随着互联网数据量呈现指数级的增长，据IDC预测，到2028年全球数据量将增长至393.8ZB，相比于2018年增长9.8倍[1]，传统搜索引擎基于关键词匹配与倒排索引的技术架构面临更加严峻的挑战。尤其是在用户需求复杂化、场景多元化的背景下暴露出以下两点缺陷：其一：短文本查询的模糊性（如如"Python多线程死锁排查"可能涉及GIL机制、锁粒度优化或协程改造等多维度的搜索意图）导致词汇不匹配问题；其二，多轮对话的上下文依赖性，使得传统算法难以精准捕捉用户的搜索意图，导致搜索结果的相关性与准召率不足。例如，基于Bing搜索日志的实证研究显示，31.6%的非导航类查询（如“低碳水化合物早餐食谱”）需用户主动修改搜索词（平均2.3次/会话）[2]，显著降低了信息获取效率。

在此背景下，AI大语言模型（LLMs）凭借其强大的语义理解与生成能力，为搜索引擎的智能化转型提供了新范式。以GPT-4、PaLM-2为代表的模型通过在千亿级语料上的预训练，实现了对复杂意图的深度解析与上下文逻辑建模，可在对话中动态调整搜索策略（如基于病史信息逐步缩小医疗建议范围）。

本课题聚焦于构建基于LLMs的对话式搜索引擎，旨在解决传统系统的三大核心问题：语义鸿沟（词汇不匹配）、交互僵化（单轮检索限制）与个性化缺失（忽略用户历史行为），从而推动搜索服务从“被动响应”向“主动理解”跃迁。

为此，本文设计实现了一种新的智能搜索系统，该系统采用了以下几点关键技术方案：混合检索架构将传统的检索技术与新一代的 AI 模型相结合，构建出“关键词检索+语义理解+对话管理”的三层架构。通过集成 Kimi、Qwen 等先进的 AI 服务以及 Bocha、Exa等专业搜索引擎，实现了检索效率与理解深度的平衡。基于检索增强生成（Retrieval-Augmented Generation）技术，系统可以动态检索和整合多源知识，显著提升了答案的准确性和可靠性。多轮对话优化通过深度学习模型的上下文记忆和意图捕捉能力，支持连续意图推理与模糊查询的语义消歧，结合对话状态追踪技术，确保搜索结果的持续相关性。该系统通过分层架构与AI技术的协同，实现了搜索效率、知识整合与交互体验的全面提升。