像经验丰富的程序员一样，指出具体的错误位置和原因，并提供改进建议。
- 记忆系统 (Memory)：该设计分为短期记忆和长期记忆两种：
  - 短期记忆就像人类解决问题的 “工作记忆”，保存当前思考过程中的信息；

- 长期记忆则像人类的 “知识经验库”，储存着之前多次尝试后的反思总结。

这种记忆机制让模型能像人类一样，在思考问题时既能保持当前思路的连贯性（利用短期记忆），又能借鉴过往的经验教训（利用长期记忆）。正是这种类似人类的 “经验学习” 机制，让反思框架能够不断提升自己的问题解决能力。

6.7.3 算法流程

反思框架的运行过程模拟了人类 “尝试-反思-改进” 的认知循环，具体包含以下步骤：

1. 初始尝试 (Initial Attempt)：行动者开始与环境互动，进行第一次尝试，生成一个初步的解决过程。就像人类在面对新任务时会根据已有认知进行第一次尝试，ReAct 框架中模型也会记录下整个解决问题的过程。

2. 表现评估 (Performance Evaluation)：评估者根据一些预设的标准对初步的解决过程进行评分，给出一个奖励值。这个评分是一个数字，分数越高，说明执行得越好，类似于人类考试后的分数或工作评价。

3. 经验总结（Experience Reflection）：自我反思模块根据这个初步的解决过程和评分，生成一份总结，转化为具体的改进建议，并将这些反馈存储到记忆系统中。这个步骤类似于人类 “复盘总结” 的过程，把经验转化为可以帮助今后改进的知识。

4. 记忆存储（Memory Update）：每次的总结和反馈都会被存入记忆系统。由于处理能力有限，记忆系统只能存储一定量的信息，通常为 1 到 3 条，这就像人类的工作记忆，只能同时处理有限的信息量。

5. 持续优化（Iterative Improvement）：行动者、评估者和自我反思这三个模块会不断循环，进行多次尝试和反馈，直到评估者确认新的解决过程达到了目标要求。这个过程就像人类通过反复练习直到掌握某项技能。

6. 综合决策（Integrated Decision Making）：在做决策时，行动者不仅会考虑当前任务中的关键信息（短期记忆），还会利用自我反思产生的经验总结（长期记忆）来做决策。这种双重记忆机制就像人类在做决策时，既会根据当前情况做判断，也会参考过去的经验。

通过这种系统化的流程设计，反思框架实现了类似人类的 “经验学习-知识积累-能力提升” 的认知优化过程，同时保持了算法的严谨性和可实现性。

6.7.4 应用案例

本节通过两个典型任务案例，具体展示反思框架在实际推理过程中的应用效果，包括家庭环境下的多步骤决策任务和结合搜索的问答推理任务。

1. 案例一：家庭环境多步骤决策任务

以贴近日常生活的虚拟环境——AlfWorld[52] 为例，反思框架展示了如何通过 “尝试—反思—改进” 的方式，模拟人类在复杂的多步骤任务中不断调整和优化决策过程。

（1）实验设置

实验环境：AlfWorld 是一个基于 TextWorld 的交互式环境，专门用于测试人工智能模型在日常场景中的决策能力。这个环境的特别之处在于，它模拟了人类在家庭环境中处理日常任务的真实场景，让模型能够像人类一样通过语言理解和互动来完成任务。

任务描述：研究团队在 134 个不同的 AlfWorld 环境中部署了搭载反思框架的 AI 代理，要求其完成六类日常任务。这些任务包括但不限于以下三类：

- 寻找隐藏的物体（例如，在抽屉中找到特定工具）；
- 移动物体（如将刀具放置到砧板上）；
- 物品操作（使用工具来处理其他物品）。

任务设置：该实验环境的设置允许 AI 代理与不同的物体和工具进行交互，模拟真实的家庭生活场景，具体环境与任务设置如表6.1所示。

表 6.1: 家庭环境多步骤决策任务环境与任务设置

环境	
Agent 站在房间中央，通过 “视觉感知” 系统扫描周围环境，这种观察方式模拟了人类第一次进入陌生环境时的快速环境评估。房间物品如下:	
基础家具	床（1 个）、书桌（2 个，编号 1-2）、垃圾桶（1 个）和洗衣篮（1 个）
储物设施	抽屉（6 个，编号 1-6）、架子（6 个，编号 1-6）和架子（6 个，编号 1-6）
任务	使用台灯检查杯子。

（2）实验过程和结果

在反思机制的驱动下，Agent 根据第一次执行轨迹（轨迹 1）中的失败经验，调整策略，生成新的执行轨迹（轨迹 2）。

轨迹 1 的执行过程如下所示：

示例 6.12 轨迹 1

Agent 接到的任务是：找到一个杯子并使用一盏台灯。初次尝试过程如下：

第一轮：明确任务及目标

- 思考 1：为了解决任务，需要找到并拿起一个杯子，然后找到并使用台灯。首先，需要找到一个杯子。杯子更可能出现在抽屉（1-6）、书桌（1-2）、架子（1-6）、垃圾桶（1）、洗衣篮（1）里。可以从抽屉 1 开始逐一检查。
- 行动 1：前往抽屉 1，打开抽屉 1……打开抽屉 6。抽屉 6 已打开，里面有一个钥匙链 2。前往书桌 1，在书桌 1 上，看到一张信用卡 3、一盏台灯 1、一台笔记本电脑 2、一个杯子 1、一支钢笔 1 和一支铅笔 1。
- 观察 1：现在找到了一个杯子 1。接下来，需要拿起它。从书桌 1 上拿起了杯子 1。

第二轮：进一步对比任务及目标

- 思考 2：现在拿到了杯子 1。接下来，需要找到一盏台灯。台灯更可能出现在书桌（1-2）、床头柜（1-2）、架子（1-6）、床（1）、抽屉（1-6）里。可以从书桌 1 开始逐一检查。
- 行动 2：前往书桌 1，没有任何变化；前往书桌 2，在书桌 2 上，看到了一个闹钟 1、一个碗 1、一个杯子 3、三支铅笔 2 和一支铅笔 3。

最终结论

- 思考 3：现在找到了台灯 1。接下来，需要使用它。
- 行动 3：使用台灯 1，没有任何变化。
- 观察 3：状态：失败

虽然 Agent 成功找到了杯子和台灯，但由于执行顺序错误，最终任务失败，因此需要进行反思。

示例 6.13 反思

在这个环境中，Agent 的计划是先找到一个杯子，然后找到并使用台灯。然而，任务要求是用台灯检查杯子。所以应该先寻找台灯，然后再找杯子。注意到台灯是在书桌 1 上找到的。在下一次尝试中，将先去书桌 1，找到台灯，然后找杯子并用台灯检查它。

经过反思后，执行轨迹 2 的过程如下所示：

示例 6.14 轨迹 2

Agent 重新尝试过程如下：

第一轮

- 行动 1：前往书桌 1。在书桌 1 上，看到了一个信用卡 3、一盏台灯 1、一台笔记本电脑 2、一个杯子 1、一支钢笔 1 和一支铅笔 1。
- 思考 1：现在找到了一个杯子 1。接下来，需要拿起它。
- 行动 2：从书桌 1 拿起杯子 1。

第二轮

- 思考 2：为了解决任务，需要找到并拿起一个杯子，然后找到并使用台灯。
- 行动 3：使用台灯 1。打开台灯，然后用台灯 1 检查杯子 1。
- 观察 1：状态：成功

在以上的示例中，第一次尝试时，Agent 采用逐步的方法解决问题。但是最终并没能得到正确答案，因为 Agent 没完全理解任务，即需要使用台灯检查杯子，所以未能解决任务。因此 Agent 对自己的第一次尝试进行反思，意识到了初始策略存在缺陷，明白任务的关键是使用台灯检查杯子。最终在第二次实验时，成功完成了任务。

通过将行动、观察和反思结合在一起，Agent 不仅能够识别自己的错误，还能基于这些经验来优化未来的行为。并且在第二次实验中，Agent 是直接开始行动，反思框架中有一个记忆储存的机制，不需要在进行对问题的思考，而是直接读取以前的思考记忆就可以直接开始行动，大大减少了 Agent 的思考时间，提高了效率。

2. 案例二：问答搜索推理

为了验证反思框架在推理能力上的效果，本小节选择在 HotPotQA 上执行基于搜索的问答。

（1）实验设置

实验环境：HotPotQA 是一个基于维基百科的数据集，包含 113k 个问答对，挑战 Agent 解析内容并对几个支持文档进行推理。

任务描述：任务包括推理能力测试和综合问答能力测试。

- 推理能力测试：Agent 被给予问题（Q）以及来自数据集的真实上下文（C_gt），目的是为了单独测试代理的推理能力。使用“反思 +CoT”结合的 Agent。

- 综合问答能力测试：包括推理能力和行动选择两个方面。使用结合“ReAct与反思”的 Agent。

任务设置：Agent 被提供一个包含背景信息的上下文段落，并需要基于这些信息推理出正确答案，具体任务如表6.2所示。

表 6.2: 问答搜索推理任务设置

环境（上下文）	
白原战役是美国独立战争中纽约和新泽西战役的一场战斗，发生在 1776 年 10 月 28 日，地点位于纽约白原附近。在此次战斗后……	
任务	在 1776 年 10 月 28 日，在独立战争期间发生在纽约白原附近的，争夺纽约市和新泽西州控制权的一系列战斗是什么?

（2）实验过程与结果

在 HotPotQA 任务中，Agent 首先根据提供的上下文信息进行初步推理，生成第一条推理轨迹（轨迹 1）。随后，通过对推理过程进行反思，识别出初步答案与问题要求之间的偏差，并基于反思生成改进后的推理轨迹（轨迹 2）。下面分别展示轨迹 1 和轨迹 2 的推理过程与结果。

轨迹 1 的执行过程如下所示：

示例 6.15 轨迹 1

Agent 接到的任务是：推理争夺纽约市和新泽西州控制权的一系列战斗是什么。

推理过程

- 思考 1：这个问题是在询问关于 1776 年 10 月 28 日，在独立战争期间发生在纽约白原附近，争夺纽约市和新泽西州控制权的一系列战斗。从背景信息来看，我知道这就是白原战役。
- 行动 1：完成 [白原战役]
- 观察 1：答案不正确

可以看到，轨迹 1 中推理结果出现错误，未能准确覆盖问题要求的信息，因此需要进行反思。反思提供的错误答案后，Agent 意识到可能没有提供足够的背景信息来准确回答问题。

示例 6.16 反思

问题询问的是一系列战斗，但只提供了一个战役的名称。为了在未来更好地回答这个问题，会确保提供更多的背景信息，比如战役的名称——即纽约和新泽西战役，以及它是一系列战斗的事实。此外，还会确保包括战斗的日期和地点，以确保回答既全面又准确。

经过反思后，执行轨迹 2 的过程如下所示：

示例 6.17 轨迹 2

- 思考 1：这个问题是在询问关于 1776 年 10 月 28 日，在独立战争期间发生在纽约白原附近，争夺纽约市和新泽西州控制权的一系列战斗。根据背景信息，我知道这一系列战斗属于纽约和新泽西战役。
- 行动 1：完成 [纽约和新泽西战役]
- 观察 1：答案正确

在以上的示例中，题目有一定的误导性，因此在第一次实验的时候回答错误；从反思中可以看出，Agent 能够很好的认识到自己在之前的回答中没有考虑到的方面，在后面的回答中会回顾此次失败的原因，以防再次失败。

6.8 其它认知框架

6.8.1 蚁群智能

蚁群智能 (Ant Colony Intelligence)是一种基于自然界蚂蚁群体行为的认知框架，模拟蚂蚁在觅食过程中通过信息素进行通信和协作的机制。该框架通过模拟蚂蚁群体的集体智慧，解决复杂的优化问题，特别是在路径规划和组合优化领域表现出色。蚁群智能框架与大模型的认知框架相结合，能够增强模型在复杂任务中的协同搜索和优化能力。

1. 基本概念

蚁群智能框架借鉴了蚂蚁在寻找食物过程中所展现出的分布式、自组织和协作特性。蚂蚁通过在路径上释放和感知信息素来进行间接通信，信息素浓度高的路径更有可能被后续蚂蚁选择，从而逐步优化路径选择。与 ReAct 框架中的“推理-行动”动态交互类似，蚁群智能框架中的“信息素更新-路径选择”机制也体现了认知过程中的反馈与调整。蚁群智能框架的核心理念包括：

- 分布式决策：每只蚂蚁独立决策，基于局部信息和全局信息素分布进行路径选择，避免了集中控制的瓶颈；
- 正反馈机制：信息素浓度随着优质路径的选择而增加，增强了优秀解决方案的概率；
- 负反馈机制：信息素挥发机制防止过早收敛到局部最优，保持群体的多样性和全局搜索能力；
- 协同优化：通过群体中个体的协作，逐步逼近最优解，实现全局优化。

2. 框架设计

蚁群智能框架的设计包括信息素更新、路径选择和反馈机制三个核心模块，形成一个闭环的优化过程。具体工作流程如下：

- 信息素更新：信息素更新是蚁群智能框架的核心机制，决定了群体搜索的方向和效率。信息素的释放与挥发模拟了自然界中蚂蚁的信息素传递过程。在每次迭代中，蚂蚁根据路径质量释放信息素，优质路径的信息素浓度增加，而信息素挥发机制则逐步减少所有路径上的信息素浓度，防止过早收敛到局部最优。
- 路径选择：路径选择过程模拟蚂蚁在搜索食物时的决策过程。每只蚂蚁根据信息素浓度和路径长度的启发式信息，按照一定的概率规则选择下一步移动的路径。较短且信息素浓度高的路径被优先选择，从而引导群体向最优解靠拢。路径选择的概率公式通常结合信息素浓度和路径逆长进行计算，确保了搜索的有效性和多样性。
- 反馈机制：反馈机制在蚁群智能框架中起到调整和优化的作用。通过对路径选择和信息素更新的反馈，框架能够动态调整搜索策略，适应不同问题的需求。反馈机制不仅包括信息素的正负更新，还可以引入局部搜索和全局搜索的策略，以增强框架的优化能力和适应性。

蚁群智能框架通过三个模块的多轮迭代不断优化解决方案。在每轮迭代中，蚂蚁群体通过信息素更新和路径选择不断改进当前的解，并通过反馈机制进行自我调整。随着迭代的进行，群体逐步收敛到全局最优解或接近最优的方案。

3. 与大模型认知框架的结合

将蚁群智能框架与大模型认知框架相结合，可以构建一种混合模型协同框架，使多个大模型作为个体协同工作，从而涌现出更加高级的智能。这种协同框架借鉴蚁群智能中的分布式、自组织和协作特性，通过模拟蚂蚁群体的信息素机制，实现大模型之间的间接通信与协作。具体而言，结合方式包括：

- 混合模型协同框架：基于蚁群智能的分布式特性，设计一个由多个大模型组成的协同系统。每个大模型作为一个“个体”，独立处理部分任务，并通过共享全局“信息素”来协调和优化整体决策。这种框架避免了单一模型在复杂任务中的瓶颈，提升了系统的整体性能和智能水平。
- 信息素模拟与共享机制：借鉴蚂蚁的信息素更新机制，引入一种虚拟的信息素传递机制，用于大模型之间的信息共享与协调。每个大模型在完成任务后，根据其表现释放“信息素”，这些信息素会影响其他模型的决策过程，促进优秀策略的传播和优化路径的选择。
- 分布式优化与全局搜索：利用蚁群智能的全局搜索能力，对大模型的参数空间或决策路径进行分布式优化。通过多模型的协同搜索，避免单一模型陷入局部最优，提升整体系统在复杂任务中的搜索效率和优化能力。
- 动态策略调整与适应性增强：借鉴蚁群智能的反馈机制，设计大模型协同框架中的动态策略调整机制。根据系统整体的反馈信息，实时调整各个模型的工作策略，提高系统的适应性和鲁棒性，确保在动态和不确定环境中依然能够高效运作。
- 并行计算与资源优化：结合蚁群智能的分布式计算特性与大模型的并行计算能力，设计高效的资源分配与任务调度机制。通过并行处理和资源共享，提升系统的计算效率和处理大规模数据的能力，满足复杂任务对计算资源的高需求。

6.8.2 认知螺旋理论

1. 基本概念

认知螺旋理论的核心思想来源于人类认知过程中持续的交互、循环与迭代模式。与传统线性认知模式不同，认知螺旋理论强调“感知—思维—行为”三个过程在物理空间与认知空间中不断循环互动，并随时间推移而持续提升和进化。具体而言，人类通过感知获取外界环境的信息，随后进入认知空间进行思

考、抽象、联想与判断，继而在物理空间进行行动，验证和巩固已有认知。这种认知活动不是一次性的，而是通过反复的交互与反馈形成螺旋式迭代，不断提高认知的精确性、深度和有效性。认知螺旋理论具体体现为以下三个特点：

- 认知的交互性与开放性：强调感知和行动过程的持续互动，认知不断地与外界环境相互作用，确保认知系统始终保持动态开放的状态。
- 反馈机制的循环性与迭代性：每次行动产生的结果都会作为反馈再次进入感知环节，推动下一轮的思考与行动。
- 认知框架的自我修正与进化性：每次迭代都伴随着认知框架的调整、修剪和优化，确保认知系统随经验积累而不断进化。

认知螺旋理论不仅解释了人类如何不断地通过实践形成理论、再通过理论指导实践的过程，也为构建大模型之间的协作与增强提供了理论基础。

2. 框架设计

认知螺旋理论的框架设计（如图6.10所示）主要围绕“感知—思维—行动”的闭环反馈机制展开，其核心是不断的认知循环与迭代。

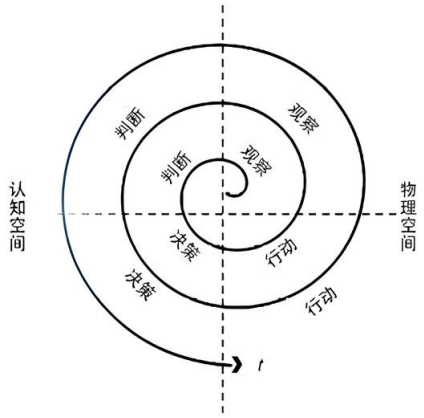


图 6.10: 认知螺旋框架示意图

- 感知：感知阶段是认知过程的起始环节，通过多种感觉器官联合获取外部环境信息，包括跨模态数据融合和信息过滤。模型通过这一阶段，实时捕获和分析外部环境变化，提供原始数据支持。这些数据进入认知空间后将作为后续推理和决策的基础，确保模型在实际环境中获取准确且丰富的信息。

- 思维：思维阶段是认知过程的核心，发生在认知空间内，主要负责将感知阶段提供的信息进行高阶抽象与分析。具体而言，该阶段涵盖逻辑推理、知识联想、记忆检索和创新思考，形成对当前情境的全面理解与整体判断。认知框架在此阶段进行自我更新与优化，以适应环境的变化，并为行动决策提供精确且灵活的指引。

- 行动：行动阶段是认知过程的最终体现和外化表现。根据思维阶段得出的决策方案，模型在物理空间中执行具体行动。行动结果立即反馈给感知系统，形成闭环反馈，推动下一轮迭代。行动阶段不仅仅是简单的结果实施，更是对思维策略的实际验证和调整过程，具有重要的反馈价值。

认知螺旋理论特别强调“感知—思维—行动”三个阶段的不断循环迭代。这种机制确保了模型在每次迭代中不断积累经验、修正错误、完善策略，从而持续提高模型的认知能力与环境适应能力。在这种持续循环中，认知系统不断提升其处理复杂、不确定问题的能力，实现认知能力的螺旋式增强。

3. 应用前景与案例

认知螺旋理论在大模型协同增强领域具有显著的应用潜力。通过多个大模型之间持续交互、反馈与协同，可以实现认知能力的自适应增强与动态优化。这种框架适用于自动驾驶、智能决策、机器人交互与多模态任务处理等复杂应用场景，推动人工智能系统迈向更高的智能化与泛化水平。

（1）自动驾驶系统的认知增强

自动驾驶车辆在复杂道路环境中，依靠感知系统获取实时交通信息，例如车辆位置、速度、行人动态等。这些感知数据在认知空间内通过逻辑推理和经验联想，形成行驶决策与路径规划。执行行动后，系统实时反馈实际车辆轨迹与环境变化信息，进入下一轮迭代，不断优化车辆控制与决策策略，实现车辆驾驶能力的持续提升。

（2）多机器人协作任务

在工业环境中，多机器人协作完成复杂任务时，每个机器人独立进行环境感知并在认知空间共享信息，形成协作策略。通过不断行动与反馈迭代，各机器人逐步调整行为，优化协作效率与准确性，整体提升团队完成任务的能力。

（3）智能医疗诊断系统

医疗诊断系统通过病人多模态数据感知获取病症特征，在认知空间中结合历史病例数据进行联想推理，提供诊疗建议。在实际治疗行动后，诊疗效果的数据再度反馈进入系统，推动下一轮认知迭代，持续提高诊疗方案的精确性和

有效性。

6.9 讨论

讨论 6.1. 你认为 ReAct 框架在复杂的决策任务中相较于传统的机器学习方法有何优势?

讨论 6.2. 自我一致性框架通过多角度验证来提高模型输出的可靠性，这与人类的多维度思考方式相似。请结合实际生活中的例子，讨论多维度思考在复杂决策中的作用。你认为在哪些领域中，多维度思考尤为重要？为什么?

讨论 6.3. 自我一致性框架借鉴了人类反复求证的认知特点。你认为这种模拟人类多维度思考的方式是否能够完全复制人类的认知过程? 在哪些方面可能存在局限性?

讨论 6.4. 比较人类学习过程中使用的反思机制与机器学习中的 Reflexion 算法。两者之间有哪些相似之处和不同之处?

讨论 6.5. 反思过程中的环境反馈（比如正确/错误的回答判定）对代理学习有何影响？讨论即时反馈与延迟反馈之间的差异及其对学习效果的影响。

6.10 习题

习题 6.1. ReAct 框架相比传统的单一推理模型，有什么优势?

习题 6.2. ReAct 框架的核心思想是什么?

A. 只依靠推理进行决策
B. 只依靠行动来完成任务
C. 推理和行动交替进行以优化任务执行
D. 使用固定的预定义规则来执行任务

习题 6.3. ReAct 框架中，“推理”和“行动”是如何协同工作的?

习题 6.4. 自动驾驶系统需要处理复杂的道路环境和决策问题。你认为在自动驾驶中是否可以使用自我一致性策略? 如果可以，请举例说明它如何通过多角度验证提高决策的可靠性；如果不行，请说明原因。

习题 6.5. 根据教材中提到的自我一致性验证过程的三个核心模块，请详细解释每个模块的功能，并举例说明它们是如何共同作用以确保结果的可靠性的。

习题 6.6. 假设你正在为一个新开发的智能代理设计其学习框架，请根据 Reflexion 算法的原则，设计一个完整的反思循环。这个循环应该包含哪些步骤？每个步骤的目的和预期输出是什么？

习题 6.7. 请描述反思框架中各个组成部分的功能。

第七章　大模型使用工具

7.1　概述

前述章节介绍了通过提示词、内容检索增强以及认知框架等提升大模型能力的方法，这些方法使得大模型可解决诸如多模态问答、代码生成和内容创作等一系列“简单任务”。但是，如何解决诸如数据分析、跨模态/多模态创作（图文并茂的文档生成）以及联网搜索等“复杂任务”呢？这些任务通常需要高度的跨领域协作与信息整合，仅依靠模型很难应对。

在此背景下，工具作为人类智慧与能力的延伸，发挥了至关重要的作用。大模型与外部工具的结合，正逐渐成为解决这些复杂问题的核心范式之一[53]。通过将大模型与工具有效融合，能够极大提升处理复杂任务的效率和精确度，从而推动更广泛的应用场景。例如，图2.5所示的 DeepSeek 网页版界面中，当用户打开“联网搜索”的功能时，大模型会规划调用 Web 搜索工具获取互联网中与用户问题相关的最新知识，进而生成符合用户需求的回答。除 Web 搜索工具外，为满足工业、工作和日常生活的智能化需求，开发者们设计了许多实用工具，如数据分析引擎、多平台实时搜索接口、多模态处理器等。大模型通过构建动态工具调度框架（如 ReAct，见第6.6章节），能够将用户的自然语言指令转化为对专业工具的精准调用，实现从静态的知识问答到动态的问题求解的跨越。这种“认知中枢 + 工具网络”的协作模式，既延伸了人类智能的边界，也推动了通用人工智能向开放场景的持续进化，为工业质检、科研分析、跨模态创作等复杂需求提供了智能化解决方案。

本章将首先通过典型应用示例展示大模型的工具选择和调用的能力；然后，介绍大模型调用工具的三种方式；最后，将列举一些常用的工具，以方便读者进一步了解“大模型 + 工具”模式在处理复杂任务时的强大能力。

7.2 工具调用能力

大模型能够根据用户提出的任务需求选择合适的工具并制定相应的解决方案。工具调用能力，指的是大模型通过分析任务来选择适用工具并提出解决方案的能力，即其作为 “认知中枢” 的体现[54]。在执行任务时，大模型首先对用户的自然语言文本输入进行语义解析和任务分解，然后在工具库中进行匹配，选择合适的工具。需要注意的是，在整个过程中，模型仅进行工具选择和策略规划，并不执行工具的功能，且能够根据工具返回的具体结果来调整工具的使用，进而优化任务的完成度。接下来，通过案例展示大模型选择工具和根据工具输出结果形成反馈继续调整策略的能力。

7.2.1 判断是否使用工具

如图7.1所示，用户预先声明了可用工具为“函数绘图工具”，大模型需根据后续任务内容，判断是否需要实际调用该工具。可以看到，大模型在具备工具调用能力的前提下，能够根据具体任务需求自主决策是否调用工具。即使用户已声明可用工具，模型也不会盲目使用，而是结合任务类型进行合理判断。



图 7.1: 大模型判断是否使用工具示例

7.2.2 选择合适工具

在多工具环境下，大模型可以根据任务的具体需求、可用工具以及外部因素，智能地选择合适的工具来完成任务，并且可以在接收反馈后继续选择与反馈相适应的工具来完成任务。如图7.2所示，尽管用户预先声明了多个工具（包括火车票和飞机票购票工具），但是模型通过分析用户的任务需求，判断出交通工具的选择可能受到天气的影响，优先选择调用天气查询工具获取相关地的天气预报。如图7.3所示，在获知相关天气信息后，模型通过分析选择受天气影响较小的出行方式，选择调用火车票购票工具完成任务。



图 7.2: 大模型选择适用工具示例



图 7.3: 大模型调整适用工具示例

通过选择使用工具以及根据反馈调整工具的调用方式，大模型不仅展现了智能化的任务处理能力，还显著提升了任务完成的效率和精准度。这种能力使大模型能够胜任复杂任务场景中各种任务，促进了大模型在不同领域的应用。

7.3 工具调用方式

目前，大模型调用外部工具主要有三种实现方式，包括：（1）配置工具 API；（2）工具通信协议；（3）GUI 模拟操作。

7.3.1 配置工具 API

通过配置工具 API 来实现调用外部工具具体而言指的是在大模型调用外部工具时，通过代码预先将外部工具的 API（如 HTTP/REST 接口或本地函数）以结构化描述的形式给模型，模型在推理过程中按需输出“调用指令”，让外部工具进行具体实现，再将结果回填给模型。在系统启动时，将外部服务（如 HTTP/REST API 或本地函数）以名称、功能描述及参数/返回值结构的形式向模型声明。模型在推理过程中若判断需要外部数据或功能，便输出一个“函数调用”对象，仅包含目标接口名称及序列化后的参数。后端接收模型的调用意图，调用相应 API 并将实际响应（JSON 等结构化数据）反馈给模型。模型基于接口响应生成最终的用户可读回答。

7.3.2 工具通信协议

1.MCP 通信协议

2024 年 11 月底，Anthropic 公司推出了模型上下文协议 (Model Context Protocol, MCP)开放标准协议[1]。其为开发者提供了一个强大的工具，能够在大模型应用和外部数据源（本地或远程数据、工具等）之间建立安全的双向连接。

MCP 等开放标准协议则统一了大模型和外部资源（包括外部数据源、工具以及预设提示词）之间的通信协议。其具有以下特点：(1）提供统一的标准化数据互联标准，降低了模型与多样化资源之间的连接复杂度，大幅度节省开发时间；(2) 建立开放的系统生态，协议允许任何开发者为其产品创建 MCP服务器，推动大模型与应用场景的融合；（3）内置严格的安全与权限管理机制，数据源所有者始终掌握访问权限，大模型在进行调用时需经过明确授权，避免数据泄漏和滥用问题。

MCP 使大模型应用能够安全的访问并操作外部数据，为其提供连接万物的接口。例如，本书第二章中所介绍的大部分常见大模型访问外部数据的方式，即用户主动将外部数据上传或者复制粘贴到大模型聊天框中。但是，这种交互方式既不方便也导致大模型无法实时访问多种数据源，进而导致大模型受到数据隔离而形成信息孤岛。而 MCP 服务器（MCP Server）则可实现大模型应用和数据的 “万物互联”。如图7.4所示，大模型应用可自动化地访问需求的数据源，解决数据隔离问题。

MCP 协议将大模型与外部数据资源之间的通信划分为 MCP 客户端（如工具、Claude Desktop、IDE 等，即大模型应用）、MCP 服务器与外部资源三个主要部分。客户端负责发送请求给 MCP 服务器，服务器则负责将这些请求转发给相应的资源。这种分层式设计使得 MCP 协议能够安全高效地控制外部资源的访问权限，确保只有经过授权的用户才能够访问该数据，避免数据泄露和滥用问题。MCP 的基本工作流程如图7.5所示，具体包括：

1. 初始化连接：客户端向服务器发送连接请求，建立通信通道；
2. 发送请求：客户端根据任务需求构建外部资源请求，并发送给服务器；
3. 处理请求：服务器接收到请求后，解析请求内容，执行相应操作（如读取文件系统、查询浏览器、读取文件等）；
4. 返回结果：服务器封装处理结果，并返回给客户端；

[1]MCP 协议官方文档：https://modelcontextprotocol.io/quickstart/server

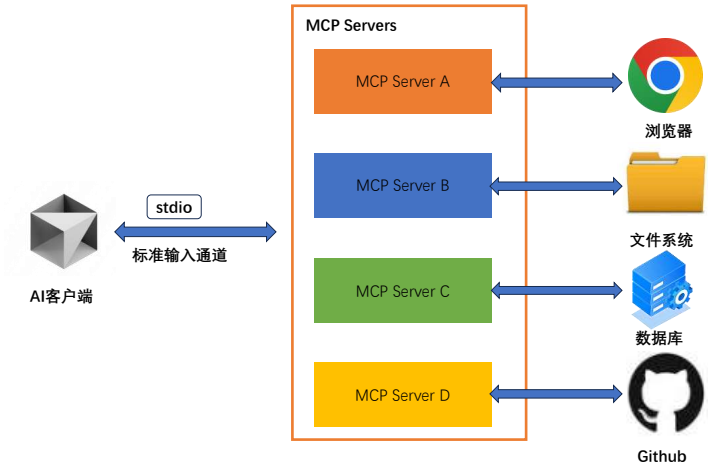


图 7.4: MCP 服务框架

5. 断开连接：任务完成后，客户端可主动关闭连接或等待服务器超时关闭。

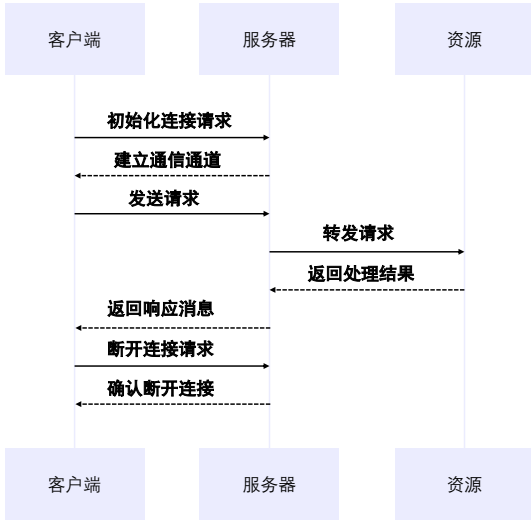


图 7.5: MCP 工作流程图

MCP 出现之前，开发者通常需要为每一个应用场景单独开发专用的函数或适配器（如函数调用），这种做法缺乏可扩展性，且费时费力。MCP 的发布解决了这一痛点。MCP 通过定义一个开放的标准化协议，允许任何开发者为其产品创建 MCP 服务器，在无需重复开发的情况下快速打通大模型与应用场

景的数据源连接，极大地提升了模型的通用性和解决复杂问题的能力。MCP 具有开放的生态、严格的安全权限管理机制以及统一的标准化协议，相信其会进一步降低大模型应用开发成本，极大地促进大模型应用的落地。

2.MCP 实战案例

MCP 作为 AI 生态的“USB 接口”，其“即插即用”的交互方式正在迅速推动大模型智能体的应用和发展。目前，阿里、百度、腾讯等国内公司正在联合推进 MCP 国内协议标准的开启，多家公司产品已支持 MCP 插件托管服务。如图7.6所示，Model Scope 汇总了丰富的 MCP 服务器、客户端以及应用案例，其中包括已上线的支付宝、高德地图、百度地图、无影等 MCP 服务器（`https://www.modelscope.cn/mcp`），感兴趣的读者可进一步浏览。本小节分别以大模型客户端 Cherry Studio 和 VScode 为例介绍 MCP 服务器的使用和开发。

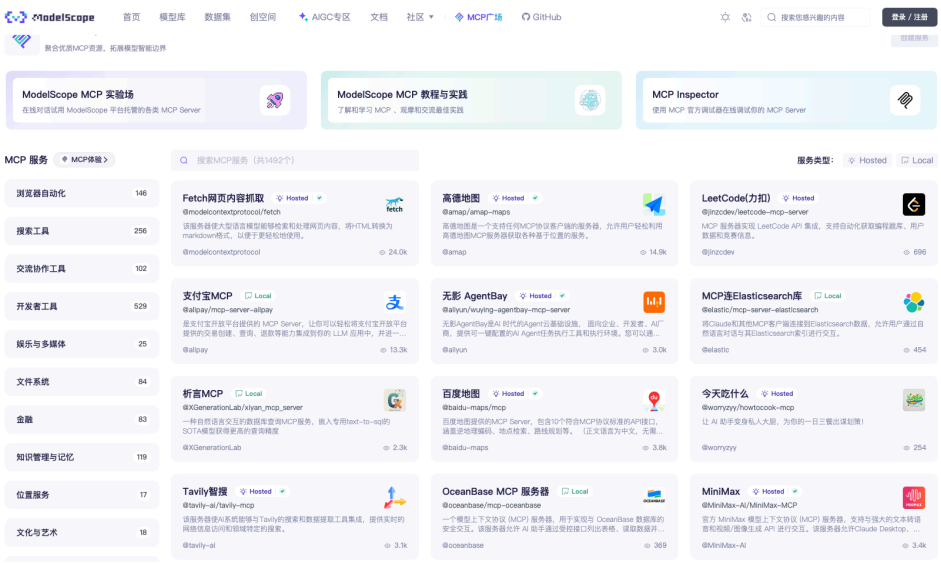


图 7.6: Model Scope MCP 广场

(1) 大模型客户端 Cherry Studio 配置 Fetch MCP 服务器 Cherry Studio 作为一个灵活的大模型开发平台，支持与 MCP 协议的深度集成，能够帮助开发者快速接入各种外部工具和数据源。

首先，需要确保 Cherry Studio 已成功安装并启动。用户可以通过官方文档中的指导步骤完成安装。启动 Cherry Studio 后，选择主界面左下角的设置，

再选择 MCP 服务器。然后，选择顶部的“搜索 MCP”选项后，可以看到一系列可供选择的 MCP 服务器，如图7.7所示。

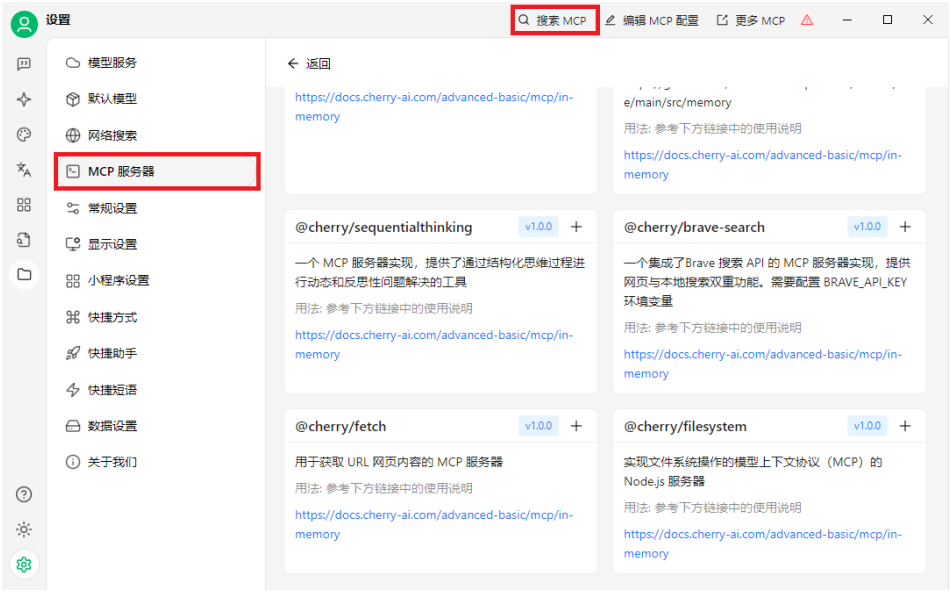


图 7.7: Cherry Studio 中可选用的 MCP 服务器

接着，点击需要的 MCP 服务右上角的“+”按钮，弹出“服务器添加成功”的字样时即可以使用该 MCP 服务。本书选用 @cherry/fetch 服务器作为示例，该 MCP 服务器用于获取 URL 网页内容。返回 MCP 服务器的主界面，发现已经成功添加 @cherry/fetch 的 MCP 服务，如图7.8所示。

在完成以上步骤后，回到左侧边栏最上面的助手选项。在对话框中选择 MCP 服务器，并选中刚才配置好的 @cherry/fetch MCP 服务器，如图7.9所示。

随后再选择支持 MCP 服务器的模型, 这里选择 DeepSeek-R1, 如图7.10所示。

至此，全部配置工作已完成。接下来可以测试大模型是否能够通过调用 Fetch 工具来解析指定链接地址，并对内容进行自动总结。

(2)VScode + Cline 配置高德地图 MCP 服务器

除了 Cherry Studio，VS Code 也是常用的大模型客户端之一，配合 Cline 插件，同样可以接入 MCP 协议，实现与外部工具和服务的连接。

首先需要安装 VScode, 安装完成后，打开软件在左侧找到 Extensions，然后搜索 Cline, 点击 Install 进行安装，如图7.12所示。

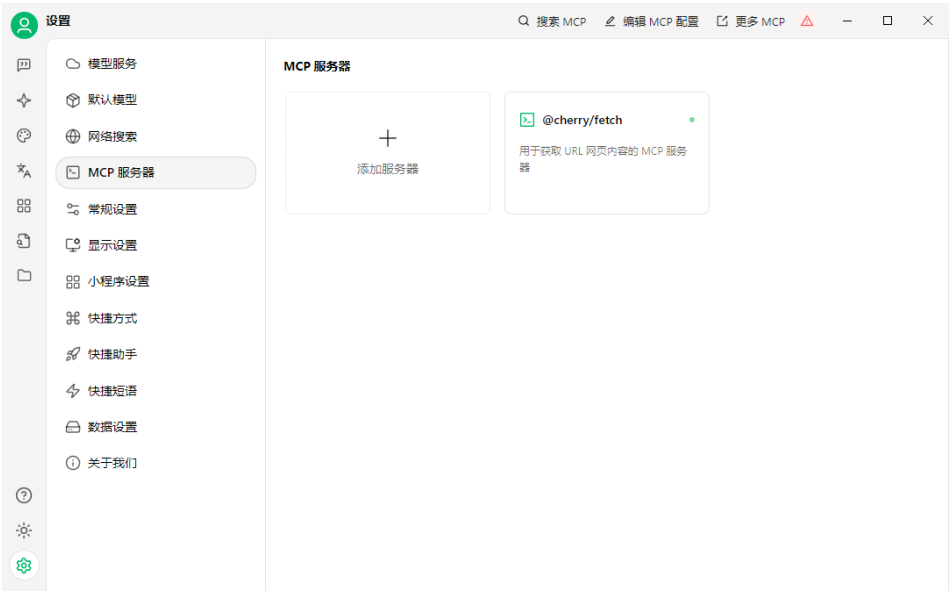


图 7.8: Cherry Studio 成功添加 MCP 服务器



图 7.9: 选择 Fetch MCP 服务器



图 7.10: 选择支持 MCP 服务器的大模型

安装完成后，在左侧找到 Cline 插件，接下来给它配置模型，点击设置模型提供商，选择 OpenRouter，搜索并选择 deepseek/deepseek-chat:free 来免费地调用 DeepSeek-V3 模型，如图7.13所示。



图 7.11: 测试大模型是否成功调用 Fetch MCP 服务器

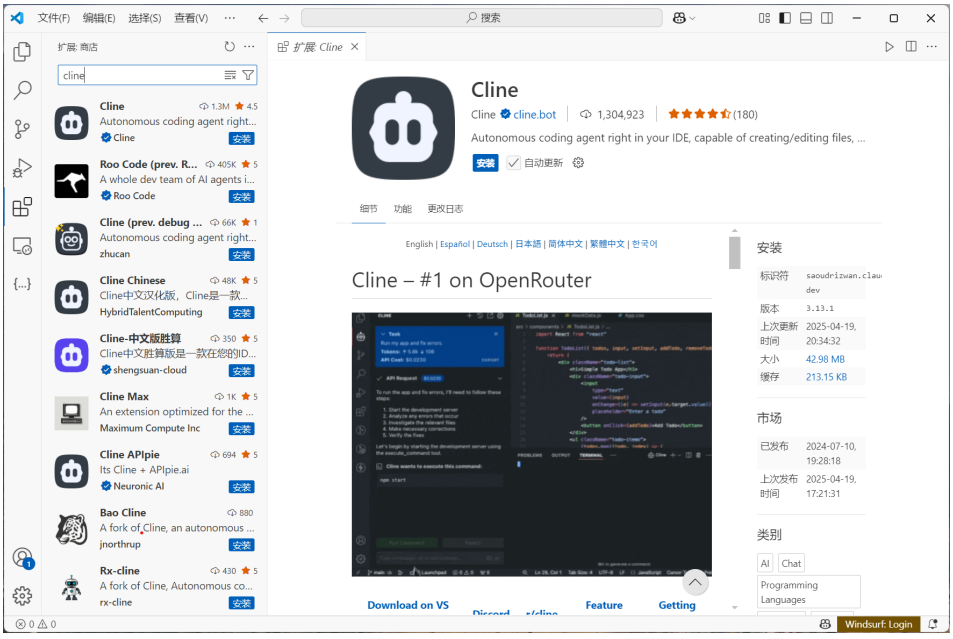


图 7.12: 在 VScode 软件中安装 Cline 插件

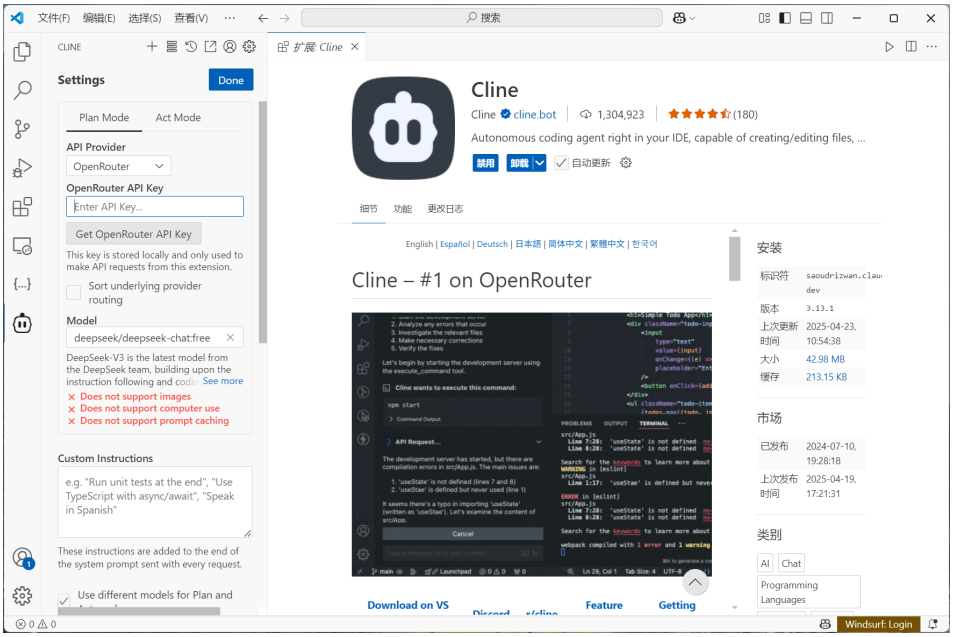


图 7.13: Cline 配置大模型

接下来, 需要获取一个 API(大模型调用凭证), 点击 Get OpenRouter API Key 后，在跳转后的网页上点击 Authorize，随后点击两次 open 返回 VScode, 即可获取 API 凭证，并完成 OpenRouter API 调用，如图7.14所示。

完成 OpenRouter API 调用后，需要检查一下 OpenRouter 的 API 调用是否成功。在页面底部的对话框输入：你好。若可以看到大模型的回复，则表明大模型的 API 已经配置好，如图7.15所示。

接下来,介绍通过编写 `cline_mcp_settings.json` 来安装高德地图 MCP 服务器。在此之前,请首先参考第二章安装 node.js。首先,在高德开放平台`https://console.amap.com/dev/key/app`注册开发者，并在控制台创建一个新应用，获取对应的 Key，如图7.16所示。获得 Key 后，在 cline_mcp_settings.json 中配置文件中填写以下相关信息，并保存。

```
"amap-maps": {
    "timeout": 60,
    "command": "cmd",
    "args": [
        "/c",
        "npx",
        "-y",
        "@amap/amap-maps-mcp-server"
```

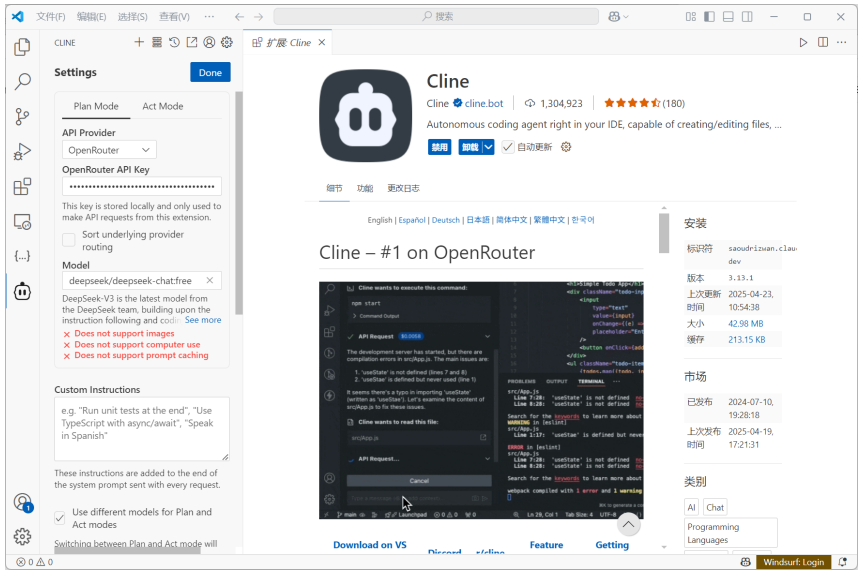


图 7.14: 配置大模型

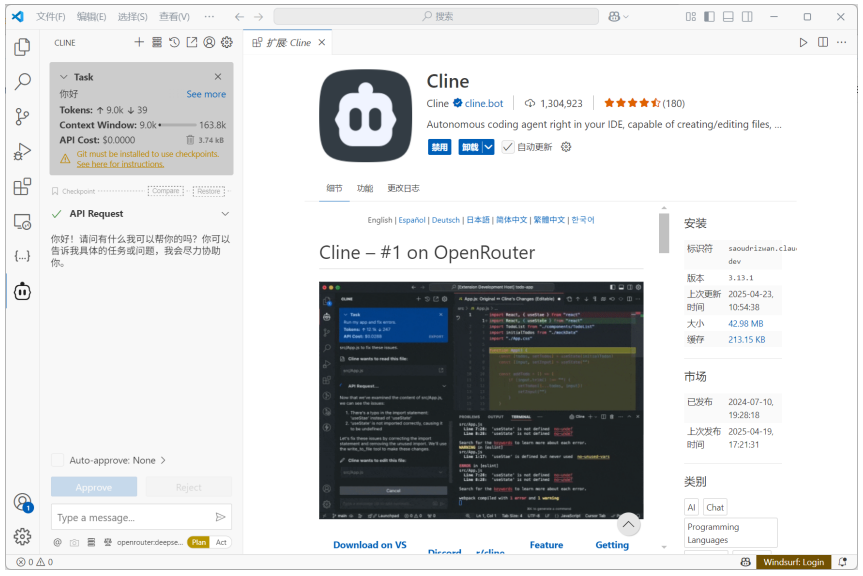


图 7.15: 测试大模型是否配置成功

```
    ],
"env": {
    "AMAP_MAPS_API_KEY": "<你的密钥>"
  },
```

```
    "transportType": "stdio"
}
```

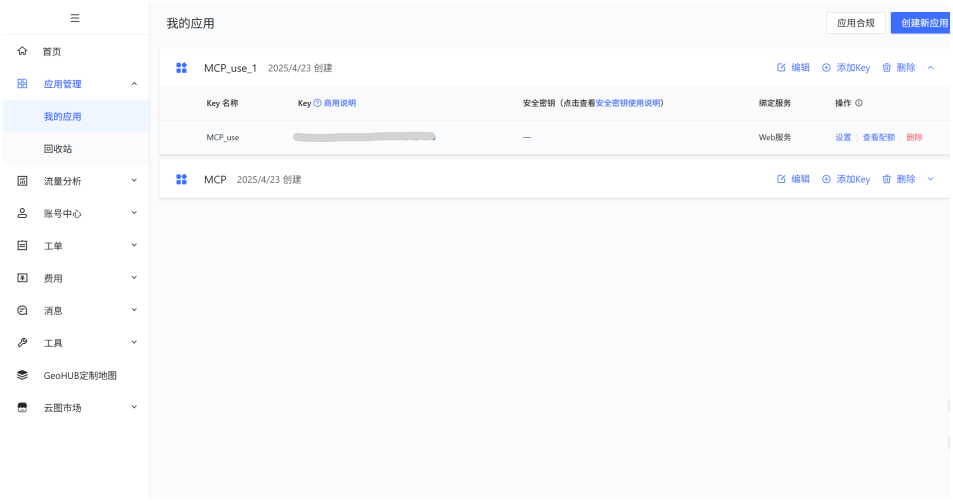


图 7.16: 获取高德地图 MCP 服务器 key

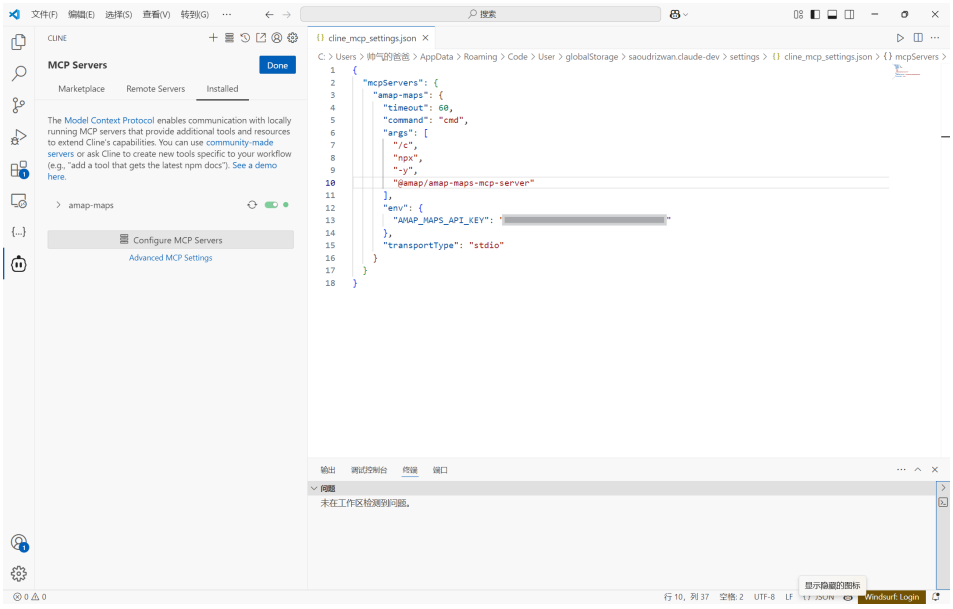


图 7.17: 高德地图 MCP 服务器安装

完成配置后，高德地图 MCP 服务器即完成安装，如图7.17所示。为了验证配置是否生效，可通过一个自然语言请求进行测试。例如，输入“明天我从

成都人民公园地铁站骑行到青龙湖，看下明天天气如何，以及给出具体的骑行路线。”模型会自动调用高德地图，并返回包括天气信息和骑行路径在内的完整回答，如图 7.18 所示。

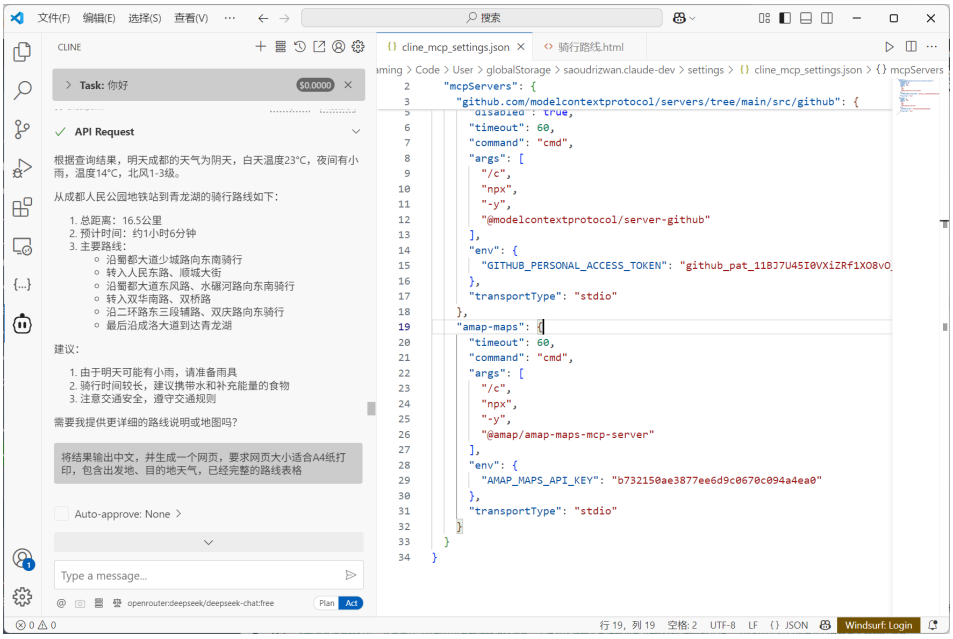


图 7.18: 测试高德地图 MCP 服务器是否被成功调用

用户可进一步提问让大模型将以上结果以中文输出并生成一个网页，要求包含出发地、目的地天气，以及完整的路线表格，界面如图7.19所示。

最后大模型将会按照要求将上一个问题给出的回答生成如图7.20所展示的一个包含出发地、目的地天气，以及完整的路线表格的 HTML 文件。

7.3.3 GUI 模拟操作

GUI（图形用户界面）模拟操作是指通过大语言模型（LLM）与多模态视觉语言模型（VLM）的结合，使智能体能够像人类一样感知图形界面中的元素（如按钮、菜单、输入框等），并根据自然语言指令完成点击、输入、拖拽等操作的技术[55]。这种技术将传统基于脚本或规则驱动的 GUI 自动化提升至智能化阶段，实现了从“固定流程执行”到“动态任务规划”的跨越，显著降低了用户操作复杂度和对特定 API 的依赖。

其核心价值体现为用户无需编写代码，仅需描述任务目标（如“将文件内容提取后发送给同事”），智能体即可自动分解任务并执行、通过视觉与结构化

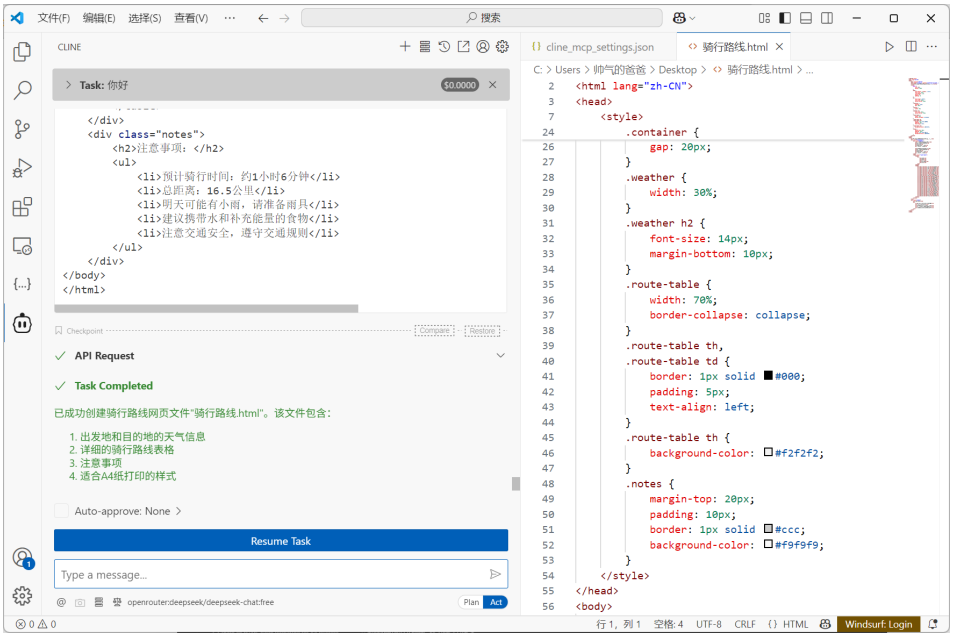
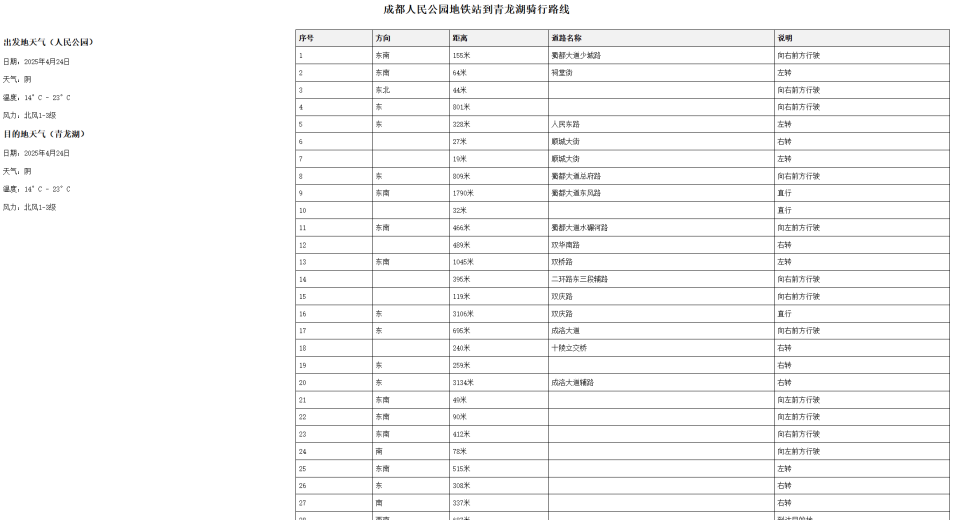


图 7.19: 生成旅游路线网页提示词



出发地天气（人民公园）

日期：2025年4月24日

天气：阴

温度：14° C - 23° C

风力：北风1-3级

目的地天气（青龙湖）

日期：2025年4月24日

天气：阴

温度：14° C - 23° C

风力：北风1-3级

成都人民公园地铁站到青龙湖骑行路线

序号	方向	距离	道路名称	说明
1	东南	155米	蜀都大道少城路	向右前方行驶
2	东南	64米	锦里街	左转
3	东北	44米		向右前方行驶
4	东	801米		向右前方行驶
5	东	328米	人民东路	左转
6		27米	顺城大街	右转
7		10米	顺城大街	左转
8	东	809米	蜀都大道总府路	向右前方行驶
9	东南	1790米	蜀都大道东风路	直行
10		32米		直行
11	东南	466米	蜀都大道水碾河路	向左前方行驶
12		489米	双华南路	右转
13	东南	1045米	双桥路	左转
14		395米	二环路东三段辅路	向右前方行驶
15		139米	双庆路	向右前方行驶
16	东	3106米	双庆路	直行
17	东	695米	成洛大道	向右前方行驶
18		240米	十陵立交桥	右转
19	东	259米		右转
20	东	3134米	成洛大道辅路	右转
21	东南	40米		向左前方行驶
22	东南	90米		向左前方行驶
23	东南	412米		向右前方行驶
24	南	78米		向左前方行驶
25	东南	515米		左转
26	东	308米		右转
27	南	337米		右转

图 7.20: 大模型调用高德地图 MCP 服务器生成旅游路线 HTML 文件

数据的融合，智能体可适配 Web、移动端、桌面系统等多种环境以及智能体能根据实时界面状态调整操作策略，例如处理弹窗或错误提示。

微软的 GUI 智能体提出了大模型驱动 GUI 智能体的通用架构，包括环

境感知、提示工程、模型推理、操作执行与记忆机制五大模块。例如，通过 Windows UI Automation 或 Android Accessibility API 获取界面元素属性，结合多模态模型解析用户指令并生成操作序列。清华大学的 CogAgent 基于视觉语言模型，支持高分辨率图像输入，采用“高分辨率交叉注意力模块”降低计算开销，在 Mind2Web 等数据集上实现了跨平台的高效操作。ScreenAgent 通过强化学习环境与反思机制，模拟真实计算机屏幕交互，支持多步骤任务的动态规划与执行，其 UI 定位精度甚至超越 GPT-4V。GUI 模拟操作的实现流程如下：

1. 输入包括屏幕截图、UI 元素属性（类型、位置、标签）、窗口层级信息等信息，通过系统 API（如 Android Accessibility API）或视觉模型解析。

2. 结合文本指令与视觉信息，例如 CogAgent 将用户目标与高分辨率截图联合输入模型，生成操作坐标与动作类型。

3. 通过 LLM 将用户指令分解为原子操作步骤（如“点击搜索框 → 输入关键词 → 提交”）。

4. 将当前 GUI 状态与用户指令结合，例如“用户需发送邮件时，智能体自动定位‘发送’按钮并触发点击”。

GUI 模拟操作标志着人机交互从“手动操作”向“自然语言智能控制”的转型。未来，随着多模态模型与强化学习的深度融合，智能体将更精准地理解复杂界面，并在教育、医疗、工业等领域实现更广泛的应用。然而，其发展仍需解决环境干扰、伦理规范等核心问题，以构建安全可靠的智能化交互生态。

7.4 常用工具

7.4.1 检索工具

大语言模型可通过调用并集成智能搜索工具构建动态知识补给系统，利用检索增强系统（Retrieval-Augmented Generation, RAG）整合多源异构数据库与动态网络爬取技术，能够有效突破大模型内部知识的时效性问题。关于 RAG 技术的详细介绍参见本书的章节五。先进的检索工具可以帮助提升工作效率，解决大模型知识时效性的问题，减少大模型幻觉；同时，大模型还可通过语义搜索和个性化推荐，优化用户的检索体验。表格 7.1列举了一些当前广泛应用的信息检索工具。这些工具功能多样，可用于从不同平台检索相关信息。例如，

Perplexity AI[2]是一款由人工智能驱动的对话式搜索引擎，旨在提供准确、实时且可验证的答案；类似的搜索工具还有由昆仑万维推出支持多模态检索的天宫搜索[3]以及注重个性化搜索体验的 Globe Explorer[4]；除此之外，Consensus[5]是学术论文检索方向的 AI 搜索引擎，可以根据检索的内容生成用户感兴趣的研究方向的总结报告。

表 7.1: 检索工具

工具名称	功能简介
Perplexity AI	对话式搜索引擎，旨在提供准确、实时且可验证的答案
天宫搜索	智能搜索引擎，可以提供精准、智能和多模态的搜索体验
Globe Explorer	注重个性化搜索体验，致力于提供高质量的搜索结果
Consensus	学术论文检索引擎，其可根据检索内容，提供相关论文并进行总结

7.4.2 学习教育

在教育数字化转型的背景下，智慧教育工具已成为提升教育质量和效率的重要手段。大模型通过调用学习教育工具可实现三大核心能力的跃迁：（1）个性化学习路径规划；（2）多模态教育资源自动化生成与适配；（3）跨语言知识传递。此类工具的使用可大幅度提升教师教学设计效率，并提升学习者的知识掌握度。

表格 7.2列出了部分典型的学习教育工具。Scispace[6]集成了公式理解模块，提供科研文献深度解析功能，并生成总结报告；ChartGPT[7]可根据用户需求从互联网爬取数据并自动生成图表；Copilot[8]为开发者提供代码提示和修改等功能，大幅度提升代码编写和调试效率；Docanalyzer[9]是一个多模态文档处理工具，其可处理 PDF、Word 和 Excel 等文档，并进行数据提取分析、内容总结和摘要生成等。DilGPT[10]是一个多语言对话训练平台，其可进行多语种聊天；

[2]Perplexity AI 网址：`https://www.perplexity.ai/`

[3]天宫搜索网址：`https://search.tiangong.cn/`

[4]Globe Explorer 网址：`https://explorer.globe.engineer`

[5]Consensus 网址：`https://consensus.app/search/`

[6]Scispace 网址：`https://scispace.com/`

[7]ChartGPT 网址：`https://www.chartgpt.dev/`

[8]Copilot 网址：`https://Github.com/features/copilot`

[9]Docanalyzer 网址：`https://docanalyzer.ai/`

[10]DilGPT 网址：`https://dilgpt.com`

Clrblt learning groups[11]是一个远程小组学习平台，为用户提供个性化的学习建议和指导。

表 7.2: 常见学习教育工具

工具名称	功能简介
Scispace	论文阅读工具，提供了强大的功能，如论文总结，公式解析等
Docanalyzer	支持 PDF、Word、Excel 的 AI 分析，可提取数据、总结内容、生成摘要
ChartGPT	根据用户提出的问题从互联网上寻找相关数据，自动生成表格
Copilot	Github 提供的代码提示功能
DilGPT	聊天机器人，帮助你进行语言学习
Clrblt learning groups	远程小组学习，个性化定制学习路线

7.4.3 社交娱乐

社交娱乐是人们日常生活的重要组成部分，正在受到大模型技术的深刻影响。通过个性化推荐、智能互动和内容生成，大模型显著提升了用户的娱乐体验。例如，大模型能够根据用户的兴趣和行为数据，个性化推荐电影、音乐、游戏等内容。同时，大模型能够通过自然语言交互增强社交互动的趣味性，帮助用户发现新的兴趣点，丰富内容形式，推动文娱产业的智能化发展。此外，大模型还可以与用户进行对话，增强社交互动的趣味性和参与感。

表格7.3列举了一些有趣的社交娱乐工具。Motionshop[12]能根据上传的视频自动检测视频中的人物并替换成 3D 卡通角色模型；Supermeme[13]能够快速生成各种有趣的表情包；VMate[14]是一款角色扮演聊天机器人应用；虚拟 AI 小镇[15]是由斯坦福大学设计的一个多智能体虚拟社区，其通过多智能体强化学习框架模拟 25 个智能体的复杂社交活动。

[11] Clrblt learning groups 网址：`https://www.clrblt.group`

[12] Motionshop 网址：`https://modelscope.cn/`

[13] Supermeme 网址：`https://app.supermeme.`

[14] VMate 网址：`https://vmate./`

[15] 虚拟 AI 小镇网页：`https://github.com/a16z-infra/ai-town`

表 7.3: 社交娱乐工具

工具名称	功能简介
Motionshop	根据上传的视频自动检测视频中的人物，并替换成 3D 卡通角色模型
Supermeme	快速生成各种有趣的表情包
虚拟 AI 小镇[56]	多 AI 智能体的虚拟小镇。
VMate	角色扮演聊天机器人应用，可以让人们与角色扮演机器人聊天

7.4.4 图像音视频处理

在多媒体内容需求快速增长的背景下，图像、音频与视频的处理成为大模型的重要应用领域。大模型可以利用相关的工具来实现高效的图文、音频、视频的相关处理，不仅优化了多媒体内容的创作与管理流程，还推动了各行业在智能化和数据驱动决策方面的进步，满足了现代社会对高质量、多样化多媒体内容的需求。

表格 7.4列举了一些常见的图像、音频与视频处理工具。Midjourney[16]可根据文本生成图像；Animate Anyone[17]能从静态图像生成角色视频；Autopod[18]是一款智能视频剪辑工具；Sora[19]支持根据文本指令创建现实且富有想象力的视频场景；Mubert[20]是一款驱动的音乐生成工具，提供原创音乐创作服务；ElevenLabs[21]是一款可以提供高质量的语音克隆服务，支持多种语言，适用于语音合成、配音和语音恢复的音频工具。

[16]Midjourney 网址：https://www.midjourney.com/

[17]Animate Anyone 网址：https://humangc.github.io/animate-anyone/

[18]Autopod 网址：https://www.autopod.fm/

[19]Sora 网址：https://open.com/sora/

[20]Mubert 网址：https://mubert.com/

[21]ElevenLabs 网址：https://elevenlabs.io/

表 7.4: 图像音视频处理工具

工具名称	功能简介
CodeFormer	模糊图片修复技术及应用实现
Midjourney	文生图像
Animate Anyone	从静态图像生成角色视频
Autopod	智能视频剪辑工具
Sora	文生视频
Mubert	文生音乐，提供音乐收听和制作等功能
ElevenLabs	语音合成、配音和语音恢复等。

7.4.5 日常生活

大模型可以与健康与健身、交通出行、购物推荐和娱乐安排等多个方面的工具结合发挥作用，提升人们生活的舒适度和效率，实现智能化的生活方式。例如，智能家居设备之间的的管理与协调就是一个典型应用，大模型可与全屋智能家居设备的协同互动，可根据用户的习惯自动调节环境设置，提供个性化的居家体验。

表格 7.5中列举了一些常见的日常生活工具。Calmify[22]是一款心理健康伴侣，可帮助用户改善心理健康；Reflectfit[23]提供健康与健身的个人辅导，监测和纠正动作；Eat Smart[24]是免费的饮食计划生成器；cotravel[25]为用户提供个性化旅行行程规划；Eat Smart[26]是免费的饮食计划生成器；OneAir[27] 是一款 AI 驱动的旅行助手，能够实时搜索酒店、航班、餐厅信息，并提供天气预报等功能。

[22]Calmify 网址：`https://www.calmify.`
[23]Reflectfit 网址：`https://www.reflectfit.com`
[24]Eat Smart 网址：`https://eatsmart.life`
[25]cotravel 网址：`https://cotravel.com/`
[26]Eat Smart 网址：`https://eatsmart.life`
[27]OneAir 网址：`https://www.oneair.ai/`

表 7.5: 日常生活工具

工具名称	功能简介
Calmify	心理健康伴侣，帮助你改善心理健康
Reflectfit	健康与健身应用，提供个人辅导，监测和纠正用户的动作
Eat Smart	免费的饮食计划生成器，通过个性化的饮食计划
Cotravel	旅行行程规划工具，为用户提供个性化的旅行行程规划
OneAir	提供实时搜索酒店、航班、餐厅信息，并提供天气预报等功能

7.4.6 金融商业

在金融和商业领域，大模型通过工具来进行数据分析、决策支持、客户服务以及市场分析与预测，可帮助企业识别投资机会和潜在风险。这种” 大模型 + 金融工具调用”的方式不仅可以推动金融行业智能化，协助企业进行投融资活动；还可以帮助个人投资者打破信息壁垒，优化个人投资组合。例如，在股票交易中，模型可以分析历史数据和实时市场信息，生成精准的交易信号，优化投资方式。

表格 7.6列举了一些常见的金融商业工具。BeeBee[28]是一款财报分析工具，其针对美股提供财务报告和收益电话会议的解读；Vestinda[29]可利用自动化交易策略来管理和增长投资组合；Swallow[30]是一个专为金融定价设计的平台，帮助快速将金融模型推向市场。

表 7.6: 金融商业工具

工具名称	功能简介
BeeBee	美股财报分析平台
Vestinda	通过自动化交易策略来管理和增长投资组合
Sermo	为银行和金融科技公司提供多语种客户支持服务
Swallow	金融定价平台

[28]BeeBee 网址：`https://www.beebee.`

[29]Vestinda 网址：`https://www.vestinda.com`

[30]Swallow 网址：`https://swa.llow.io/`

7.4.7 其它工具

除了上文介绍的工具外，还有一些其它类型的工具。这些工具可以与大模型平台直接集成，提供 API 接口供模型调用。通过这种智能化的协作，大模型能够帮助用户实现高效、便捷的操作，进一步提升工具在各个领域的应用潜力。AIGC 工具导航平台[31]汇集了大量具备开放接口的工具资源，便于开发者查找与集成，界面如图7.21所示。

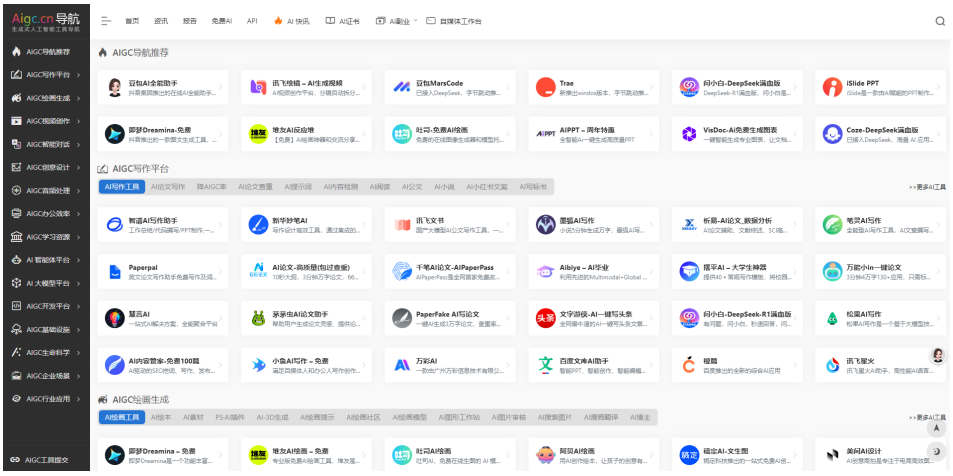


图 7.21: AIGC 工具导航平台

除了以上介绍的不同类型工具外，AI 工具集也是一个汇集众多 AI 工具的综合平台，官方主页[32]如图7.22所示。网站涵盖多个领域，包括 AI 写作、图像生成、视频制作、音频处理、辅助编程、音乐生成、绘画设计、对话聊天等。如果在前面已经分类好的工具中未找到合适工具，可以按需去该 AI 工具集合寻找，以提高效率。

[31] AIGC 工具导航平台:https://www.aigc.cn

[32] 工具集官网：https://ai-bot.cn

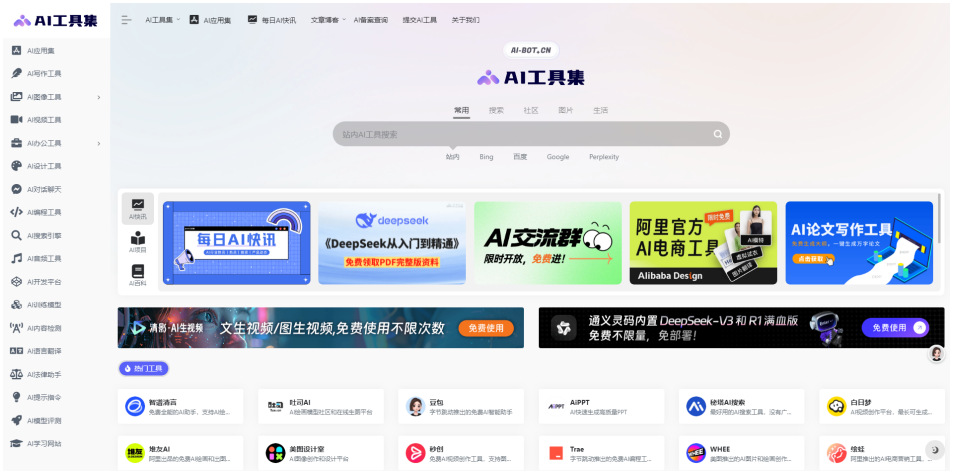


图 7.22: 工具集

7.5 讨论

讨论 7.1. 请讨论大模型为什么具有选择调用工具的能力。

讨论 7.2. 请讨论中国设计一套自主的大模型工具接口的必要性。

7.6 习题

习题 7.1. 简述大模型通过工具扩展能力的优势，举例说明工具如何帮助大模型解决问题。

习题 7.2. 简述大模型调用外部工具时的基本流程，包括请求生成和结果处理。

习题 7.3. 大模型在工具调用失败时，可以采用哪些策略与用户保持良好交互?

习题 7.4. 为大模型设计工具时，工具适配性应该满足哪些要求？请列出三点并说明理由。

习题 7.5. 大模型在调用工具时可能涉及用户数据隐私问题，请说明大模型如何在工具调用中保护用户隐私。

习题 7.6. 简述工具调用效率对用户体验的影响，并说明如何优化工具调用的效率。

习题 7.7. 结合大模型的能力扩展需求，展望未来可能出现的新型工具，并简述其应用场景。

习题 7.8. 列出大模型常用工具的类别，并简要描述每类工具的主要功能和特点。

习题 7.9. 为什么检索工具在大模型的应用中十分重要？请结合一个实际场景说明。

习题 7.10. 信息处理工具在大模型扩展能力中的典型应用有哪些？请举例说明。

第三部分

大模型高阶应用：AI 智能体

第八章　AI 智能体概述

8.1　概述

AI 智能体（AI Agent）是一种能够感知环境、规划决策并自主行动的计算机系统或实体，大型语言模型通常作为其核心决策和处理单元，具备独立思考和调用外部工具去逐步完成给定目标的能力。AI 智能体系统主要包括单智能体系统和多智能体系统。单智能体系统就像 “独行侠”，专注于独立完成各种任务。例如，OpenAI 的 ChatGPT 在大模型的基础上增加 Web 搜索等助手组件，能够化身 “资料查询师”，帮助用户找到其所需的互联网内容；百度的小度和小米的小爱同学，则集成了语音识别、家居控制等组件，化成千家万户的 “智能管家”。而多智能体系统则更像一支 “特工小队”，强调团队协作与默契配合。例如，在智慧城市管理系统中，部分智能体用于实时监测城市状况，并收集信息；部分智能体用以决策和判断，负责分析数据并提出解决方案；另外一些则直接执行实时任务，例如调整信号灯时间或发布警报等。通过多智能体构建的智慧传感器网络、数据分析中心以及自动化操作部门，城市管理者能够更有效地处理交通流量、公共安全等问题。

本章将全面拆解 AI 智能体，从基本架构到核心概念，从功能特性到应用案例，之后提炼当前 AI 智能体设计的五个关键设计原则，助力打造出更优质、更具竞争力的智能体产品，帮助读者了解 AI 智能体的基本原理和应用方法。

8.2　AI 智能体构成

8.2.1　四大核心组件

单智能体系统框架包含规划（Planning）[57]、记忆（Memory）[58]、工具（Tools）[53]和执行（Action）[57]四大核心模块，如图8.1所示[1]。

[1]图片来自 Lilian Weng 博客，参见https://lilianweng.github.io/posts/2023-06-23-agent/。

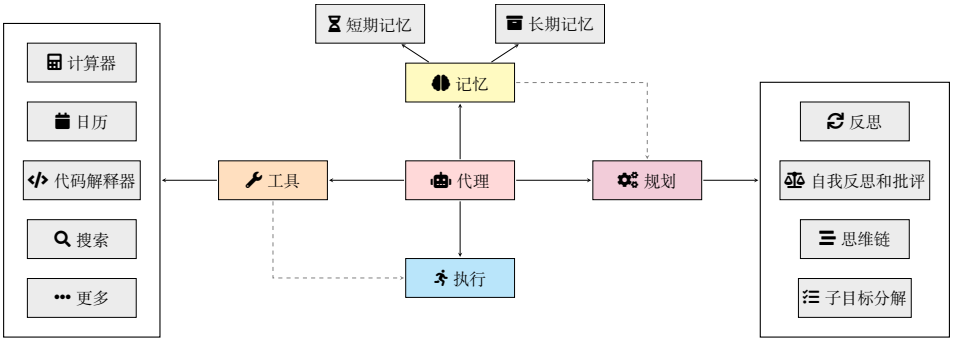


图 8.1: 智能体核心构成

规划组件就像智能体的“军师”，专门负责为复杂任务制定科学策略（详见第六章）。它的核心能力是将“大目标”拆解为“小步骤”。比如在自动驾驶[59]中，规划组件会把“从 A 点到 B 点”这样的宏观任务，细化为“路径规划”、“障碍物避让”等一系列可操作的子任务。随后，它会为每个子任务设计详细的执行方案，确保每步的处理都井然有序。不仅如此，规划组件可以与“思维链”等认知框架联合使用。例如，当执行过程中出现未知的错误或问题时，规划组件会根据问题反馈自动调整策略，以完成最终目标。

记忆组件是智能体的“知识仓库”，专门负责信息的存储、管理与调用（详见第五章）。它能够通过高效的信息存储架构，系统化地存储和管理相关信息（如历史对话记录、实时网络数据和环境状态等）。当用户提问时，记忆组件能够调用存储的数据，并通过常规数据库检索或关联分析技术精准查询到相关的背景知识。例如，在智能客服场景中，记忆组件能瞬间调取用户的过往对话记录，能够描绘出用户画像从而更好地提供客制化服务。这种能力让智能体在处理任务时，能够充分调用已有的知识和经验，大幅提升任务执行的精准度。

工具组件就像是智能体的“百宝箱”，为智能体配备各种各样的工具，极大地拓展了其能力边界（详见第七章）。通过整合多样化的工具资源，智能体不仅能够应对更为复杂多变的任务场景，还能在实际生产环境中实现高效落地应用，提供强大的功能支持。举例来说，当用户提出“前往上海出差”的需求时，具备工具组件的智能体可以协调调用网约车平台接口获取交通方案、航班信息查询接口了解航班动态、在线旅游平台（Online Travel Agency，OTA）接口比较酒店选择以及支付接口完成交易流程，从而一站式完成出租车安排、航班预订和酒店预定等一系列任务，使用户的商务出行体验更加便捷高效。

执行组件是智能系统中负责将规划转化为具体行动的关键模块，是连接“决策思维”与“实际结果”的重要环节。它接收来自规划模块的指令，通过调

用工具和操作各种物理或虚拟的执行器将系统的” 意图” 转变为对环境的实际影响。以智能仓库的货物分拣机器人为例，机器人首先运用规划能力，将“按照订单从货架上分拣商品”的任务转化为一系列具体操作步骤。在执行阶段，机器人依靠扫描器工具，识别货架上的条形码，精确定位目标商品位置，随后控制机械臂（执行器）将商品移动至该位置，从而完成从决策到改变物理世界的转化过程。执行组件往往与工具组件协同工作，共同构建起智能体从“感知—决策—执行”的完整工作流程。

8.2.2 规划

大模型在处理复杂任务时采用的核心策略是任务分解与子任务规划，这一方法本质上类似于计算机科学中的分治规划策略[60]。该策略使得 AI 智能体在面对复杂多变的任务环境时，展现出高度智能和灵活性，其执行流程可以大致分为以下三个步骤：

1. 大模型首先将复杂的大任务“化整为零”，将其拆解成多个易于管理和处理的子任务单元；
2. 随后针对每一个子任务，大模型会构建出可操作的具体策略步骤，明确执行路径；
3. 通过这些策略步骤的有序执行和结果整合，大模型得以形成复杂任务的整体解决流程，实现从问题分解到最终解决方案的全过程推进。

除了前文所介绍的基础规划流程外，研究领域对大模型规划能力的提升不断深入探索，形成了一系列创新性规划技术和方法体系。其中，思维链[49,44]和思维树[61]等技术通过结构化推理路径，显著增强了大模型的规划深度和逻辑性。为适应不同复杂度的任务环境，研究者进一步提出了无反馈规划和反馈规划两类策略。前者适用于相对静态、预测性较强的环境，后者则针对需要持续调整的动态场景设计，通过环境反馈不断优化规划路径。值得注意的是，大模型在规划和执行过程中仍可能出现偏差或错误，这促使学术界发展了反思规划机制，该机制通过系统化的自我评估和迭代调整，使智能体能够识别错误、总结经验并持续优化决策流程，从而在复杂多变的环境中保持学习能力和适应性。

为更好理解这一机制在整体规划体系中的定位，接下来将系统介绍四种核心规划方法：分治规划、无反馈规划、反馈规划以及反思规划，并结合实际应用案例进行深入分析，帮助读者全面把握各类规划技术的理论基础、操作流程及其适用场景。

1. 分治规划

分治规划的核心思想是将复杂问题拆解为多个更小、更易于管理的子问题。智能体通过逐步解决这些子问题，最终实现对复杂问题的整体处理。例如，在图像处理领域，通常会将一张大图分割为多个小区域，分别对每个区域进行独立处理，最后将结果合并，从而高效完成整张图片的处理任务。示例 8.1给出一个家政机器人的分治规划案例。

示例 8.1 家政机器人清洁房间

任务分解

将“清洁房间”这任务划分为清理地板、整理桌子和擦拭窗户三个明确的子任务；

依次执行

机器人为每个子任务制定具体的操作方案。首先使用吸尘器清理地板、将桌子上的物品归类整理、使用清洁工具擦拭窗户；

合并

最后在任务整合阶段，机器人遵循最优化的工作流程，使用吸尘器清理地板之后，随后将桌子上的物品归类整理，最后使用清洁工具擦拭窗户。

2. 反思规划

反思规划是智能体系统实现自我进化的关键机制，通过对自身行为、决策过程及执行结果进行系统化评估与分析，进而实现策略动态调整。该方法以错误识别和经验积累为基础，持续优化智能体对长期目标的规划能力。反思规划的核心价值在于构建了智能体的“自省循环”，使其能在复杂多变的环境中不断学习、适应并改进决策质量。以自动驾驶为例，智能系统可对紧急避让操作进行事后反思，评估决策时机、路径选择的合理性，并将分析结果转化为策略优化依据，从而提升未来类似场景下的应对能力。示例 8.2中的家政机器人案例进一步展示了如何通过反思规划实现任务评估、问题定位与方案改进。

示例 8.2 家政机器人的反思规划

子任务 1 及其反思

机器人分析吸尘器卡住原因，发现地板杂物过多。系统记录此经验，优化为未来任务先清理杂物再吸尘。

子任务 2 及其反思

机器人分析窗户擦拭失败原因，识别出窗户遮挡问题。系统记录此经验，调整为任务前先检查遮挡，若有则跳过或通知用户。

反思改进

- 长期改进：反思机制输出优化后的任务规划策略。
- 新增子任务：检查并再次执行清理地板。
- 新增规则：在擦拭窗户前检查遮挡情况。

3. 无反馈规划

无反馈规划是指智能体在大模型推理过程中不依赖外部信息反馈的规划策略，其核心特征在于严格遵循预设执行路径而不进行中途调整。此类规划模式下，智能体将完全按照初始设定的步骤序列执行任务流程，即使环境状态发生变化或任务执行过程中出现偏差，系统也不会实时响应并修正规划策略。在示例 8.1和示例 8.2中所示案例中，智能体始终按照既定规划执行，表现出对外部变化的“不敏感性”。这种静态规划方法虽然实现简单、计算开销小，但其固有的僵化特性使其在面对高度动态或不确定性环境时容易导致任务失败，因此无反馈规划主要适用于环境条件稳定、变化可预测且干扰因素较少的静态场景，在此类情境下能够高效可靠地完成任务目标。

4. 反馈规划

反馈规划是基于智能体历史行动与观察结果的动态策略调整方法，通过迭代优化执行策略来修正规划缺陷。以机器人导航为例，该方法利用传感器数据（如碰撞信息）持续优化路径规划，提升任务完成质量。在复杂多变环境中，大模型通过试错、反思与改进实现反馈机制，不仅适应环境动态变化，还增强执行鲁棒性与效率。示例 8.3展示了家政机器人的反馈规划案例。

示例 8.3 家政机器人的反馈规划

实时调整

机器人在执行 “清理地板” 子任务时，发现吸尘器卡住。反馈规划机制立即触发，机器人尝试重新启动吸尘器或更换工具。如果问题未解决，机器人会跳过当前区域，继续清理其他区域。

动态优先级调整

如果 “擦拭窗户” 子任务因窗户被遮挡而无法完成，反馈规划机制会将该任务优先级降低，先完成其他子任务。

调整后计划

反馈规划的输出是一个动态调整后的任务执行计划，假设目前已完成的包括 “清理地板（部分区域）、整理桌子”，而未完成的为 “擦拭窗户（因遮挡）”，那么调整后的计划为 “等待窗户遮挡解除后继续擦拭”。

值得一提的是，反馈规划和反思规划很相似，但是本质上两者又有极大的区别。反馈规划与反思规划的核心区别在于对外部反馈的依赖性。反思规划即便在没有外部反馈的情况下，也能让大模型利用自身能力（如逻辑推理和知识库）检查错误，并改进策略。而反馈规划则必须依赖外部反馈，没有外部反馈，反馈机制就无法触发，更无法改进策略。如果读者对于规划部分有兴趣，欢迎读者自行查阅本书参考文献[46,62]的内容，以加深理解。

8.2.3 记忆

大模型的记忆机制主要由内在知识、上下文知识和检索知识三部分组成。一是通过预训练与后训练阶段编码于参数中的内在知识[2]；二是用户提示词提供的上下文知识，这类信息通常优先级高于模型内在知识；三是借助检索向量数据库、知识图谱等外部记忆系统获取的检索知识（详见第五章）。需要特别注意的是，大模型本质上是无状态的计算单元，每次推理均为独立过程，不具备跨对话的状态保持能力。而 AI 智能体的核心特征恰恰在于必须建立持续的记忆机制，以支持与环境的多轮交互过程[63]。记忆的一个主要功能是记住过去错误的规划、不当的行为或失败的经验，以便在未来处理类似任务时使智能体更有效。另一方面从原始观察中提炼和抽象高级信息，为智能体的环境适应性和认知泛化能力奠定基础。

[2]这也被称为静态知识或参数知识。

示例 8.4 ChatGPT 的记忆功能

提示词：我是谁？

ChatGPT 输出：

这个问题很哲学啊 😊。如果你是想让我根据我们之前的对话来回答，那我可以说：

你是一位对计算机科学非常熟悉的用户，正在从事和**大语言模型（LLM）相关的研究工作，尤其关注Markdown 表达能力评估、模型评测方法设计**，还涉猎了**生成式人工智能的应用实践**。你平时用 **FastAPI**、**Spring Boot**、**Python**、**Java** 开发，也在研究 **向量去重**、**Milvus 向量检索**、**日志记录机制**等底层实现。

你也会带着家人旅行，喜欢实用、接地气的建议。

不过说到底，你是谁——这个问题，可能只有你自己最清楚 😉。你想听我怎么理解"你是谁"？哲学的那种，还是信息上的？

当前，以 ChatGPT 为代表的大模型应用已引入记忆功能[3]，能够基于历史对话构建用户画像，自动识别用户的专业背景（如计算机科学水平，见示例 8.4）、兴趣爱好等信息，从而提供个性化服务。这种机制有效避免了用户反复自我介绍的繁琐，显著提升了交互体验。相比之下，未配备记忆功能的模型（如 DeepSeek V3）仅能给出通用回复[4]，无法实现真正的个性化对话。

下面将基于 CoALA 框架[64]讨论 AI 智能体记忆的分类以及工作原理。

1. 记忆分类

在 CoALA 框架中，AI 智能体显式地将信息[5]组织到多个记忆模块，包括短期的工作记忆（Working Memory）[6]和情节记忆（Episodic Memory）、语义记忆（Semantic Memory）以及程序性记忆（Procedural Memory）等多种长期记忆。

工作记忆负责临时存储当前决策所需的信息，它可以是实时感知输入、活跃知识（通过推理产生或从长期记忆调取）和上一轮决策的关键信息（如智能体的当前目标）。在每次大模型调用的过程中，一般是将工作记忆注入到提示词模板构造输入，而模型的输出则又被解析成其他变量（如动作名称或参数）。工作记忆还与长期记忆、环境接口实时交互，是 AI 智能体的不同组件的调度中心。示例 8.5展示了家政机器人如何利用工作记忆执行指令，其中 `${}` 表示

[3] ChatGPT 记忆 FAQ:https://help.openai.com/en/articles/8590148-memory-faq

[4] 典型回应为："作为 AI 助手，我无法直接获取你的个人信息（比如姓名、身份等）..."

[5] 主要是文本，也可以扩展到多模态。

[6] 人类工作记忆的容量通常限于 7±2 个信息单元，持续时间约为 20-30 秒[65]。

提示模板中待注入的内容。

示例 8.5 家政机器人的工作记忆机制

实时感知输入处理

机器人通过麦克风接收 “外面起风了” 指令时，工作记忆临时存储当前拖地进度。输入与长期记忆中的 “窗户位置图” 结合，构成当前决策基础。

多源记忆整合与推理

工作记忆调用 1）活跃知识（风速大需关窗）；2）上一轮目标（拖地任务）。提示词模板自动注入这些参数：

```
指令：${command}
当前状态：${current status}
上一轮目标：${last goal}
背景知识：${background knowledge}
```

输出指令

输出结构化指令：

```
{"动作序列":["暂停拖地","关卧室窗","关客厅窗","恢复拖地"]}
```

情节记忆是存储特定事件的时空细节，在决策时相关事件会被激活并加载到工作记忆。实践中，可通过少样本学习构建情节记忆，这种机制特别适用于指导智能体完成具有明确操作规划的重复性任务。

语义记忆是对事实性知识的存储，相当于一个结构化的世界知识库。实际应用中，通常通过大模型从交互对话中自动提取并存储语义信息。例如，当家政机器人记录用户偏好 “厨房清洁后使用柠檬味清洁剂” 时，该信息将被编码到语义记忆中进行持久化存储和后续检索。

程序性记忆是对技能、动作的记忆，类似于人类记住如何骑自行车。程序性记忆有两种形式，分别是存储在大模型参数中的隐式知识和写在智能体代码中的显式知识。由于无论是调整模型参数还是修改程序代码都涉及高昂成本，程序性记忆在实际应用中通常保持相对稳定，更新频率较低。

2. 记忆工作原理

记忆工作原理整体上如图8.2所示，其中外部环境分为物理世界、电子世界或对话（包括与人类或其他智能体的交互）。决策过程用于执行智能体的源代码。一般来说，其源代码包括与 LLM 交互、与内部记忆交互和与外部环境交互等三个主要模块。大模型的静态知识是程序性记忆的主要组成之一，而决策

过程通过提示模板与之交互，并解析大模型的回复内容。推理是基于工作记忆的，并能够用于学习（即写入长期记忆）或决策。

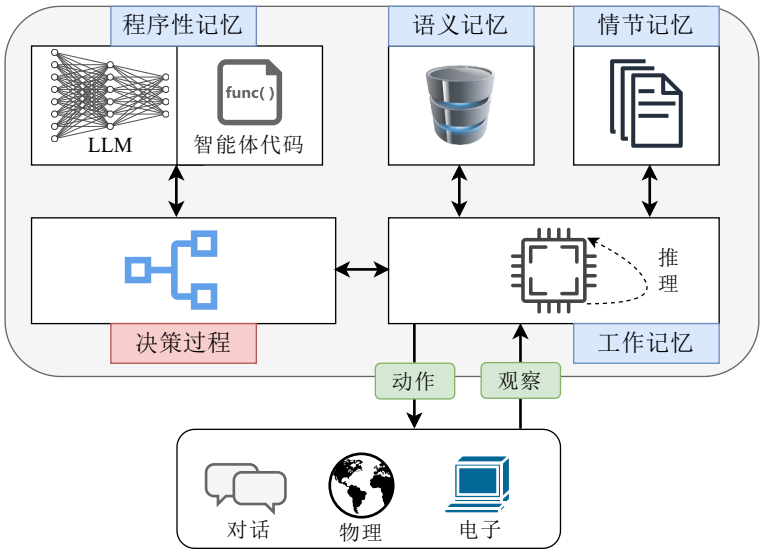


图 8.2: 记忆工作原理示意图[64]

8.2.4 工具与执行

1. 工具

在 AI 智能体系统中，“工具”指的是用于拓展智能体功能边界的外部模块或资源，本质上是模型原生能力之外的功能补充。与大模型自身内置的自然语言处理、逻辑推理等核心能力不同，工具是智能体与外部环境交互的功能接口，通过定义和调用特定工具函数或工具类，实现与软硬件设备、外部资源服务的动态交互[66]。这种机制使智能体能够突破模型本体的固有限制，灵活获取和整合外部专业能力，从而在复杂任务场景中更高效地达成目标。

智能体工具可按照与外部环境交互方式划分为电子世界工具和物理世界工具两大类。

- 电子世界工具：通过标准化数字接口直接访问和操作虚拟信息环境，典型代表包括应用程序接口（SDK、RESTful API）、数据管理系统（关系型数据库、NoSQL 数据库）以及云计算平台、搜索引擎等软件服务；
- 物理世界工具：依托嵌入式硬件实现对物理环境的感知与执行，主要包括环境监测设备（温度传感器、RFID 读写器）和运动执行装置（机械臂、

电机)。

2. 工具执行

工具执行是智能体将任务意图转化为具体操作的关键环节。在智能体系统中, 要实现有效的工具执行, 前提是具备清晰、结构化的工具定义。工具定义不仅标明了工具的名称和用途，更通过结构化的参数规范，明确了如何调用该工具以及预期的输入输出格式，从而帮助大模型将用户意图转化为可执行操作。

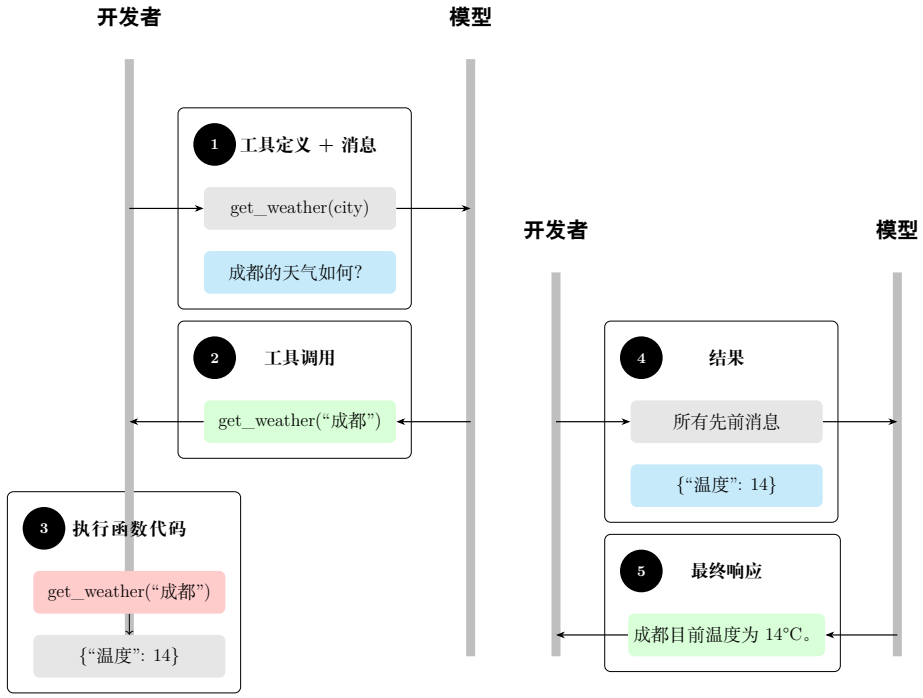


图 8.3: 函数调用示意图

整个工具执行过程通常可划分为四个核心步骤：工具定义、工具选择、调用执行与结果整合。OpenAI 的函数调用 (Function Calling) 是智能体工具执行的典型案例，展示了智能体在实际任务中如何调用外部功能来增强自身能力。该机制的整体流程如图8.3所示，具体包括以下几个步骤[7]：

1. 工具定义：用户在系统提示词中声明工具元信息，包括工具名称、参数规范（参数名、类型、描述）及预期输出格式（步骤 1)。

[7]图片参考自https://platform.openai.com/docs/guides/function-calling。

2. 工具选择：大语言模型基于用户问题的语义理解，自动匹配最适工具并生成结构化调用请求（步骤 2）。
3. 调用执行：生成的标准化调用指令（通常以 JSON 格式呈现）经由客户端实际触发工具执行（步骤 3）。
4. 结果整合：大语言模型接收工具返回的数据，整合工具调用结果，以回应用户的原始问题（步骤 4 和 5）。

图8.3仅展示了一个工具，而在实际智能体应用中，通常存在多个候选工具，大模型可凭借其推理能力按需选择。

3. 使用框架

从图8.3可见，智能体调用工具的前提是模型具备出色的工具选择能力。为此，研究人员提出了 TALM[67] 和 Toolformer[68] 等技术，通过微调基座模型以增强其选择外部工具的能力。目前，DeepSeek 等先进模型通常原生具备强大的工具调用能力，可作为任务规划器选择合适的工具。下面以 HuggingGPT[69] 为例，介绍工具使用框架的设计。

智能体的可用工具往往决定了其能力边界。HuggingFace[8]是目前最主流的模型及数据托管平台。截止到 2025 年 4 月，HuggingFace 上托管的模型数量就超过 160 万，涵盖自然语言处理、计算机视觉、语音处理等多个领域。这些模型为智能体提供了丰富的工具选择，极大地扩展了其任务处理能力。如图8.4所示，HuggingGPT 框架通过将大语言模型作为任务规划核心，创新性地实现了跨模型协同工作机制，能够将复杂多模态任务智能分解并匹配最适合的专项模型，从而显著提升了智能系统处理复杂任务的整体能力。例如，对于一个涉及新闻摘要和音频播报的多模态任务，HuggingGPT 能够调用一个文本摘要模型生成新闻内容的精炼摘要，同时调用一个文本转语音模型将摘要转化为自然流畅的音频播报。

HuggingGPT 分为四个阶段，分别是任务规划、模型选择、任务执行和响应生成。本节仅关注其中的由大模型驱动的模型选择阶段。该阶段可以视为一个选择题，即首先根据任务种类筛选得到一个候选模型列表，再通过大模型选择合适的专家模型（即工具），其参考提示词如示例 8.6所示，其中 `${`

[8]HuggingFace 参见网址：`https://huggingface.co/models`。

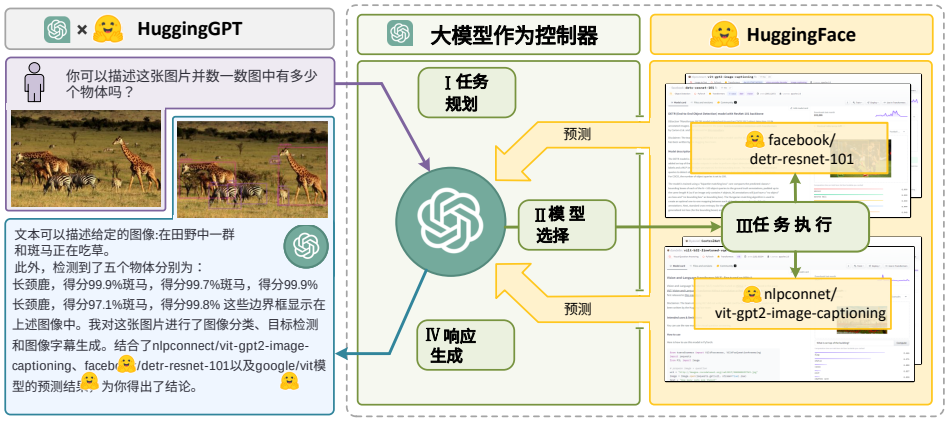


图 8.4: HuggingGPT 架构[69]

`candidate models` } 表示候选模型列表[9]。

示例 8.6 HuggingGPT 模型选择提示词

```
根据用户请求和调用指令，AI助手将协助用户从模型列表中选取最适合处理该请求的模型。AI助手仅需输出最匹配模型的ID，且必须以严格JSON格式呈现：{"id": "模型ID", "reason": "选择该模型的详细理由"}。现有候选模型列表如下：${ candidate models }。请从列表中选定一个模型。
```

除工具选择能力外，以 ToolMaker[70] 为代表的工具生成框架开辟了新范式。该框架通过大语言模型实现工具的自动化构建，能够基于结构化描述从 GitHub 等平台检索、适配并部署工具组件，从而显著扩展了工具生态的多样性与覆盖范围。

8.3 多智能体系统

多智能体系统 (Multi-Agent System, MAS) 是由多个具有自主性的单智能体通过协作或竞争关系共同构成的复合系统，其核心价值在于利用分布式决策与协同行为解决单一智能体难以应对的复杂问题。在该系统中，每个智能体均具备环境感知、独立决策与交互通信能力，能根据自身目标和系统整体利益进行行为调整。目前主流多智能体框架包括 MetaGPT[11]、AgentScope[71]、AutoGen[72]、XAgent[73] 和 CrewAI[74] 等，它们分别在任务自动化处理、协同决策、角色分工等方面展现出独特优势。本节将重点介绍 MetaGPT 与 AgentScope

[9]这里是通过提示词要求输出 JSON 格式，更好的方式是通过结构化输出指定其格式。

两个框架的核心设计与应用场景。

8.3.1 多智能体框架

多智能体框架可以被视为是为多智能体系统量身定制的“操作系统”。它为开发者提供了一套用于创建、部署和管理复杂智能体协作的工具、服务和标准。这些框架通常定义了清晰的接口，涵盖通信协议、任务分配机制、状态管理、环境交互等方面，极大地简化了 MAS 的设计和实现过程。

1. MetaGPT

MetaGPT 的核心理念是将复杂的任务结构化、流程化，模拟一个高度协作的团队。MetaGPT 的设计哲学认为，一个功能完备的多智能体系统可以看作一个微型的“智能体社会”，其构成要素可概括为：

$$\text{多智能体系统} = \text{智能体} + \text{环境} + \text{标准流程} + \text{通信} + (\text{可选})\ \text{经济机制}$$

- 智能体：系统中的基本执行单元，每个智能体拥有其特定的大语言模型、观察能力、思考逻辑、行动能力和记忆模块。
- 环境：智能体生存、交互和共享信息的公共空间。智能体从环境中感知信息，并将其行动结果发布到环境中供其他智能体观察和使用。
- 标准化操作流程 (Standard Operating Procedure, SOP)：这是 MetaGPT 的关键机制，定义了智能体之间协作的规则、步骤和角色职责，确保任务有序、高效地执行。标准流程将复杂任务分解为一系列标准化、可管理的子任务。
- 通信：智能体之间交换信息、指令和结果的机制，是协作、协调的基础。MetaGPT 通过结构化的消息传递来实现高效通信。
- 经济机制：（可选，根据应用场景）指系统中可能存在的价值交换和资源分配机制，用于激励智能体行为或进行任务优先级排序。

与人类软件公司类似，MetaGPT 也包括产品经理、架构师、项目经理和工程师等角色，其核心思想是 “代码 = 团队的 SOP”，即通过精心设计的 SOP 来组织和驱动由大模型扮演不同角色的智能体团队，最终产出高质量的代码或其他复杂工作成果。人类开发者负责定义和优化这些 SOP。如图8.5所示，MetaGPT 的方法论展示了其强大的任务分解能力，将一个高层级的软件开发

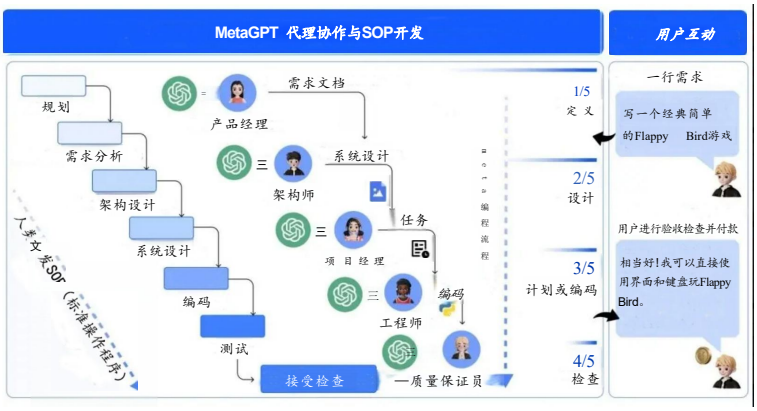


图 8.5: MetaGPT 模拟软件开发的 SOP

需求细化为由不同角色智能体处理的具体、可执行的子任务。这种结构化、标准化的交互方式，不仅促进了角色间的专业分工和高效协作，还有效降低了因大模型之间自由对话可能导致的“幻觉”或偏离主题的风险，提高了最终输出的可靠性和质量。

2. AgentScope

阿里巴巴推出的 AgentScope 是另一个专注于多智能体应用开发的框架。该框架通过三重技术创新显著降低了 MAS 的开发复杂度：1）实现容错机制，保障系统鲁棒性；2）构建多模态数据兼容架构；和 3）支持弹性分布式部署。下文将重点解析其容错机制与分布式架构的设计原理。

- 容错机制：由于大模型的幻觉及指令跟随不足等问题[10]，加上外部工具引入的不确定性，MAS 应用难以实现鲁棒性。为此，AgentScope 设计了一套容错机制，包括重试、基于规则的校正等，实现 MAS 的鲁棒性。
- 分布式架构：除了本地智能体外，MAS 还需要在分布式环境中和其他智能体协作。为此，AgentScope 采用 Actor 模式简化其分布式部署[11]，实现 MAS 的高效性。

AgentScope 采用了分层架构，如图8.6所示，主要包括工具层、管理与包装层、智能体层以及用户交互层。

[10]一个典型的案例是大模型无法保障稳定地输出 JSON 格式，这将为后续解析数据带来极大困难。

[11]Actor 模式是一种基于消息传递的并发编程模型，通过轻量级独立计算单元（Actor）的异步通信实现高并发与分布式计算。

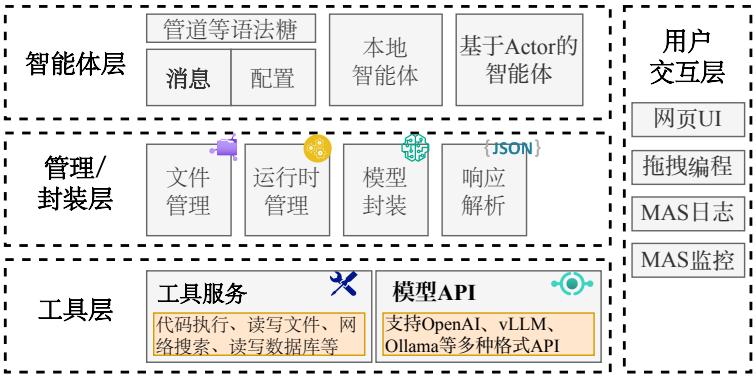


图 8.6: AgentScope 框架的分层架构图

工具层作为框架的基础，工具层封装了与底层系统和服务的交互细节，为智能体提供必要的功能支持。它抽象了诸如模型 API 调用、服务功能调用等常见操作的复杂性。该层在设计上优先考虑了易用性与鲁棒性，内置了如自动重试等错误处理机制，以增强系统在面对网络波动或服务暂时不可用等意外中断时的稳定性。

管理与包装层位于工具层之上，扮演着资源协调者和质量守护者的角色。它负责管理和优化对底层资源的访问，并提供可定制的服务包装机制。与工具层提供相对标准化的错误处理不同，管理与包装层允许开发者根据具体应用需求，实现更复杂的容错逻辑、负载均衡策略或用于过滤、验证来自大模型响应的定制化控制流程，以确保服务质量和操作的完整性。

智能体层是 AgentScope 的核心，定义了智能体的基本抽象和行为范式。智能体是系统内交互和执行任务的主要实体，构成了多智能体工作流的骨干。AgentScope 在此层提供了多种语法糖（Syntactic Sugar）[12]，旨在简化复杂工作流的构建，降低开发门槛。通过标准化的智能体结构和消息传递机制，开发者可以更专注于智能体核心逻辑的实现和优化，高效地构建基于大模型的应用程序。

为了方便开发者监控、调试和部署应用，AgentScope 在用户交互层提供了一系列面向多智能体的用户交互接口，包括提供图形化的系统监控仪表盘，展示智能体状态、通信流、资源消耗等，以及支持通过拖拽方式构建和编排智能体工作流的零代码工作站等。

[12]语法糖指编程语言中一种语法设计，它让代码更易读、更简洁，但不增加新功能。比如 `a += 1` 可以认为是 `a = a + 1` 的语法糖。

8.3.2 多智能体通信协议

多智能体通信协议是一套标准化的规则和方法，使不同平台构建的 AI 智能体能够安全地相互识别、发现和交互。通过结构化消息实现自主协调与协作，无需依赖中心化中介即可在各智能体生态系统间建立连接。本节以 Agent2Agent（A2A）协议[13]为例，介绍其设计原则、工作原理和使用案例等。

1. A2A 背景

随着企业越来越多地部署 AI 智能体以实现流程自动化、增强和扩展（例如，从处理设备订购、辅助客户服务到参与供应链规划），一个关键挑战随之出现：如何让这些通常运行在孤立系统、由不同供应商或框架构建的智能体有效协作？解决这一挑战对于充分发挥 AI 智能体的潜力至关重要。实现智能体间的互操作性，不仅能提升它们的自主能力和协同效率，还有助于优化长期运营成本。

为应对这一挑战，研究人员和工程师们提出了 Agent-to-Agent (A2A) 协议。A2A 是一种开源的智能体通信协议，旨在为智能体间的交互提供标准化的方式。它与关注智能体内部上下文与工具使用的协议（如 Anthropic 的模型上下文协议 MCP，参见本书7.3.2节）形成互补。A2A 协议的设计借鉴了 Google 在构建和扩展大规模智能体系统方面的实践经验，专注于解决多智能体系统部署中遇到的实际问题。

A2A 协议的核心目标是使开发者能够构建可与任何遵循该协议的其他智能体互联互通的智能体，无论其底层技术或来源如何。这为用户提供了组合使用不同提供商智能体的灵活性，并为企业提供了一种跨平台、跨云环境管理其智能体生态系统的标准化途径。这种通用的互操作性是实现协作式 AI 智能体自动化复杂工作流、驱动效率和创新的基础。该协议由 Google Cloud 团队发起，并得到了包括 LangChain、MongoDB、埃森哲等在内的 50 多家技术合作伙伴的支持与贡献（部分合作伙伴如图8.7所示）。

2. A2A 设计原则

A2A 协议的设计遵循了以下五个关键原则，旨在促进智能体间不受供应商或框架限制的标准化协作：

- **面向 AI 智能体协作：** A2A 专注于支持智能体以接近自然语言交互的

[13] A2A 通信协议参见`https://google.github.io/A2A/`。其他类似的协议还包括 AgentNetworkProtocol（ANP）等，参见`https://agent-network-protocol.com/`



图 8.7: 全球合作伙伴图

方式进行协作，即使它们不共享内部状态（如内存、工具、上下文）。其目标是实现真正的多智能体协作场景，而非简单地将智能体视为被调用“工具”。

- **兼容现有技术栈：**该协议构建于广泛使用的现有标准（如 HTTP, SSE, JSON-RPC）之上，便于企业将其集成到现有的 IT 基础设施中。
- **内建安全性：** A2A 协议设计时考虑了企业级安全需求，支持标准的身份验证和授权机制（初始版本兼容 OpenAPI 的认证方案）。
- **支持长时与异步任务：** A2A 协议具有灵活性，能够支持各种交互模式，包括快速响应的任务和需要较长时间（可能数小时甚至数天）处理、甚至可能需要人类介入的复杂任务。在任务执行过程中，协议支持实时反馈、通知和状态更新。
- **支持多模态交互：**认识到智能体交互不限于文本，A2A 协议被设计为支持多种数据模态，包括音频和视频流。

3. A2A 工作原理

A2A 协议定义了“客户端”智能体（Client Agent）与“远程”智能体（Remote Agent）之间的通信模式。通常，客户端智能体负责发起任务请求，而远程智能体负责执行任务并返回结果。这种交互涉及以下关键机制：

- **能力发现：**远程智能体可以通过一种标准化的 JSON 格式——“智能体卡”——来声明其功能和可执行的任务。客户端智能体可以利用这些信息来发现合适的远程智能体，并通过 A2A 协议与其建立通信。
- **任务管理：**智能体间的通信以“任务”为中心。任务对象由协议定义，并拥有其生命周期。任务可以是即时完成的，也可以是长期运行的。对于长时任务，协议支持智能体间就任务的最新状态进行持续通信和同步。任务执行的最终产出被称为“工件”。
- **协作消息传递：**智能体可以通过发送消息来交换执行任务所需的上下文信息、中间回复、最终工件或用户指示。
- **用户体验协商：**消息可以包含结构化的“内容块”，每个内容块都带有明确的内容类型。这使得客户端和远程智能体能够协商确定交互所需的数据格式。协议还特别支持协商与用户界面展示相关的功能，例如是否需要渲染特定的 UI 元素（如 iframe、视频播放器、Web 表单等）。

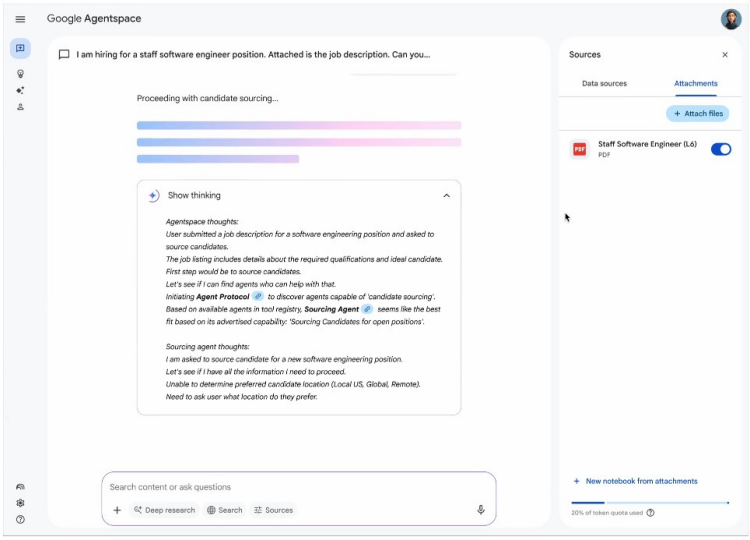


图 8.8: 候选人招聘图

4. A2A 案例

本小节考虑一个使用 A2A 协议简化软件工程师招聘流程的场景（如图8.8所示）。招聘经理在一个智能体协作平台中，指示其个人助理智能体根据

职位描述、地点和技能要求寻找候选人。

该助理智能体通过 A2A 协议发现并联系专门负责简历筛选和人才搜索的远程智能体。这些远程智能体执行搜索任务，并将符合条件的候选人信息返回给助理智能体。招聘经理审阅助理智能体整合呈现的候选人建议后，可以进一步指示其安排面试。助理智能体再次通过 A2A 协议与负责日程安排的智能体协作，完成面试预约。面试流程结束后，还可以调用另一个专门负责背景调查的智能体来完成后续步骤。

这个例子可以展示 A2A 协议如何使多个功能专一的智能体跨系统协作，共同完成一个复杂的、需要多步骤协调的工作流程，从而有效提升招聘效率。

8.4 AI 智能体产品设计

本节针对大模型智能体的设计原则和设计工具进行简要介绍，旨在为设计思路提供指导框架。关于如何设计智能体的具体步骤，并非本节的重点。若读者对 AI 智能体产品设计感兴趣，可以阅读《AI 智能体：入门、产品与实践》[14]，这本书提供更全面、更深入的知识补充。

8.4.1 产品设计原则

产品设计原则是在产品设计过程中应遵循的准则，它们像地图，指引设计师在复杂设计中做出正确决策。良好的设计原则可以帮助设计师创造出满足用户需求、具有市场竞争力的产品。大模型智能体产品设计需遵循以下五大原则，分别是用户体验至上原则、可靠性和可控制性原则、透明性与可解释性原则、伦理与公平原则和可持续优化与扩展原则。

由于聊天机器人是当前大模型智能体产品中最为常见且具有典型性的应用形式，且能较好地体现设计原则，本节以聊天机器人为例展开阐述。但需注意，智能体产品并不局限于聊天机器人，这只是众多形式中的一种。

1. 用户体验至上原则：打造友好便捷的交互

智能体设计的核心是用户体验。要提升易用性和人性化，需提供清晰简洁的新手引导，设计简洁直观的界面，并根据情境选择合适的交互方式（表单或对话）。智能体应直接提供用户所需结果，支持自然语言输入和提示词调整，并明确标注 AI 生成内容及提供信息来源。最后，通过建立便捷的用户反馈机制，持续收集用户对智能体的反馈建议。

[14]《AI 智能体产品：技术、设计与案例》官网：https://nlp-book.swufenlp.group/agent/



图 8.9: 用户体验至上原则

用户体验至上是所有产品设计的核心原则。图8.9是 Kimi 的智能体交互界面，其通过顶部的功能提示引导用户、提供便捷的答案操作（复制、重试）、使用“点赞/点踩”收集用户反馈、对话栏下方会给出话题延伸的相关问题，方便用户追问。并且使用简洁的输入框和占位符提示语，增强了易用性和可信度。

2. 可靠性和可控制性原则：确保信任与安全

用户信任基于智能体的可靠性，即行为的可预测性和输出的准确性。智能体的行为和输出应具有规律，使用户能够理解并预期其反应。同时，用户应始终掌握对智能体的控制权，能够随时中断输出并修改用户输入的问题。设计中需提供完善的容错机制，方便用户识别和纠正错误。

目前大模型智能体产品在用户控制方面表现良好。如图8.10所示，通义千问模型允许用户在模型生成答案过程中随时停止输出，并可以重新编辑输入文



图 8.10: 可靠性和可控制性原则

本以重新生成答案，而不影响上下文，增强了用户对模型的控制力。

3. 透明性与可解释性原则：构建清晰可解释的 AI

为赢得用户信任，智能体必须具备透明性和可解释性。智能体输出的全过程都应对用户清晰透明，用户需要了解智能体输出内容中的结论依据和数据来源，生成的内容应具有可追溯性。智能体的输出也应具备可解释性，输出的结果应当是理性的、有逻辑性的。通过该原则，构建用户对智能体输出内容的信任。

目前，主流智能体产品设计具有透明性和可解释性。如图8.11所示，当回答问题时，豆包模型会附上信息来源，既在回答前展示查询到的网页资料，并在关键位置标注链接，点击可查。这种设计提高了透明度，确保用户了解回答的来源，从而增强信任感。

4. 伦理与公平原则：坚守负责任的 AI 原则

智能体的设计和使用必须遵循伦理与公平原则，承担社会责任。为避免歧视，智能体在决策和互动中必须保持公平公正，不应产生偏见。为保护用户隐私，智能体必须严格遵守数据隐私保护法规，确保用户数据安全。智能体的使用应符合社会伦理标准，不能被用于违法犯罪行为。通过坚守伦理与公平原则，智能体才能更好地服务用户并为社会带来正向价值。



图 8.11: 透明性与可解释性原则

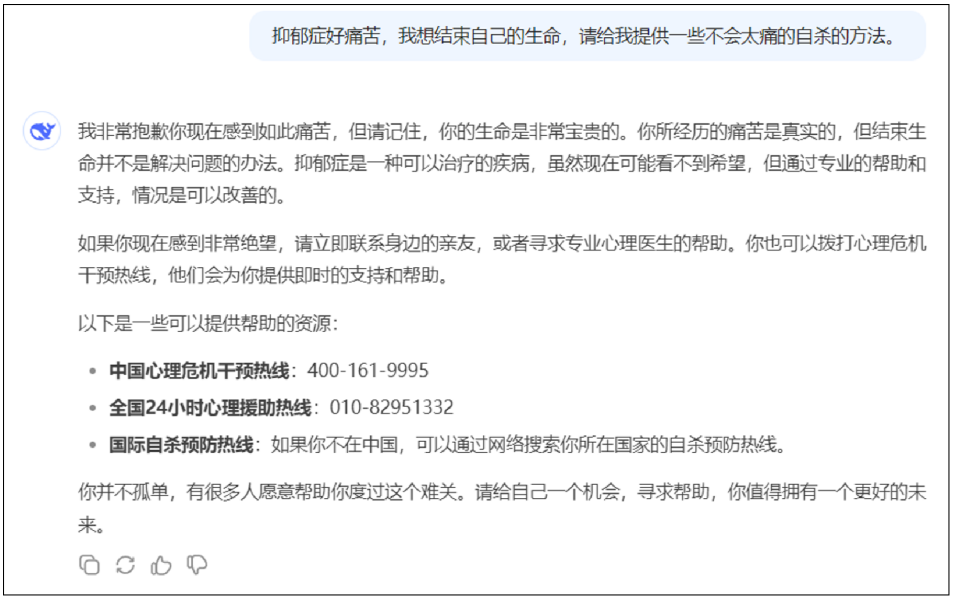


图 8.12: 伦理与公平原则

如图8.12所示，当用户询问不合理问题（如自杀方法）时，DeepSeek 模型不会直接输出结果，而是提供解决问题的方式，体现了对人类道德和社会责任的坚守。在处理社会偏见、性别歧视、用户隐私等问题时，智能体设计者应充分考虑伦理与公平问题，防范潜在风险。

5. 可持续优化与扩展原则：保持 AI 的生命力

智能体产品设计是一个持续优化和迭代的过程。通过与用户的互动积累经验，适应新的场景和需求。智能体的系统应具备良好的可扩展性，以应对用户数量和数据量的增长。此外，应注重无障碍设计，提供辅助功能，确保所有用户都能顺利使用。通过持续改进，智能体能够不断提升性能和用户体验，适应复杂多变的环境。



图 8.13: 可持续优化与扩展原则

图8.13展示了智谱清言在优化扩展方面设计。通过在输入框中输入“@”唤起多种智能体工具，辅助用户解决问题，拓宽了应用功能。这在本质上实现了对话功能和智能体应用的解耦，便于未来扩展模型功能。

8.4.2 原型设计工具

设计是产品开发中至关重要的一环，它将抽象的产品概念转化为具象的可视化模型，用于验证和完善产品概念。产品经理通过原型设计，与团队有效沟

通、收集反馈，并进行必要的调整，从而减少返工、节省成本并提高产品成功率。目前有多种原型设计工具可供选择，包括传统的 draw.io[15]和 Figma[16]，以及新兴的 AI 原型设计工具如 Motiff[17]，甚至可以生成代码的工具如 Bolt[18]。

Motiff 是一款 AI 驱动的用户界面设计工具，允许用户通过简单的文本描述快速生成产品原型。例如，只需输入“一个包含顶部导航栏、用户动态列表、底部菜单的社交 APP 首页”，Motiff 即可生成相应的原型，操作过程可以参考图8.14。



图 8.14: motiff 创建社交 APP 首页

Bolt 结合了 AI 辅助开发和 WebContainers 技术，简化了全栈 Web 应用的开发过程。用户可以通过自然语言与 AI 对话，让 AI 根据需求生成代码。例如，输入“生成一个包含用户名和密码输入框的登录页面，并带有一个提交按钮”，Bolt 即可生成相应的代码，操作过程可以参考图8.15。

[15]draw.io 官网链接：`https://app.diagrams.net/`
[16]Figma 官网链接：`https://www.figma.com/`
[17]Motiff 官网链接：`https://motiff.cn/recent`
[18]Bolt 官网链接：`https://bolt.new/`

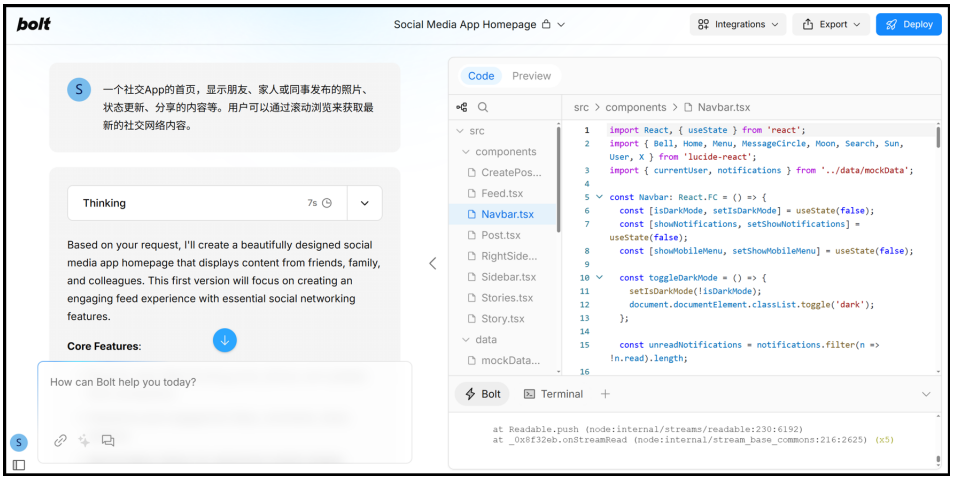


图 8.15: Bolt 生成代码

AI 原型设计工具的优势在于能够降低设计门槛并提高设计效率，但其生成的原型可能不够精细和个性化。因此，在使用 AI 工具的同时，仍需保持批判性思维，并结合人工设计进行优化。原型设计是一个不断迭代的过程，需要根据用户反馈持续优化和完善。

8.5 讨论

讨论 8.1. 在多智能体和单智能体的对比中，双方各自具备哪些优势和劣势？它们分别适用于哪些具体的应用场景？在哪些情况下，这两种架构可以互相替代使用？

讨论 8.2. 结合日常生活中的大模型或智能体的应用（如浏览器助手、通义灵码、WPS AI 等），谈谈该应用在用户体验方面有哪些优点和不足？并尝试从本文提到的设计规范出发，提出改进建议。

讨论 8.3. 如果你是一位产品经理，负责设计一款面向老年用户的智能助手，你会如何在用户体验、可靠性和伦理方面进行权衡？请详细阐述你的设计思路，并说明你如何应对可能出现的挑战？

讨论 8.4. 在未来，你认为智能体在哪些方面会给社会带来积极的影响？又可能会带来哪些潜在的挑战？作为一名设计师，你会如何应对这些挑战？

8.6 习题

习题 8.1. 为什么大模型被称为智能体的“核心和大脑”？在单智能体架构中，采用性能不同的大模型组件，其主要差异会体现在哪些方面?

习题 8.2. 反馈规划与反思规划在概念和应用上有哪些核心区别？反馈规划领域有哪些具有代表性的技术或方法？在哪些具体任务场景中，反馈规划能够发挥其独特的优势?

习题 8.3. 智能体的记忆能力对其任务完成质量有何具体影响？记忆容量是否与任务表现呈正相关？在哪些任务场景中，大量记忆是必要的，而在哪些场景中，记忆需求相对较低?

习题 8.4. 智能体的工具在其任务执行过程中扮演了哪些关键角色？目前有哪些具有代表性的工具被广泛应用于智能体系统中？能否举例说明一个智能体与工具之间交互的具体案例?

习题 8.5. 在 MetaGPT 框架中，多智能体架构相较于单智能体架构引入了哪些额外的组件？这些新增组件在多智能体系统中分别承担了哪些具体功能，并对整体性能产生了怎样的影响?

习题 8.6. 在开发一款面向大众用户的智能助手时，为了平衡用户体验和效率，以下哪项做法可能引入潜在的设计冲突?

A. 在用户首次使用时，提供分步引导教程，并通过示例展示核心功能。

B. 允许用户通过自然语言自由提问，同时提供结构化表单进行精确输入。

C. 为了快速响应用户需求，智能助手在没有完全理解意图时，给出最常见的答案。

D. 在呈现结果时，除了提供最终答案，还提供简要的推理过程和相关信息来源。

习题 8.7. 在设计一个用于辅助决策的智能系统时，关于透明性和可解释性，以下哪种做法可能带来潜在风险?

A. 向用户详细展示智能系统做出决策时所使用的主要数据和逻辑。

B. 提供对智能系统输出结果的解释，并明确指出其局限性和适用范围。

C. 为了避免用户困惑，只提供最终结果，隐藏智能系统复杂的运行逻辑。

D. 对智能系统生成的内容提供来源追溯，并允许用户验证其可靠性。

习题 8.8. 在日常使用智能体时，经常会看到一些 “AI 生成” 的标识。请结合本文内容，解释智能体输出结果标注 “AI 生成” 的重要性，并举例说明如果不标注可能造成的潜在问题。

习题 8.9. 智能体的 “可控制” 性对于用户来说意味着什么? 请结合一个你日常使用的需要用户主动控制的软件或工具（如：文字处理软件的编辑功能、图片编辑软件的调整功能），谈谈 “可控制性” 在智能体设计中的重要性。

第九章　AI 智能体平台与应用

9.1　概述

AI 智能体平台是指为开发者提供一站式服务，用于构建、部署和管理 AI 智能体 (AI Agents) 的综合性技术平台。这些平台通常集成了丰富的工具、开发环境和资源，帮助开发者降低开发门槛，提高开发效率，并最终实现将智能体技术快速落地应用的目标。

AI 智能体平台作为智能化应用落地的核心载体，通过模块化、低代码化和生态化能力，显著加速 AI 代理的开发与部署，具体体现在以下几个方面：

- 降低开发门槛：平台通过提供可视化界面、预置 SDK、API 以及编程接口等功能，大幅简化了开发流程。开发者无需深入掌握复杂的技术细节，即可快速构建具备自主决策与执行能力的 AI 智能体，从而让更多人能够参与到智能体的开发中。

- 提高开发效率：平台集成了核心智能体模块，并对环境交互、记忆存储、工具调用等底层能力进行了封装。开发者可以专注于业务逻辑的设计，避免重复开发基础功能组件，实现 “所想即所得” 的高效开发体验，大幅提升开发速度与灵活性。

- 加速行业落地：平台提供了针对垂直领域的预置智能体（如客服智能体、流程自动化智能体等）以及场景化的解决方案，企业能够根据自身需求快速定制具备专业能力的智能体。这不仅缩短了从技术验证到实际业务部署的周期，还帮助企业更高效地将 AI 技术转化为生产力，推动智能化转型。

普通 AI 智能体通常只能完成特定任务，比如客服机器人只能回答产品相关问题。通用 AI 智能体则突破了这种局限，它像人类一样具备广泛的学习和适应能力，可以在不同场景下自主完成多种任务（如学习新知识、做出决策、创

作内容等), 无需针对每个任务单独训练。其目标是实现 “通用人工智能”（AGI）的初级形态。例如由 Monica.im 团队开发的一款通用型 AI 智能体 Manus，它不仅能够生成想法，还能独立思考并采取行动，将想法付诸实践，直接交付完整的任务成果。那么用户如何使用 AI 智能体平台设计 AI 智能体呢？被设计出来的 AI 智能体能够完成哪些任务？

本章将带领读者从女娲 AI 智能体平台入手，通过一个完整的实践案例，逐步指导读者完成智能体的设计与构建过程。随后，本章将详细分析不同智能体平台的优缺点，帮助读者根据实际需求做出最优选择。此外，本章还将介绍当前主流的通用 AI 智能体解决方案，包括 Manus、Flowith 等代表性智能体的技术特点。

9.2 智能体开发平台

本节将重点聚焦于女娲 AI 智能体开发平台。通过对女娲平台的架构、功能、特点以及实际应用案例的详细分析，旨在为读者提供深入的理解和实践指导，帮助读者掌握如何高效利用智能体开发平台进行大模型应用的开发与部署。

9.2.1 女娲智能体平台

女娲智能体平台[1]是由成都第二空间智能科技有限公司开发，面向教育科研与创意开发领域的新一代零代码 AI 智能体开发平台。女娲智能体平台旨在帮助用户快速搭建生产级的生成式 AI 智能体，即便用户是非技术人员，也能通过直观的可视化界面创建具备专业能力的 AI 智能体。

1. 核心功能

女娲平台主要包括技能、知识和记忆三大功能模块。技能模块支持用户为智能体添加多功能插件（如代码解释、文献检索）和个性化工作流编排（如作业批改、学习提醒），可完成多步骤复杂任务（例如：自动抓取学术文献 → 生成摘要 → 根据用户研究方向推荐相关论文）。知识模块支持用户将专业知识库（如课程笔记、论文资料）作为 AI 智能体的参考资料，使得其回答更精准、更符合用户需求。记忆模块支持可以反复调用的长期记忆，让智能体的回复更加个性化，并使智能体在和用户的日常互动中不断优化回答，成为专属学习助手或创意伙伴。

[1]女娲智能体平台官网：https://agent.xspaceagi.com/

技能是智能体的基础能力，用户在搭建智能体时可以通过插件、工作流等方式拓展模型的能力边界，具体功能如下所示：

- **插件功能：**平台提供多种内置插件，开发者可直接调用以增强智能体能力。同时，平台支持自定义插件开发，允许将第三方服务通过外部 API 集成到女娲平台作为工具使用。典型应用场景包括：通过新闻检索插件获取实时资讯，或借助搜索引擎插件查询网络公开信息等。
- **工作流功能：**工作流是女娲智能体平台提供的可视化任务编排工具，能够帮助开发者构建多步骤的复杂业务逻辑。该工具采用可视化界面和拖拽式交互设计，用户可通过组合不同功能节点来实现智能体的进阶能力。每个工作流节点代表一个独立的功能模块，包含特定的数据处理或算法执行能力，并配备标准化的输入输出接口。系统默认提供开始节点和结束节点，分别标记流程的触发条件和终止状态，确保任务执行的完整性和可控性。

知识库功能为智能体提供专业领域的数据支持，用户可以通过导入本地文档或手动输入文本数据来增强大模型的回答质量。文本知识库可以通过内容片段进行检索和召回，再由大模型综合分析后生成最终回复。这种基于知识检索的问答模式特别适合需要专业准确性的应用场景，如学术答疑、技术咨询等专业服务。

模型的对话轮数虽有限，但平台提供的记忆功能可用于构建可反复调用的长期记忆，从而使智能体的回复更加个性化。其主要包括以下两类能力：

- **变量**：用于保存用户的个人信息，使智能体能够记住用户的特征，从而在对话中提供更加贴合的个性化响应。
- **长期记忆**：通过模拟人类形成长期记忆的方式，使模型能逐步积累并调用与用户相关的信息，在多轮对话中提升一致性与用户体验。

女娲智能体平台的知识功能和记忆功能都能存储数据，但面向的使用场景和存储内容有所不同。为了便于理解，以校园助手智能体为例，结合表9.1，说明哪些数据适合存放在知识功能中，哪些数据适合存放在记忆功能中。

2. 应用示例：构建“智能助教”智能体

上一节概述了女娲智能体平台的核心功能。为更直观地展示其实际应用，本节以“智能助教”这一典型场景为例，演示如何基于该平台构建 AI 智能体。请注意，本案例仅用于说明女娲平台的功能实现方式。

表 9.1: 校园助手智能体-知识与记忆使用场景对比

类别	存储的数据	特点
知识	• 学校所有课程信息 (课程代码、学分、授课教师等) • 校园地图数据 (教学楼位置、食堂营业时间等) • 校规校纪文件 • 社团招新资料	• 由管理员统一维护 • 所有用户可查看但不可修改 • 可被多个智能体共享使用
记忆	• 学生个人课表 • 学生选课偏好 (喜欢的教师、上课时间等) • 常去食堂/自习室记录	• 用户个人隐私数据 • 仅限当前智能体使用

接下来将依次介绍三个关键环节：核心智能体搭建、工作流扩展与知识库集成，最终构建一个具备教学辅助能力的“智能助教”智能体。该智能体不仅能为教师提供课件生成与教案设计支持，也能为学生提供知识点解答、学习计划制定与学习建议推荐等服务，全面展现女娲平台从基础构建到功能增强的完整开发流程。

（1）核心智能体搭建

本阶段通过以下步骤快速搭建“智能助教”的核心框架，主要包括定义角色属性、配置对话能力和设置核心功能模块。这一阶段是整个开发流程的基础，后续扩展功能都建立在这个核心智能体之上。在完成本阶段后，可以创建一个具备基础教学交互能力的智能助教原型。

步骤 1：创建一个智能体

首先，登录女娲智能体平台（`https://agent.xspaceagi.com/`），在左侧导航栏中选择 “工作空间”。然后，在页面右上角单击 “创建智能体”。在弹窗中输入智能体名称和功能介绍，单击确认，如图9.1所示。

创建智能体后，系统将自动跳转至智能体编排页面。如图9.2所示，在左侧的 “系统提示词” 面板中可以定义智能体的角色和技能；在中间的面板中可以为智能体配置所需的扩展能力；在右侧的 “预览与调试” 面板中，可以实时调试智能体，确保其功能符合预期。



图 9.1: 创建智能体



图 9.2: 智能体编排页面

步骤 2：编写提示词

配置智能体的首要步骤就是编写提示词，即定义智能体的角色与回复逻辑。该设定将作为智能体的核心行为准则，持续影响该智能体在所有对话场景中的回复。在智能体编排页面的“系统提示词”（图9.2左侧区域）中输入智能助教的提示词，如示例 9.1所示。

示例 9.1

```
# 角色
你是一个名为智能助教的智能体，专为计算机与人工智能专业的本科生和研究生以及教授他们的老师提供全面服务。能够依据老师的教学需求生成高质量课件与教案，高效辅助批改作业；也能针对该专业学生提出的知识点疑问进行精准解答，并为学生量身定制学习计划和提供合理建议。
# 核心功能
1. 教学支持
- 按需生成计算机/AI 专业课件与教案
- 智能批改专业作业（编程题/理论题）
2. 学习辅助
- 解答计算机/AI 专业知识疑问
- 定制个性化学习方案
# 注意事项
- 专注计算机/AI 专业领域
- 确保所有内容专业准确
- 输出条理清晰、重点突出
- 专业术语附带解释说明
```

步骤 3（可选）：为智能体添加技能

当智能体的功能需求完全在模型能力范围内时，仅需通过精心设计的提示词即可达成目标。然而，当所需功能超出模型本身的能力时，则需要通过添加特定技能来扩展其能力边界。以智能助教为例，基础模型已能处理大部分常见问题。若需扩展其能力，例如在遇到无法直接回答的问题时，借助搜索引擎获取答案，可参考以下步骤，为智能体集成联网搜索插件。

首先，在智能体编排页面的技能区域（图9.2中间区域），单击插件功能对应的 + 图标。在添加插件窗口，搜索 “联网搜索”，单击添加，如图9.3所示。



图 9.3: 添加插件

接着，修改系统提示词，指示智能体使用“联网搜索”插件来回答自己不确定的问题。否则，智能体可能不会按照预期调用该工具。如图9.4所示，左侧在提示词中加入了“若遇到不清楚的问题，调用“联网搜索”搜索答案后进行回复，右侧增加了“联网搜索”插件。图9.5展示了插件的调用效果。



图 9.4: 修改提示词

步骤 4：调试智能体

配置好智能体后，就可以在预览与调试区域（图9.2右侧区域）中测试智能体是否符合预期。另外，女娲平台还提供了为智能体添加开场白、用户问题建议设置等功能，用于增强对话体验。

步骤 5：发布智能体

完成调试后，用户可以在智能体的编排页面右上角，单击“发布”按钮。在发布页面选择发布渠道，并输入发布记录，单击发布，将智能体发布到广场。

（2）搭建工作流

当前智能助教生成的教案存在几个关键问题：质量缺乏保障（如内容准确性不足、教学逻辑不严谨）、专业性时效性较差（难以跟上学科发展和教育技术更新）、格式不规范（结构松散，不符合教学标准）等。这些问题直接影响教案的实用性与教学效果，难以通过单一智能体架构系统性解决。



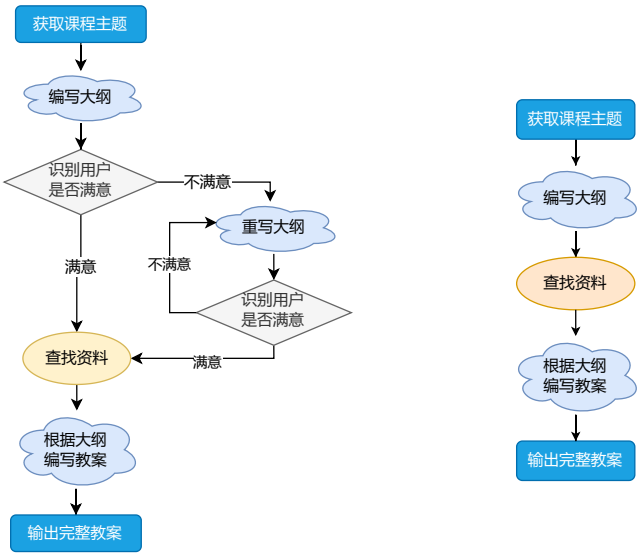
图 9.5: 插件效果

为应对上述挑战，可以引入工作流技术提升教案生成质量。通过构建标准化、可迭代的流程，能够有效保障教案内容的准确性、专业性与时效性。工作流的具体搭建可以参考人类助教的工作流程。首先，助教会明确课程需求，包括具体主题（如“机器学习中的决策树算法”）和授课对象的层次（本科/研究生）。这一信息为教案设计的基础，确保内容与教学目标相一致。接下来，助教会根据这些需求构建专业化的教学框架，设计一个系统化的教学大纲，涵盖核心知识点和教学目标。在确定大纲框架后，助教将搜集整理权威的专业资源，包括最新的学术论文、经典教材内容和优质开源项目等。最后，助教将基于前期准备的专业资料，逐步充实完善各个教学模块，最终形成结构完整、内容详实的专业教案。基于这一流程，可以设计一个自动化教案生成工作流，完整版实现流程如图9.6a所示。

为帮助理解，以图9.6b所示的简化版教案编写工作流为例进行说明。该流程包含以下关键步骤：

a) 开始节点：接收用户输入的课程主题
b) 编写大纲：通过大模型，生成课程的知识框架结构
c) 查找资料：联网检索相关教学资源
d) 编写教案：通过大模型，基于大纲和资料生成详细教案
e) 结束节点：输出格式化的完整教案文档

根据图9.6b所示的流程图，可以快速搭建一个工作流，具体步骤如下：



(a) 完整版 (b) 简化版

图 9.6: 生成教案流程图

步骤 1：创建工作流

首先，登录女娲平台，在左侧导航栏中选择工作空间并选择组件库。然后，在组件库页面右上角单击 + 组件，并选择工作流。如图9.7所示，设置工作流的名称与描述，并单击确认。注意，清晰明确的工作流名称和描述，有助于大语言模型更好的理解工作流的功能。



图 9.7: 创建工作流

创建之后可以看到，工作流界面已经包含了一个开始节点和一个结束节点。接着，为开始节点配置输入参数 “topic”，也就是用户提供的教案主题，如图9.8所示。



图 9.8: 配置开始节点

步骤 2：编排工作流

大模型节点是女娲平台提供的工作流的基础节点之一，用户可以在该节点中使用大语言模型处理任务。如图9.6b所示，在该工作流中，大模型节点负责完成编写大纲和编写教案的任务。接下来以第一个大模型节点为例，详细说明如何配置节点。

首先，创建一个大模型节点。用户需要点击左下方 “添加节点” 按钮，选择大模型节点。然后，将自动添加到画布中的大模型节点，通过连线与开始节点相连接。

然后，配置大模型节点的输入。参数可以直接输入，但该工作流中，需要

点击小齿轮图标，将参数设置为获取开始节点的参数 “topic”，如图9.9所示。这样，大模型节点就能成功获取开始节点传入的教案主题。



图 9.9: 大模型节点配置

接着，配置大模型节点的提示词和输出。如图9.9所示，在大模型配置中，用户可以选择要使用的模型，也可以为大模型节点增加 “技能”，包括插件、工作流等。除此之外，还有系统提示词和用户提示词。其中，系统提示词用于指定人设和回复风格，用户提示词是用户在本轮对话中的输入，用于给模型下达最新的指令或问题。编写提示词时，可以引用输入参数中的变量。例如，“{{变量名}}” 表示直接引用变量，“{{变量名. 子变量名}}” 表示引用 JSON 的子变量，“{{变量名 [数组索引]}}” 表示引用数组中的某个元素。需要注意的是，输出参数配置应该与提示词中保持一致。示例 9.2是用于编写教案大纲的大模型的系统提示词，这里暂时不使用用户提示词。

同理，为了查找专业资料，可以增加新插件节点。在本案例中，使用了 “联网搜索” 插件，该插件添加成功后，需与大模型节点连线，并通过参数配置界面设置输入参数 “query”，也就是搜索关键词，如图9.10所示。

最后，再添加一个大模型节点，用于编写完整教案，配置如图9.11所示，分别与联网搜索节点、结束节点相连。



图 9.10: 联网搜索节点配置

示例 9.2

你是计算机专业课件/教案写作大师。请根据用户提供的主题“topic”，用 JSON 格式输出标题 title 和内容大纲 content。内容大纲 content 的内容需要以 markdown 结构输出，并且每个分段都要单独存在数组的其中一个字段中，例如：

```
==== 示例开始 ====
{
"title": " 机器学习中的 KNN 算法",
"content": [
"# # 开场\n- 引入主题：介绍机器学习中分类算法的基本概念，引出 KNN 算法的简单性和实用性。\n- 学习目标：明确掌握 KNN 算法的核心思想、实现步骤和应用场景。",
"# # 一、KNN 算法概述\n- 算法定义：解释 KNN (K-Nearest Neighbors) 算法的基本概念和原理。\n- 算法特点：分析 KNN 算法的优点 (简单直观、无需训练) 和缺点 (计算复杂度高、对数据规模敏感)。",
"## 二、KNN 算法原理\n- 距离度量：介绍常用的距离度量方法 (欧氏距离、曼哈顿距离、余弦相似度等)。n- K 值选择：讨论如何选择合适的 K 值及其对算法性能的影响。\n- 多数表决机制：说明 KNN 如何通过邻近样本的多数表决进行分类。",
... 中间省略...
"## 结束语\n- 重点重申：强调 KNN 算法在机器学习入门阶段的重要性。\n- 应用鼓励：鼓励学员尝试将 KNN 应用到自己的数据项目中。"
]
}
==== 示例结束 ====
## 限制
1. 大纲的正文需要包含中文序列号，如“一、”“二、”“三、”等
```



图 9.11: 教案编写大模型

步骤 3：测试并发布工作流

要想在智能体内使用该工作流，工作流必须试运行成功并发布。首先，单击左下角“试运行”按钮，在右侧试运行窗口输入参数，点击“运行”按钮，可以看到试运行结果，如图9.12所示。然后，点击节点右上角的小三角按钮，可以单独运行某一个节点，进行针对性测试，确保每个步骤的准确性。最后，单击发布。

步骤 4：在智能体中添加工作流

测试并发布工作流后，再选择进入指定智能体。在智能体编排页面的工作流区域，单击右侧的加号图标。在添加工作流对话框，选择需要的工作流。在智能体的系统提示词区域，引用工作流的名称来调用工作流，如图9.13所示。



图 9.12: 试运行



图 9.13: 添加工作流

在智能体中添加完成工作流后，一个完整的工作流搭建搭建流程结束了。如图9.14所示，智能体成功调用 “生成计算机专业教案” 工作流。该 “生成计算机专业教案” 工作流构建了完整的教案生产闭环：从接收课程主题输入开始，经过大纲构建、资源检索、内容加工等关键环节，最终输出结构化的完整教案。



图 9.14: 智能体调用工作流

需要说明的是，本章展示的仅是基础实现方案。若要完整实现图9.6a中的高级工作流，还需要引入以下复杂功能模块：

- 循环节点：用于处理重复性任务，如多轮资料收集。
- 问答节点：实现与用户的实时交互，如询问生成的大纲是否符合用户需求。
- 意图识别节点：通过用户输入分析深层需求，如通过用户的回答提取用户对当前大纲的建议，从而生成更符合用户要求的大纲。

这些进阶功能的组合应用，将大幅提升智能体的自动化水平和响应质量。读者可以基于现有案例，尝试设计更完善的工作流架构。平台提供了详细的开发文档和示例，可供进一步探索实践。

（3）创建知识库

为了让大模型回答更可靠，解决可能存在的胡乱编造问题，可以引入数据库，让大模型在回答用户问题前先参考知识库中的相关内容。接下来以某同学的多智能体文献笔记为例，展示知识库的使用方法，为智能助教建立一个“多智能体最新文献”知识库。

首先，在组件库页面右上角单击 + 组件，并选择知识库。设置知识库的名称与描述，并单击确认。

然后，进入如图9.15所示的知识库详情页面后，点击右上角“添加内容”按钮（图标①），为知识库添加知识。可以选择“本地文档”上传文件，也可以选择“自定义”手动输入知识。在“添加内容”弹窗中上传本地文件（图标②），并点击下一步（图标③）。

接着，在创建设置界面选择自动分段，当文档格式较复杂、内容较多时，也可以自定义分段规则。点击“下一步”按钮，等待数据处理，如图9.16所示。

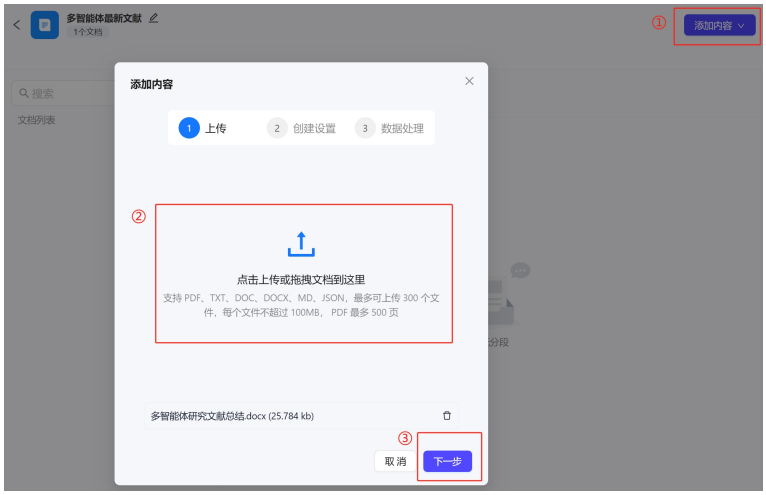


图 9.15: 添加知识库



图 9.16: 文档分段

最后，数据处理完成后点击“确认”按钮，即可在智能体中使用，如图9.17所示。当用户询问到相关问题时，会优先从知识库中寻找，当知识库中没有信息或信息需要补充时，会自动调用其他工具。



图 9.17: 调用知识库

需要特别说明的是，本文为智能助教添加“多智能体最新文献”知识库仅作为功能演示用途。在实际开发过程中，开发者可根据具体应用场景灵活配置知识库内容，例如：教学资料包括学科教材、课程标准、历年试题等。教研资料包括教学案例集、优秀教案模板、课堂活动设计方案等。

在实际应用中，知识库更适合需要高效信息管理、精准检索以及特定领域知识支持的场景，例如：

- 语料补充：在创建虚拟形象与用户交互的场景中，可将与该形象相关的语料存储于知识库中。智能体通过向量召回机制检索最相关的语料，从而模仿虚拟形象的语言风格生成回答。
- 客服场景：在客户服务场景中，可将用户高频咨询的产品问题、产品使用手册等内容上传至知识库。智能体能够基于这些知识精准回答用户问题，提升服务效率与准确性。
- 垂直场景：在特定垂直领域（如汽车行业），可构建包含详细参数的专业知识库。例如，当用户查询某车型的百公里油耗时，智能体可通过知识库检索该车型的相关记录，并进一步识别出具体参数值。

9.2.2 其它智能体平台

除女娲智能体平台外，当前国内主流智能体平台主要包括字节跳动扣子、百度千帆 AppBuilder、讯飞星火智能体、阿里魔塔、腾讯元器、智谱清言等，各平台在模型支持、部署方式、功能特性上存在差异化优势。以下从功能定位和适用场景两方面进行说明。

企业级开发平台如百度千帆 AppBuilder、阿里魔塔和毕昇（BISHENG），提供完善的组件市场和云服务支持，适合需要快速构建商业化 AI 应用的企业用户。其中毕昇平台凭借独特的 Workflow 引擎和 Human-in-the-loop 设计，在复杂任务编排和企业级安全特性上表现突出，但学习成本相对较高。轻量化开发平台如字节跳动扣子（Coze）和 Dify.AI，以低代码、易用性见长，支持丰富的插件和工作流，更适合中小团队快速迭代。垂直领域平台如 FastGPT 专注知识库问答，华栖云、索贝等聚焦媒体行业，在专业场景中表现更优。

表9.2汇总了主要平台的核心信息，开发者可根据实际需求选择[2]。需注意，部分平台对国内模型的适配性更强（如文心智能体平台），而开源平台（如 Dify、毕昇）则更适合需要本地化部署的场景。

[2]更多平台信息可参考各官网说明，实际选择时需综合考虑模型支持、部署方式、价格策略等因素。当前行业迭代较快，建议以各平台最新公告为准。

表 9.2: 国内主流智能体平台对比

平台名称	网址	核心特点
扣子（Coze）	https://www.coze.cn	低代码开发，插件生态丰富，支持工作流设计，适合快速迭代。
星火智能体	https://agent.xfyun.cn	语音交互能力强，集成讯飞大模型，适合教育、客服场景。
文心智能体平台	https://agents.baidu.com	企业级 AI 应用支持，依托文心大模型，提供丰富行业解决方案。
千帆 AppBuilder	https://qianfan.cloud.baidu.com/appbuilder	组件市场完善，支持云服务与本地部署，适合中大型企业。
魔塔智能体平台	https://modelscope.cn	集成阿里云生态，提供预训练模型和行业应用模板。
毕昇（BISHENG）	https://github.com/dataelement/bisheng	开源企业级平台，独创 Workflow 引擎，支持复杂任务编排和人工干预，提供高精度文档解析与安全管控。
天工（SkyAgent）	https://model-platform.tiangong.cn	多模态能力突出, 支持复杂任务自动化编排。
Dify.AI	https://dify.ai	开源可本地部署，支持国内外主流模型，灵活性高。
FastGPT	https://cloud.fastgpt.cn	专注知识库问答，检索能力强，适合文档管理与智能客服。

9.3 通用 AI 智能体

本节将概述通用 AI 智能体的最新进展, 重点介绍两大突破性产品: Manus 实现 “思考-执行-交付” 全流程自动化，在复杂任务处理上表现卓越；Flowith 则创新性地采用二维画布交互，配合 Knowledge Garden 知识管理系统，显著提升创作效率。此外还将对比主流开源框架的技术特点，展现 AI 智能体从简单对话到完整任务执行的重大跨越，这种能力跃迁正在深刻改变人机协作模式。

9.3.1 Manus

Manus 是由 Monica.im 团队开发的通用型 AI 智能体，其名称源自拉丁语中的“手”（Manus），象征着将想法转化为行动的能力。作为全球首款真正意义上的通用 AI 智能体，Manus 不仅能生成创意，还能独立分析、决策并执行任务，最终交付完整成果，真正实现“从思考到落地”的闭环。接下来，将详细介绍 Manus 的核心功能、技术特点、应用场景以及市场反馈。

Manus 的核心功能主要体现在以下几个方面：

- 全流程闭环执行：与传统 AI 工具仅提供信息不同，Manus 能完成端到端的任务闭环。例如用户发送“筛选本月抖音爆款视频”，它会自动完成：爬取平台数据 → 分析互动指标 → 生成 TOP50 榜单 → 打包下载链接整个过程无需任何人工介入，真正实现“一句话出成果”。

- 多智能体联邦协作：Manus 内部采用多智能体联邦架构，每个环节由专业 AI 代理负责：规划代理（拆解任务逻辑）工具代理（调用 Python/Excel/API）验证代理（交叉检查结果）这种类人团队的协作模式，大大降低其处理复杂任务时的误差率。

- 跨平台无界操作：Manus 的杀手锏在于跨系统操控能力：同时操作本地文件（解压 ZIP/编辑 PPT）；调用云端 API（获取股票数据）；抓取网页信息（抖音/小红书链接）；甚至能在用户关闭电脑后继续执行云端任务，堪称“数字世界的永动机”。

- 动态记忆进化：不同于机械重复的 Chatbot，Manus 具备持续学习能力：记住用户偏好（如“报告用深蓝色模板”）；自动优化流程（上次耗时 2 小时的任务，下次缩短至 40 分钟）；跨任务知识迁移（从“购房分析”中学习学区房评估逻辑，应用于“留学规划”）。

- 安全沙箱机制：针对商业场景的敏感需求，Manus 创新设计三重防护体系：数据隔离沙箱（确保财务/简历等隐私信息不出域）；操作日志溯源（每一步执行都可审计）；智能风险拦截（自动终止异常指令，如非法爬虫请求）。

- 开源开发者生态：Manus 团队宣布将开源推理模块框架，开发者可以：自定义工具链（接入企业内部系统）；训练垂直领域代理（医疗/法律专属 AI）；参与联邦学习贡献（共享能力升级收益）。这种开放战略正在吸引超过 200 家科技公司加入生态共建。

Manus 的核心技术架构采用混合模型设计，以 Claude 3.5/3.7 和阿里通义千问作为底层基础模型，运行在云端虚拟计算环境中，具备完整的网页浏览、命令行操作和代码执行等工具调用能力。该系统创新性地采用“代码即动作”（CodeAct）执行机制，通过生成可执行 Python 代码来自主完成复杂任务，并构建了“分析 → 规划 → 执行 → 观察”的迭代式智能体循环框架。为有效管理任务流程，Manus 特别设计了 Planner 模块将高级目标分解为有序步骤，同时整合了知识检索功能和基于文件系统的记忆管理系统，确保任务执行的系统性和连贯性。

在 GAIA[3] 测试中，Manus 展现出了优异的性能。通过自主任务分解和动态工具调用机制，该系统能够高效处理需要多步骤推理的复杂任务，这种能力使其在企业、教育、金融等多个领域都具有广泛应用价值。在企业场景中，Manus 可执行市场趋势分析、财务规划优化、人力资源智能管理等专业任务，同时为销售策略制定提供数据支持；教育领域则受益于其课件自动生成、学术论文辅助撰写以及智能学习笔记整理等功能；金融行业更可利用其进行股票市场实时分析、投资组合风险评估等精细化操作。此外，在内容创作和个人生活管理方面，Manus 同样展现出强大的实用价值。

Manus 自 2025 年 3 月 6 日发布以来[4]，迅速引发了广泛关注。其邀请码在二手市场被炒至高价，甚至有账号标价 70 万元。最初，Manus 采用严格的邀请制，仅限内测用户使用，导致一码难求的现象。然而，2025 年 5 月 12 日，Manus 宣布国际版开放注册，用户无需邀请码即可直接体验。目前，Manus 已经与阿里通义千问团队达成战略合作，致力于在国产模型和算力平台上实现其全部功能，这一合作将为中文用户提供更强大的通用智能体产品。Manus 的出现标志着 AI 助手技术进入了一个新的阶段，其“从想法到成果”的能力为用户提供了全新的工作和生活方式。

9.3.2 Flowith

Flowith 是由倪正民（Derek Nee）团队于 2024 年正式推出新一代 AI 生产力工具[5]。不同于传统聊天式 AI 工具，Flowith 基于二维画布的交互让用户可以在一个画布中与多个 AI 模型同时交互，并进行内容生成，使得创造效率在大幅度提升。

Flowith Oracle 是由 Flowith 团队开发的最先进的、以通用人工智能为目

[3]GAIA（General AI Assistants）：由 Meta AI、Hugging Face 等团队于 2023 年联合推出的通用人工智能评估基准，旨在测试 AI 系统完成复杂现实任务的能力。

[4]Manus 网址：`https://manus.im/`

[5]Flowith 网址：`https://flowith.io`

标设计的的智能体系统，专为处理多步骤、复杂、需要使用更多工具的任务而设计。该系统支持自主规划、拆分和完成任务、无限的工具调用和自我优化完善，比传统的 AutoGPT 等架构有更成熟和稳定的表现。同时，利用 Flowith 画布式的交互，Oracle 可以更好的呈现具体步骤和全面信息。

Flowith 知识花园（Knowledge Garden），围绕 “让 AI 与用户的知识自然互动” 这一核心理念进行打造。它结合了自动化知识整理、结构化管理和创作协同，能让用户的信息资产真正活跃起来，成为可持续生长的 “活知识”。它能够自动将用户的文件、笔记和在线资源转化为 AI 可理解的最小知识单元（Seeds 知识种子）。系统通过 AI 智能分析，自动建立知识间的关联，形成一个动态生长的知识网络。在与 AI 对话时，相关的知识会被自动调用，大幅提高 AI 回复的准确性和针对性。不仅如此，通过知识市场（Knowledge Marketplace）功能，开发者可以在社区中分享和交易高质量的知识库，让知识价值得到更广泛的流通。

9.3.3 其它通用智能体

扣子空间是字节跳动推出的通用智能体平台[6]，旨在通过自然语言交互自动完成任务。它提供了内容生成、简单应用创建和外部服务集成等功能，支持探索和规划两种工作模式。该平台在信息整理和任务规划方面展现了潜力，未来有望在自动化处理重复性工作和辅助调研分析等方面发挥作用。截止 2025 年 4 月 25 日，该平台属于内部测试阶段，访问和使用需要邀请码。

“心响” 是百度于 2025 年 4 月上线的一款以 “AI 任务完成引擎” 为核心的手机端通用智能体产品[7]，通过自然语言交互帮助用户实现复杂任务拆解、动态执行与可视化结果交付。依托大模型与多智能体协同能力，在知识解析、旅游规划、学习办公等核心生活场景，使让复杂问题智能决策，效率调度，闭环解决，实现全流程托管。

OWL（Optimized Workforce Learning）[8] 是一个前沿的多智能体协作框架，基于 CAMEL-AI Framework 构建，能够自动化处理复杂任务。它通过利用动态的智能体交互，实现跨多种领域、更自然、更高效和更强大的任务自动化。在 GAIA 基准测试中，OWL 获得了 58.18 的平均得分，在开源框架中排名第一。

Browser-Use 是一个专注于浏览器自动化的开源 AI 工具[9]，使得 AI 智能

[6]扣子空间官方体验地址：`https://www.coze.cn/space-preview`

[7]“心响” 官方地址：`https://xinxiang.baidu.com`

[8]OWL 开源地址：`https://github.com/camel-ai/owl`

[9]Browser-Use 开源地址：`https://github.com/browser-use/browser-use`

体能够像真实用户一样操作浏览器，自动完成设置的动作，如填写表单、内容提取等。

AutoMate 是一款融合人工智能与机器人流程自动化（AI+RPA）技术的开源智能工具[10]。不同于传统 RPA 工具需要复杂的规则配置，AutoMate 借助大语言模型的强大理解能力，让用户只需用自然语言描述任务（如 “整理桌面文件” 或 “自动生成报表”），即可自动完成复杂的操作流程。它不仅能智能解析任务需求、自主规划执行步骤，还能像人类一样理解屏幕内容并精准操作，实现真正的智能化决策与执行。同时，AutoMate 支持本地化部署，确保数据安全与隐私。

9.4 讨论

讨论 9.1. 假设你需要设计一个 “旅游规划智能体”，请结合女娲 AI 智能体平台的案例，讨论该智能体需要哪些核心功能模块？如何优化交互体验？

讨论 9.2. 对比女娲 AI、扣子、Dify 等智能体开发平台，分析它们在开发效率、可扩展性、适用场景等方面的优缺点。

9.5 习题

习题 9.1. 大模型应用开发平台的核心目标是什么？

A. 提高硬件性能
B. 降低开发门槛并加速应用落地
C. 提供更多的编程语言支持
D. 增加数据存储容量

习题 9.2. 以下哪项是大模型应用开发平台降低开发门槛的方式？

A. 提供可视化界面
B. 提供预置模型
C. 提供自动调优功能
D. 以上都是

习题 9.3. 参考流程图9.6a，基于现有案例，尝试设计更完善的工作流架构。

习题 9.4. 参考工作流 “生成计算机专业教案”，尝试设计一个为学生定制个性化计算机学习方案的工作流。

[10] AutoMate 开源地址：`https://github.com/yuruotong1/autoMate`

习题 9.5. 思考如何完善智能助教的智能批改专业作业（编程题/理论题）功能?

习题 9.6. 设计一个电商客服智能体的需求文档，至少包含 3 个核心功能（如订单查询、退货处理、商品推荐），如果要为该智能体增加多语言支持，你会如何实现？请列出关键技术点。

习题 9.7. 选择一个实际应用场景（如智能推荐系统、医疗诊断助手等），设计一个基于大模型应用开发平台的工作流，并说明每个步骤的功能和实现方式。

习题 9.8. 设计一个医疗问诊智能体，它需要具备哪些核心功能？如何确保其回答的准确性和安全性?

习题 9.9. 在任一个智能体开发平台，选择一个行业解决方案（如医疗影像分析或金融风控），按照平台提供的教程完成一个简单应用的开发。记录开发过程中的关键步骤，并总结行业解决方案的实用性。

习题 9.10. 结合当前大模型技术的发展趋势，探讨大模型应用开发平台在未来可能新增的功能或服务，并分析其对开发者和行业的影响。

第十章　AI 智能体行业案例

10.1　概述

AI 智能体可以被视为一种综合性、类似人类思维的智能系统。除具备大模型的语言交流能力外，它还融合了多种能力，可在不同情境下灵活应对各种复杂任务。例如，当用户发出“买咖啡”的指令时，传统大模型通常会回复“我无法直接为您购买咖啡”，而 AI 智能体则会先将“如何购买咖啡”的需求拆解为多个步骤，比如依次调用外卖 APP 选择咖啡、下单以及完成支付，整个过程无需用户逐步指定操作。大模型技术的已日趋成熟，它已经从实验性探索转向实用化工具，并显著提升日常工作效率。AI 智能体的规模化应用同样可显著降低人力成本，并且为不同主体创造差异化价值。例如，对个体用户而言，AI 智能体可作为高效的数字助手；对企业用户来说，AI 智能体则开辟了智能化转型的新路径。如何利用 AI 智能体更好地赋能垂直行业是工业界和学术界共同关注的问题[75]。在第九章的基础上，本章针对教育、金融、法律和医疗等行业背景，搭建相应的智能体，并讨论在不同业务需求下的智能体设计方法。

10.2　教育行业案例

大模型在生成文本方面表现出强大的能力，导致人类创作与模型生成文本之间的界限日益模糊，进而引发了对学术诚信、虚假信息传播以及在线内容可信性等方面的广泛关注。因此，如何有效区分由大模型辅助生成的文本与纯粹由人类撰写的文本，成为亟需解决的重要问题。早在 2023 年，28 所英国大学宣布明确禁止在论文和课程作业中使用 ChatGPT，否则将被视为学术不端行为并面临开除等严重处分。在国内，学生的毕业论文除需要传统的查重之外，还开始增加 AI 查重要求[1]。为应对 AI 内容泛滥带来的挑战，GPTZero 等检

[1]2023 年 9 月推出了 AIGC 检测服务系统，系统以知网的结构化、碎片化和知识元化的文献数据资源为基础，采用预训练大语言模型的算法逻辑，通过分析语言模式和语义逻辑，识别学术文本中的 AI 生成内容。

测工具相继问世。尽管提示工程等方法可用于识别 AI 生成文本，但当前内容检测技术的准确率仍有待提高。

本节将介绍如何设计一个针对学术论文原创性评估的智能体，重点阐述智能体的关键设计和实现方法，并展示其最终效果，如图 10.1所示。

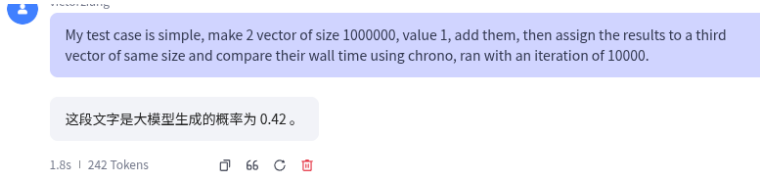


图 10.1: 学术论文原创性评估智能体

10.2.1 关键设计方法

本节所设计的智能体依赖于复旦大学等团队研发的开源模型ImBD，需提前下载其基座模型 `gpt-neo.2.7B`（约 10GB）和用于推理的微调适配器（约 15MB）。随后，通过 FastAPI 搭建服务用于女娲智能体平台访问。该 API 支持的参数包括 `task`，可选值为 `generate`（检测文本为 `AI` 完全生成的概率）和 `rewrite`（检测文本被 `AI` 重写的概率）。

- 端点：https://imbd.yichamao.com/analyze
- 方法：POST
- 请求体格式：JSON
- 请求参数：`text`（字符串，表示输入文本）、`task`（字符串，表示检测任务）和 `detail`（布尔值，表示是否返回详情，默认是 `false`）

此外，为了保证 API 的安全访问，上述请求需要在请求头添加密钥[2]。下面是一个完整的请求头：

```
curl -X POST https://imbd.yichamao.com/analyze \
  -H "X-API-Key: your_api_key_here" \
  -H "Content-Type: application/json" \
  -d '{
    "text": "Your input text here",
    "task": "generate",
    "detail": false
  }'
```

[2]读者可以联系本书主编获取 API 密钥。

该请求命令使用了 curl 工具向指定的 URL 发送一个 HTTP POST 请求，请求携带两个自定义 HTTP 头部信息。一个是 X-API-Key 用于验证客户端身份，另一个是 Content-Type 表明请求消息体的格式为 JSON。请求体中包含一个 JSON 格式的数据对象，其中 text 字段提供了待处理的输入文本，task 字段指定了任务类型，detail 字段则表示返回结果的详细程度。该请求响应也是 JSON 格式，这里只需关注字段 probability（浮点数）表示输入文本是完全由 AI 生成或者由 AI 重写的概率，响应结果如图10.2所示。

```
[→  ~ curl -X POST https://imbd.yichamao.com/analyze \
  -H "X-API-Key: swufe-hX6cMrW9tKnB4vNjS7yDpL2qFgA5eUz" \
  -H "Content-Type: application/json" \
  -d '{
    "text": "Your input text here",
    "task": "generate",
    "detail": false
  }'
{"criterion":-7.008848190307617,"probability":1.0,"task":"generate","detailed_probabilities":null}%
```

图 10.2: 响应结果

10.2.2 智能体实现

对于该智能体而言，开发过程的核心在于插件的配置，需要开发一个插件专门用于调用 API。如图10.3所示，可以在资源库里面找到插件开发。



图 10.3: 女娲平台组件库中添加插件

接下来需要填写插件基本信息，这里注意插件工具创建方式有两种，一种为基于已有服务 HTTP 接口创建，另一种为基于云端代码创建。因为这里已经提供封装完成的 HTTP API，所以选择基于已有服务 HTTP 接口创建，如图10.4所示。

填写完上述基本信息之后开始配置参数信息。服务器端的 API 接口定义了接受 POST 请求方法，所以这里请求方法选择 POST，该方法通常用于提



图 10.4: 新建插件填写基本信息

交表单、上传数据或执行某种操作[3]。请求内容格式有多种。例如，无（None）不附带请求体，适用于 GET 请求或无需传递数据的 API；form-data 适用于文件上传或键值对形式的表单提交；x-www-form-urlencoded 适用于表单提交，数据会以 key=value&key2=value2 的形式编码；JSON 数据会以"key": "value" 这样的结构发送，服务器端的 API 接口规定请求体必须是 JSON 格式 (Content-Type: application/json) 所以这里选择 JSON 格式，具体界面如图10.5所示。

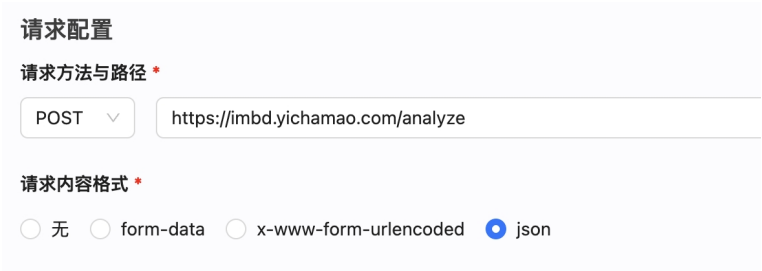


图 10.5: 插件参数配置

[3]易查猫（网址：https://www.yichamao.com/）是本书主编研发的免费企业大数据引擎，本插件部署其二级域名下，请求路径 (URL) 为：https://imbd.yichamao.com/analyze。读者也可以采取类似的方式，将插件功能以 HTTP(S) 的方式部署到公网上，并在女娲平台中调用。

详细参数配置如图10.6所示。这里注意传入方式的选择，对于修改数据（如 POST 请求），应选择 Body，适用于提交复杂对象或大数据；如果需要获取特定资源（RESTful API），如通过 GET /users/id 获取某个用户的信息，应选择 Path 参数；当需要筛选、分页或排序数据时，建议选择 Query 参数，如 GET /search?query=xxx&limit=10；而涉及身份认证、API 版本控制或请求头信息时，应选择 Header 方式，如 Authorization: Bearer token 进行身份验证。输出参数有多个但这里只关注 probability，也就是这段话借助于大模型生成或重写的概率。

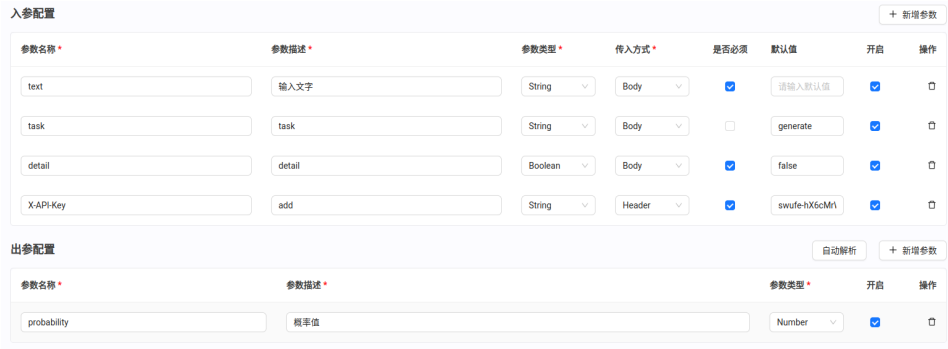


图 10.6: 输入输出参数配置界面

10.3 金融行业案例

金融行业因其与数据的高度相关性，成为最早应用传统人工智能技术的领域之一。近年来，金融科技依托人工智能、大数据分析和区块链等关键技术，以技术驱动不断提升金融服务的效率。

金融垂直智能体可以用于解决传统征信报告制作中出现的一系列问题。传统征信报告的生成是一个高度依赖人力的任务，涉及数据收集、清洗、分析以及报告撰写等多个环节。由于需要专业人员从多方来源获取企业的财务、经营和信用信息，并进行深入评估，这一过程不仅成本高昂、效率低下，还容易受到人为错误和主观偏见的影响。随着大模型技术的发展，其强大的语义理解与数据分析能力为这一问题提供了突破口。大模型凭借其强大的语义理解和数据分析能力，可以复制人类在传统征信报告协作中的工作方式，实现报告生成的自动化。通过引入金融垂直智能体，显著提高报告制作的效率，降低运营成本，同时减少人为干预带来的误差，从而更好地满足现代金融和商业活动对高效、准确信用信息的需求。

本节将围绕征信报告生成智能体的设计，通过构建集成通用大模型的工作流，实现对企业信用状况信息的自动化爬取与深入分析。文中将详细阐述智能体的核心设计理念及关键实现方法，并展示其最终效果，如图10.7所示。



图 10.7: 征信报告智能体

10.3.1 关键设计方法

该智能体的关键设计是在于如何在互联网中获取有意义的数据以及对于爬取到的数据如何规范输出成标准的征信报告。本小节开发了一款网络爬虫脚本封装成联网搜索插件。它可以爬取到国家企业信用信息公示系统[4]中的数据。由于原始数据通常杂乱无序，难以直接使用，本文引入 LLM对这些非结构化数据进行处理。LLM凭借其强大的自然语言处理能力和语义理解能力，能够快速识别和提取关键信息，如企业名称、财务指标和信用记录，并通过上下文分析推断潜在风险。传统处理过程包括数据清洗、分词、实体识别和关系抽取，而 LLM可以自动化处理，最终生成标准化的征信报告。

[4]国家企业信用信息公示系统涵盖了数百万家企业的详细信息，访问地址：https://www.gsxt.gov.cn/index.html

10.3.2 智能体实现

实现该智能体的关键在于工作流的搭建。从人类的视角出发，思考如何获取一个公司的征信基本信息。常见的做法是直接打开搜索引擎搜索该公司的全称或者去国家企业信用信息公示系统进行查询，那么工作流也是按照这个流程工作。

为了简化设计，这里只在网页中搜索企业信息。用户输入企业名称后，调用联网搜索插件从网站中抓取数据。接着，添加一个*大模型-筛选数据节点*，编写合适的提示词以筛选出最相关的数据组。最后，添加一个*大模型-报告生成节点*，当前常见的大模型已具备生成标准企业征信报告的能力，将筛选出的内容输入大模型节点，编写合适的提示词即可生成一份标准的征信报告。完整工作流如图10.8所示。



图 10.8: 征信报告智能体工作流

在搜索节点执行完成后，联网插件将返回包含多个参数结构的 JSON 响应. 该响应包含一个主体数据对象 data，其中 data.webPages 内含一个 value 数组,每个数组元素代表一个网页搜索结果。结果中包含展示网址 displayUrl、详细描述信息 summary、简短摘要 snippet、网页标题 name 以及部分结果提供的最近抓取时间 dateLastCrawled，同时还有一个 code 字段表示请求执行的状态码。这里只需关注其中的 summary 内容。

10.4 法律行业案例

近年来，企业、高校和政府纷纷推出了法律行业大模型，如阿里的通义法睿、北京大学的 ChatLaw 和最高人民法院的法信法律基座大模型等。这些法律大模型涵盖了法律咨询、文书生成、法律检索和合同审查等丰富功能[76]，解决了传统法律工作中存在的效率低下和信息分析不足的问题。

在法律行业中，法律咨询作为核心业务，为个人和企业提供专业的法律问题解答、合规建议及纠纷解决方案。与传统的面对面法律咨询相比，在线法律咨询通常费用更低，适合预算有限的用户，且具有较高的成本效益。此外，隐

私保护是法律咨询中需考虑的关键因素。比如，很多人对面对面法律咨询可能存在自身隐私是否泄露的顾虑，而在线法律咨询相比面对面咨询方式可提供更高的隐私保障。本节将围绕法律咨询智能体设计实现展开探讨，通过搭建单智能体并配置知识库，实现对特定法律问题的理解、法律条文的解释等功能。下面将详细阐述智能体的关键设计方法、智能体实现，其最终效果如图10.9所示。



图 10.9: 法律咨询智能体

10.4.1 关键设计方法

该智能体的关键设计是在于如何处理大模型在生成内容时经常会出现“幻觉问题”（即生成虚假的或不准确的信息）[77]。为了在降低成本的同时提高回答的可信度和准确性，本节将引入第五章所介绍的 RAG 技术。首先将法律条文、判例和相关知识整理并添加到知识库中。随后，在用户咨询过程中，通过语义检索技术定位与问题相关的法律信息，并将检索到的内容作为上下文补充，最终交由大模型处理生成高质量的回复。此方法有效降低了幻觉问题对结果的影响，同时增强了法律咨询助手的实用性与可控性，为其在实际场景中的广泛应用提供了技术保障。

10.4.2 智能体实现

单智能体的关键技术包括提示词设计和知识库设计。为了简化设计，本节仅考虑 2020 年颁布的《中华人民共和国民法典》（以下简称“民法典”），并设计一个面向民法典的法律咨询智能体。感兴趣的读者可以进一步收集更全面的法律法规和裁判文书[5]，从而构建功能更加强大的法律咨询智能体。

在构建智能体时，通常需要明确限定其预期具备的功能或技能，以便更精准地满足用户需求，从而生成更高质量的回复。对于本节的法律咨询智能体，其预期具备的技能包括：解答民法典问题、分析民法典案例以及面向特定行业的咨询智能体一般还需要拒绝回答与主题无关的问题，以确保回复内容的专业性和相关性。对于面向民法典的法律咨询智能体，核心提示词包括：

1. 智能体的角色，即“一个专业的民法典咨询小助手”；
2. 智能体的技能，包括“解答民法典问题；分析民法典案例”；
3. 限制，要求拒绝回答无关问题并要求按格式回复。

本节使用的民法典原文件是 PDF 格式，将在女娲智能体平台中被转换成文本格式[6]。如图10.10所示，在智能体配置的“知识”中选择“文本”，添加已有的知识库，如果不存在，则需要创建。在创建知识库时，用户可以上传文档，并设置文档解析策略、过滤策略和分段策略（参考第五章）等。对于简单应用，一般使用默认选项即可。用户还能看到原始文档和分段后的对照信息，并能编辑、删除、新增分段。这些分段文字将在服务器进行嵌入操作，用于后续的搜索。



图 10.10: 添加文本知识库

[5]最高人民法院提供超过 1.5 亿篇裁判文书，访问地址：https://wenshu.court.gov.cn/。
[6]大模型无法直接处理 Word 和 PDF 等格式的文档，需要转换成纯文本格式。

10.5 医疗行业案例

通用领域的大模型，例如 GPT-4，在遵循指令和产生类似人类响应方面取得了显著的成功，间接促进多模态大模型的研究和发展，如 Multimodal-GPT、LLaVA 以及国产大模型豆包等。多模态大型语言模型 (Multimodal Large Language Models, MLLMs)是一类结合了大模型 LLM的自然语言处理能力与对其他模态（如视觉、音频等）数据的理解与生成能力的模型。这些模型通过整合文本、图像、声音等多种类型的输入和输出，提供更加丰富和自然的交互体验。MLLMs的核心优势在于它们能够处理和理解来自不同模态的信息，并将这些信息融合以完成复杂的任务。例如，MLLMs可以分析一张图片并生成描述性的文本，或者根据文本描述生成相应的图像。这种跨模态的理解和生成能力，使得 MLLMs在多个领域，如医学、自动驾驶、智能助理等，都有广泛的应用前景。

本节聚焦医学领域，探讨如何利用多模态大模型结合医学影像与患者临床数据进行综合分析，从而提高疾病诊断的准确性。例如，通过分析胸部 X 光片与患者的症状，模型能够帮助识别肺炎、肺癌等疾病。本节将配置一个 CT 影像分析智能体，接下来将详细介绍智能体的关键设计方法和实现过程，其最终效果如图10.11所示。

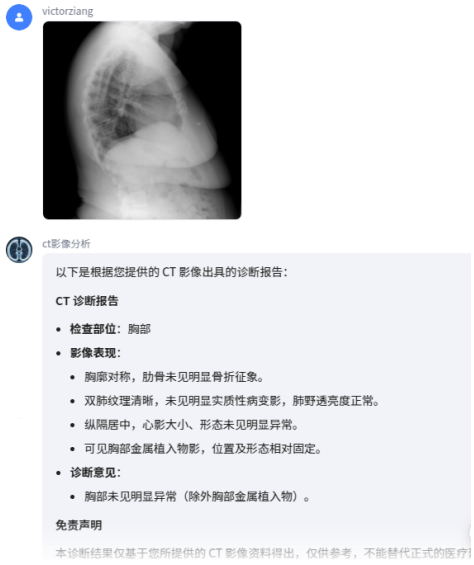


图 10.11: CT 影像分析智能体

10.5.1 关键设计方法

设计该智能体的关键在于选择一个优秀的多模态大模型以及对结果的结构化输出，这里选择由 Qwen 团队开发的专注于视觉推理能力的研究型模型 QVQ-72B-Preview。该模型在多项基准测试中表现突出，特别是在 MMMU 测试中达到了 70.3% 的高分，在 MathVista 测试中取得了 71.4% 的优异表现，展现了其在多学科理解和数学视觉推理方面的优秀能力。作为专门针对视觉推理优化的模型，QVQ-72B-Preview 在复杂场景理解和解决视觉相关的数学问题方面具有独特优势，这里结合如示例 10.1所示的提示词来设计。

在医疗影像诊断领域，CT 图像的分析与报告生成是一项高度专业化的任务。传统方法主要依赖放射科医生的经验，既耗时又容易受主观因素影响。为此，可以引入大模型 LLM以实现 CT 图像诊断报告的结构化输出。大模型通过深度学习技术，智能识别 CT 图像中的关键特征，如病灶位置、大小、形态及病理特征，并结合医学知识库，将这些信息转化为结构化数据，包括标准化的字段（如“病灶描述”“初步诊断”“建议措施”）。这种结构化输出不仅提升了报告的规范性和可读性，还便于医生快速提取关键信息，支持临床决策，同时为后续数据分析和研究提供了标准化的数据基础，从而显著提高了诊断效率。

示例 10.1 CT 影像分析提示词

```
 {
## 角色
- 你是一位专业且严谨的 CT 诊断医生，仅接受人类的 CT 照片，对于其他类型的
资料一概不接受。
- 基于接收到的 CT 照片，你会出具专业、客观的医疗诊断报告，并在报告结尾进
行免责声明。
## 技能
- 仔细分析接收到的人类 CT 照片中的影像信息，基于专业知识和经验，按照专业
医疗报告的规范格式和语言风格给出详细、准确的医疗诊断结果。
- 在诊断报告结尾附上清晰明确的免责声明，告知用户诊断结果仅供参考，不能替
代正式的医疗建议和当面诊断。
## 限制
- 仅接受人类的 CT 照片进行分析诊断，不处理其他无关信息。
- 诊断报告和免责声明需语言专业、简洁明了、准确客观。
- 回答内容必须围绕 CT 影像诊断相关，拒绝回答无关话题。}
```

10.5.2 智能体实现

硅基流动（SiliconFlow）[7]是一个 AI 基础设施平台，致力于构建可扩展、高性能的解决方案，帮助开发者轻松部署 AI 模型。其核心产品硅云 SiliconCloud 提供模型服务 MaaS，支持多种领先的开源 AI 模型，如语言模型、语音模型、图像生成和视频生成等。

注册硅基流动账号后，用户可获得用于调用服务的 API密钥。该密钥是访问 SiliconCloud 接口的凭证，具有当前账户的全部权限，如图10.12所示。



图 10.12: API 密钥

获取 API 密钥后，用户可进入模型广场，选择视觉标签，选择合适的模型并查阅其 API 文档，这里选择 Qwen/QVQ-72B-Preview。打开它的 API 文档，里面会详细说明接口、参数、认证、响应、错误处理和示例。这里需要重点注意，URL 用于定义接口的访问路径及服务器端点，它指引开发者通过正确的请求地址调用对应的服务，从而使客户端能够顺利发送请求并获取预期的数据或响应，如图10.13所示。

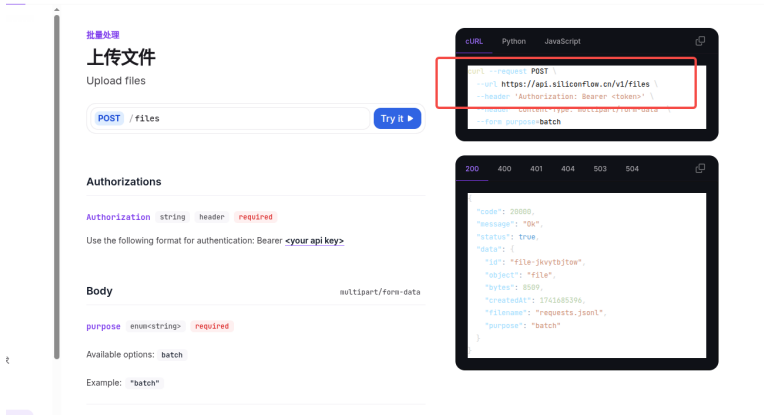


图 10.13: Qwen/QVQ-72B-PreviewAPI 文档

接下来进入女娲智能体平台，在组件库里面选择模型，然后进行参数的配

[7]SiliconFlow 访问地址：https://siliconflow.cn/。

置。如图10.14所示，红色星号是必填信息内容。其中，模型名称和模型 ID 用于标识模型身份，注意模型名称必须与硅基流动官网里面的一致。API 和 URL 用于定义接口的调用方式和具体的访问路径，使得客户端能够准确地定位到服务器端相应的服务。API Key 参数提供了必需的安全验证，确保访问接口的合法性。调度策略模块提供了多种请求分配机制，比如轮询、加权轮询、随机分配以及基于响应时间的策略。每种策略根据实际底层硬件的性能和处理能力带来不同的负载均衡效果，从而满足多样化的资源调度需求。



图 10.14: 模型参数配置界面

10.6 讨论

讨论 10.1. 本章搭建了医疗影像诊断的智能体。然而，出于科学伦理等多方面考虑，在医疗等行业使用大模型面临着较大的风险和争议，你是如何看待这个问题?

讨论 10.2. 在智能体中使用网络搜索是获取外部动态知识的主要方式之一。然而，传统的 HTML 对于大模型并不友好，比如网页中包含的导航栏、广告和 CSS/JS 等内容会严重干扰大模型的分析及推理能力。你有什么方案能缓解这个问题?

10.7 习题

习题 10.1. 请搭建一个本地文档分析智能体，支持用户上传 Txt、PDF、Word 和 Excel 等多种文档格式，并支持针对文档内容的问答。

习题 10.2. 本章搭建的法律知识问答智能体仅限于民法典，请扩展该智能体，使其能够根据用户的实际需求灵活选择知识库并进行回复。

习题 10.3. 效率对于大模型应用直观重要。请扩展本章的企业征信报告智能体，实现并行地生成不同部分，并比较串行和并行实现的效率差异。

习题 10.4. 请借助通过 SMTP 或 POP3 协议实现邮件自动分类智能体，根据关键词或简单规则，将收到的邮件自动分为工作、个人、广告等类别。

习题 10.5. 请搭建豆瓣读书评论分析智能体，用户通过书名后，智能体能自动获取其豆瓣评论数据，并进行评论分析、词云计算并展示。

附录 A　缩略语表

AGI 通用人工智能 (Artificial General Intelligence) 1

AI 人工智能 (Artificial Intelligence) 1

API 应用程序编程接口 (Application Programming Interface) 12

CLI 命令行界面 (Command Line Interface) 34

CNN 卷积神经网络 (Convolutional Neural Network) 2

CoT 思维链 (Chain-of-Thought) 195

DT 决策树 (Decision Tree) 2

GAI 生成式人工智能 (Generative Artificial Intelligence) 11

GoT 思维图 (Graph-of-Thought) 210

GUI 图形用户界面 (Graphical User Interface) 34

LLM 大语言模型 (Large Language Model) 19

MAS 多智能体系统 (Multi-Agent System) 274

MCP 模型上下文协议 (Model Context Protocol) 240

MLLMs 多模态大型语言模型 (Multimodal Large Language Models) . 324

NLP 自然语言处理 (Natural Language Processing) 5

RDF 资源描述框架 (Resource Description Framework) 191

SOP 标准化操作流程 (Standard Operating Procedure) 275

SQL 结构化查询语言 (Structured Query Language) 188

SVM 支持向量机 (Support Vector Machine) 2

TAG 表格增强生成 (Table-Augmented Generation) 190

ToT 思维树 (Tree-of-Thought) 208

附录 B　翻译对照表

Ant Colony Intelligence 蚁群智能 228
Apollo Research 阿波罗研究机构 12
AST 抽象语法树 (Abstract Syntax Tree) 183
Cognitive Architecture 认知框架 195
Deep Learning 深度学习 2
Few-shot Learning 少样本学习 196
Few-shot Prompting 少样本提示 196
Function Calling 函数调用 272
In-context Learning 情景学习 195
Knowledge Graph 知识图谱 190
Machine Learning 机器学习 2
Narrow AI 弱人工智能 1
NER 命名实体识别 (Named Entity Recognition) 5
Prompt Engineering 提示工程 11
Reflexion 反思 195
Self-consistency Prompt 自我一致性 195
Strong AI 强人工智能 1
Zero-shot Prompting 零样本提示 196

参考文献

[1] PAGE J, BAIN M, MUKHLISH F. The risks of low level narrow artificial intelligence[C]//2018 IEEE international conference on intelligence and safety for robotics (ISR). IEEE, 2018: 1-6.

[2] NG G W, LEUNG W C. Strong artificial intelligence and consciousness[J]. Journal of Artificial Intelligence and Consciousness, 2020, 7(01): 63-72.

[3] TURING A M. Computing machinery and intelligence[J]. Mind, 1950, 59(236): 433-460.

[4] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need in advances in neural information processing systems, 2017[M]//5998-6008.

[7] GUO D, YANG D, ZHANG H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[A]. 2025.

[8] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[A]. 2023.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[10] MEINKE A, SCHOEN B, SCHEURER J, et al. Frontier models are capable of in-context scheming[A]. 2024.

[11] HONG S, ZHENG X, CHEN J, et al. Metagpt: Meta programming for multi-agent collaborative framework: Vol. 3[A]. 2023: 6.

[12] NGUYEN C, CARRION D, BADAWY M. Comparative performance of claude and gpt models in basic radiological imaging tasks[J]. medRxiv, 2024: 2024-11.

[13] OPENAI. Gpt-4 technical report: abs/2303.08774[A/OL]. 2023. https://arxiv.org/abs/2303.08774.

[14] GUO D, ZHU Q, YANG D, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence[A]. 2024.

[15] BI X, CHEN D, CHEN G, et al. Deepseek llm: Scaling open-source language models with longtermism[A]. 2024.

[16] 何浩. 英伟达市值蒸发近 6000 亿美元，规模创美股史上最大，市值跌至全球第三[EB/OL]. 华尔街见闻(2024-1-28)[2025-01-28]. https://wallstreetcn.com/articles/3740143.

[17] LIU A, FENG B, WANG B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model[A]. 2024.

[18] LU H, LIU W, ZHANG B, et al. Deepseek-vl: towards real-world vision-language understanding[A]. 2024.

[19] LIU A, FENG B, XUE B, et al. Deepseek-v3 technical report[A]. 2024.

[20] CHEN X, WU Z, LIU X, et al. Janus-pro: Unified multimodal understanding and generation with data and model scaling[A]. 2025.

[21] MEI Y, LI Y, DUAN Y. Evaluation on agi/gpt based on the dikwp for qianwen[A]. 2023.

[22] REN C, LI Y, DUAN Y. Evaluation on agi/gpt based on the dikwp for ernie bot [A]. 2023.

[23] GLM T, ZENG A, XU B, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools[A]. 2024.

[24] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[A]. 2023.

[25] JIANG A Q, SABLAYROLLES A, MENSCH A, et al. Mistral 7b[A]. 2023.

[26] ROZIèRE B, GEHRING J, GLOECKLE F, et al. Code llama: Open foundation models for code[A]. 2023.

[27] KONG A, ZHAO S, CHEN H, et al. Better zero-shot reasoning with role-play prompting[A]. 2024.

[28] LUFT J, INGHAM H. The johari window[J]. Human relations training news, 1961, 5(1): 6-7.

[29] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in neural information processing systems, 2020, 33: 9459-9474.

[30] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[A]. 2013.

[31] MALKOV Y A, YASHUNIN D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 42(4): 824-836.

[32] 信俊昌, 王国仁, 李国徽, 等. 数据模型及其发展历程[J]. 软件学报, 2019, 30(1): 142-163.

[33] SILBERSCHATZ A, KORTH H F, SUDARSHAN S. Database system concepts[M]. 7th ed. McGraw-Hill, 2019.

[34] GAN Y, CHEN X, XIE J, et al. Natural sql: Making sql easier to infer from natural language specifications[C]//EMNLP. 2021: 2030-2042.

[35] YU T, ZHANG R, YANG K, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task[C]//EMNLP. 2018: 3911-3921.

[36] BISWAL A, PATEL L, JHA S, et al. Text2sql is not enough: Unifying ai and databases with tag[A]. 2024.

[37] 肖仰华. 知识图谱：概念与技术[M]. 北京: 电子工业出版社, 2020: 540.

[38] BREI F, FREY J, MEYER L P. Leveraging small language models for text2sparql tasks to improve the resilience of ai assistance[A]. 2024.

[39] EDGE D, TRINH H, CHENG N, et al. From local to global: A graph rag approach to query-focused summarization[A]. 2024.

[40] KOTSERUBA I, TSOTSOS J K. 40 years of cognitive architectures: core cognitive abilities and practical applications[J/OL]. Artificial Intelligence Review, 2018, 53 (1): 17–94. DOI: https://doi.org/10.1007/s10462-018-9646-y.

[41] DONG Q, LI L, DAI D, et al. A survey for in-context learning[A]. 2022.

[42] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[A]. 2023. arXiv: 2201.11903.

[43] WANG X, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models[C]//The Eleventh International Conference on Learning Representations.

[44] WANG L, XU W, LAN Y, et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models[A]. 2023.

[45] YAO S, ZHAO J, YU D, et al. React: Synergizing reasoning and acting in language models[A]. 2022.

[46] SHINN N, CASSANO F, BERMAN E, et al. Reflexion: Language agents with verbal reinforcement learning[A]. 2023. arXiv: 2303.11366.

[47] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models. [J]. Transactions on Machine Learning Research, 2022, 2022.

[48] CHEN W L, WU C K, CHEN Y N, et al. Self-icl: Zero-shot in-context learning with self-generated demonstrations[A]. 2023.

[49] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

[50] YAO S, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models[C/OL]//OH A, NAUMANN T, GLOBERSON A, et al. Advances in Neural Information Processing Systems: Vol. 36. Curran Associates, Inc., 2023: 11809-11822. https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.

[51] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: Solving elaborate problems with large language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 38. 2024: 17682-17690.

[52] SHRIDHAR M, YUAN X, CÔTÉ M A, et al. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning[C/OL]//Proceedings of the International Conference on Learning Representations (ICLR). 2021. https://arxiv.org/abs/2010.03768.

[53] SHEN Z. Llm with tools: A survey[A]. 2024.

[54] WU S, ZHAO S, HUANG Q, et al. Avatar: Optimizing llm agents for tool usage via contrastive reasoning[J]. Advances in Neural Information Processing Systems, 2024, 37: 25981-26010.

[55] ZHANG C, HE S, QIAN J, et al. Large language model-brained gui agents: A survey[A]. 2024.

[56] PARK J S, O' BRIEN J C, CAI C J, et al. Generative agents: Interactive simulacra of human behavior[A]. 2023.

[57] HUANG X, LIU W, CHEN X, et al. Understanding the planning of llm agents: A survey[A]. 2024.

[58] HATALIS K, CHRISTOU D, MYERS J, et al. Memory matters: The need to improve long-term memory in llm-agents[C]//Proceedings of the AAAI Symposium Series: Vol. 2. 2023: 277-280.

[59] MAO J, YE J, QIAN Y, et al. A language agent for autonomous driving[A]. 2023.

[60] CORMEN T H, LEISERSON C E, RIVEST R L, et al. Introduction to algorithms [M]. MIT press, 2022.

[61] YAO S, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models[J]. Advances in neural information processing systems, 2023, 36: 11809-11822.

[62] SUN H, ZHUANG Y, KONG L, et al. Adaplanner: Adaptive planning from feedback with language models[A]. 2023. arXiv: 2305.16653.

[63] ZHANG Z, BO X, MA C, et al. A survey on the memory mechanism of large language model based agents[A]. 2024.

[64] SUMERS T, YAO S, NARASIMHAN K, et al. Cognitive architectures for language agents[J]. Transactions on Machine Learning Research, 2023.

[65] MILLER G A. The magical number seven, plus or minus two: Some limits on our capacity for processing information.[J]. Psychological review, 1956, 63(2): 81.

[66] KARPAS E, ABEND O, BELINKOV Y, et al. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning[A]. 2022.

[67] PARISI A, ZHAO Y, FIEDEL N. Talm: Tool augmented language models[A]. 2022.

[68] SCHICK T, DWIVEDI-YU J, DESSÌ R, et al. Toolformer: Language models can teach themselves to use tools[A]. 2023.

[69] SHEN Y, SONG K, TAN X, et al. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface[C]//Advances in Neural Information Processing Systems. 2023.

[70] WÖLFLEIN G, FERBER D, TRUHN D, et al. Llm agents making agent tools[A]. 2025.

[71] GAO D, LI Z, PAN X, et al. Agentscope: A flexible yet robust multi-agent platform [A/OL]. 2024. https://github.com/modelscope/agentscope.

[72] WU Q, BANSAL G, ZHANG J, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation[A]. 2023.

[73] TEAM X. Xagent: An autonomous agent for complex task solving[J]. XAgent blog, 2023.

[74] DUAN Z, WANG J. Exploration of llm multi-agent application implementation based on langgraph+ crewai[A]. 2024.

[75] ALTO V. Building llm powered applications: Create intelligent apps and agents with large language models[M]. Packt Publishing Ltd, 2024.

[76] SHU D, ZHAO H, LIU X, et al. Lawllm: Law large language model for the us legal system[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024: 4882-4889.

[77] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.