ICS 35.020
CCS L 70

T/GDEIIA

团　体　标　准

T/GDEIIA xx—XXXX

AI 大模型中间件技术规范

Technical Specifications of Middleware for AI Large Models

（征求意见稿）

XXXX–XX–XX 发布　　XXXX–XX–XX 实施

广东省电子信息行业协会　　发布

目　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由xxx提出。

本文件由广东省电子信息行业协会归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

AI 大模型中间件技术规范

1 范围

本标准是基于大模型的中间件产品的技术规范要求，本标准所涵盖的技术内容，包括以下几个方面：

① 功能要求：提供多种模块化组件，涵盖数据处理、模型训练、工具使用、智能体和提示工程等方面，灵活定义构建垂类模型大模型，开发大模型驱动的 AI 应用。

② 架构设计：中间件采用模块化分层结构设计。

③ 接口标准：中间件与外部系统使用标准 http 接口协议，确保互操作性和集成性。

④ 性能要求：大模型能够同时处理 100 个并发请求，95%的请求在 5 秒内响应。系统可用时间达到 99%。

⑤ 安全性要求:敏感数据在传输和存储时使用 TLS/SSL 加密。系统实施基于角色的访问控制机制。本标准适用于企业构建垂类大模型应用，包括微调训练垂类模型，开发大模型驱动的 AI 应用，集成大模型与企业内、外部信息系统，赋能企业数字化转型。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 22239 《信息安全技术—网络安全等级保护基本要求》

GB-T 28168-2011 《信息技术 中间件 消息中间件技术规范》

GB/T 38676-2020 《信息技术—大数据—存储与处理系统功能测试要求》

GB/T 41867-2022 《信息技术人工智能术语》

GB/T 42382.1-2023 《信息技术 神经网络表示与模型压缩第1部分：卷积神经网络》

TC260-003《生成式人工智能服务安全基本要求》

3. 术语和定义

下列术语和定义适用于本文件。

3.1. 大模型 large model

大模型是指由具有数十亿甚至上百亿参数的神经元网络构建的人工智能模型，能够处理复杂任务，如自然语言处理、图像生成、图文分析等。这些模型通过在大量数据上进行预训练，具备强大的特征提取、内容生成和泛化能力，可以广泛应用于教育、医疗、金融等智能应用场景。

3.2. 自注意力机制 self-attention

自注意力机制是 Transformer 模型的核心组成部分，允许模型在处理每个输入单元时能够关注序列中的不同单元，并且能够学习不同单元之间的关联性，从而捕捉全局依赖关系。

3.3. 多头注意力机制 multi-head attention

多头注意力机制通过并行计算多个独立的注意力机制，将不同的注意力头的输出拼接在一起。这种方法增强了模型的表示能力，使其能够从不同的角度关注输入序列的特征，提高了对复杂任务的处理能力。

3.4. Transformer

Transformer 是一种用于处理序列数据的深度学习架构，特别适用于自然语言处理和文本生成任务。其组成包括：自注意力机制，编码器-解码器结构，位置编码，多头注意力，前馈网络，残差连接和层归一化。

3.5. 预训练模型 pretrain model

预训练模型在大规模无标注数据上进行训练，学习到通用的特征表示。随后，通过在特定任务上的小规模标注数据进行微调，使模型适应具体任务需求。

3.6. 自监督学习 self-supervised learning

自监督学习是一种机器学习方法，利用数据自身的结构生成监督信号，而无需人工标注数据。通过设计合适的任务（如预测掩码词），模型能够在大规模无标注数据上进行训练，学习到有用的特征表示，广泛应用于预训练模型中。

3.7. 人工反馈强化学习 reinforcement learning from human feedback, RLHF

人工反馈强化学习结合人类反馈信息优化模型。通过让模型生成多个候选输出并接受人类的反馈评分，模型可以调整参数，提高生成结果的相关性和质量，常用于对话系统和内容生成，使得生成的内容对齐人类价值观，并遵从法规、符合伦理和社会的约定俗成。

3.8. 监督精调 supervised fine-tuning

监督精调是用标注数据集进一步训练预训练模型，以使其更好地适应特定下游任务。通过引入人工标注的数据，模型能够校正预训练过程中可能产生的偏差，提升在应用场景中的表现。

3.9. 提示工程 prompt engineering

提示工程通过设计和优化输入提示，指导模型生成更符合预期的输出。在使用预训练语言模型能够显著提高交互效果和生成结果的质量，广泛应用于对话系统和文本生成任务。

3.10. 智能体 agent

智能体是基于大型预训练模型的智能应用程序或系统。这些大模型智能体在各种领域和任务中展示出色的性能，特别是在自然语言处理和相关的智能决策任务中。

3.11. Token

Token 是模型处理文本时的基本单位，通常是单词或字符的子单位。模型通过对输入文本进行 Token 化，将其分解为可处理的最小单元，然后进行分析和生成输出。

3.12. 词嵌入 word embedding

词嵌入是一种将词汇映射到高维向量空间的方法，使得具有相似意义的词在向量空间中相近。通过词嵌入，模型能够理解词汇间的语义关系，提高文本处理任务的效果，是自然语言处理的核心技术之一。

3.13. 位置嵌入 positional encoding

位置嵌入在大模型中加入位置信息，使其能够理解序列中各元素的位置关系。位置嵌入弥补了 Transformer 架构中缺乏位置信息的缺陷，提升了处理顺序数据的能力，广泛应用于文本和序列数据的处理。

3.14. 向量数据库 vector database

向量数据库存储和检索高维向量，常用于相似性搜索和大规模机器学习模型的高效查询。向量数据库在推荐系统、图像检索和自然语言处理等领域有广泛应用，提供快速的相似性匹配和高效的数据管理。

3.15. 搜索增强生成 retrieval augmented generation, RAG

搜索增强生成结合搜索和生成技术，利用检索到的信息增强生成模型的输出。在处理复杂问答和知识丰富的生成任务中，融合检索信息，提高回答的准确性，减少模型幻觉。

3.16. 工具 tool

工具是指各种用于辅助模型训练、评估和应用的软件、库和平台。给开发者高效构建、优化和部署大模型，从数据处理、模型训练到结果分析，提供全方位的技术支持。

3.17. 混合专家模型 mixture of experts, MOE

混合专家模型是一种模型架构，它通过组合多个子模型（即“专家”）来提高模型的预测性能和效率。每个子模型专门处理输入空间的一个子集，而一个门控网络决定每个数据应该由哪个模型进行训练，以减少不同样本类型之间的干扰。

3.18. 门控网络 gating network

门控网络用于 MOE 模型中选择哪个专家网络处理输入数据。它的输出结果是一个概率向量，表示每个专家网络被选择的概率。

3.19. 扩散模型 diffusion model

扩散模型是一种在图像处理和生成模型，特别是在深度学习领域。扩散指的是一种图像生成过程，其中图像的生成是稳定且逐步进行的，而不是突然或随机地出现。

3.20. 循环神经网络 recurrent neural network, RNN

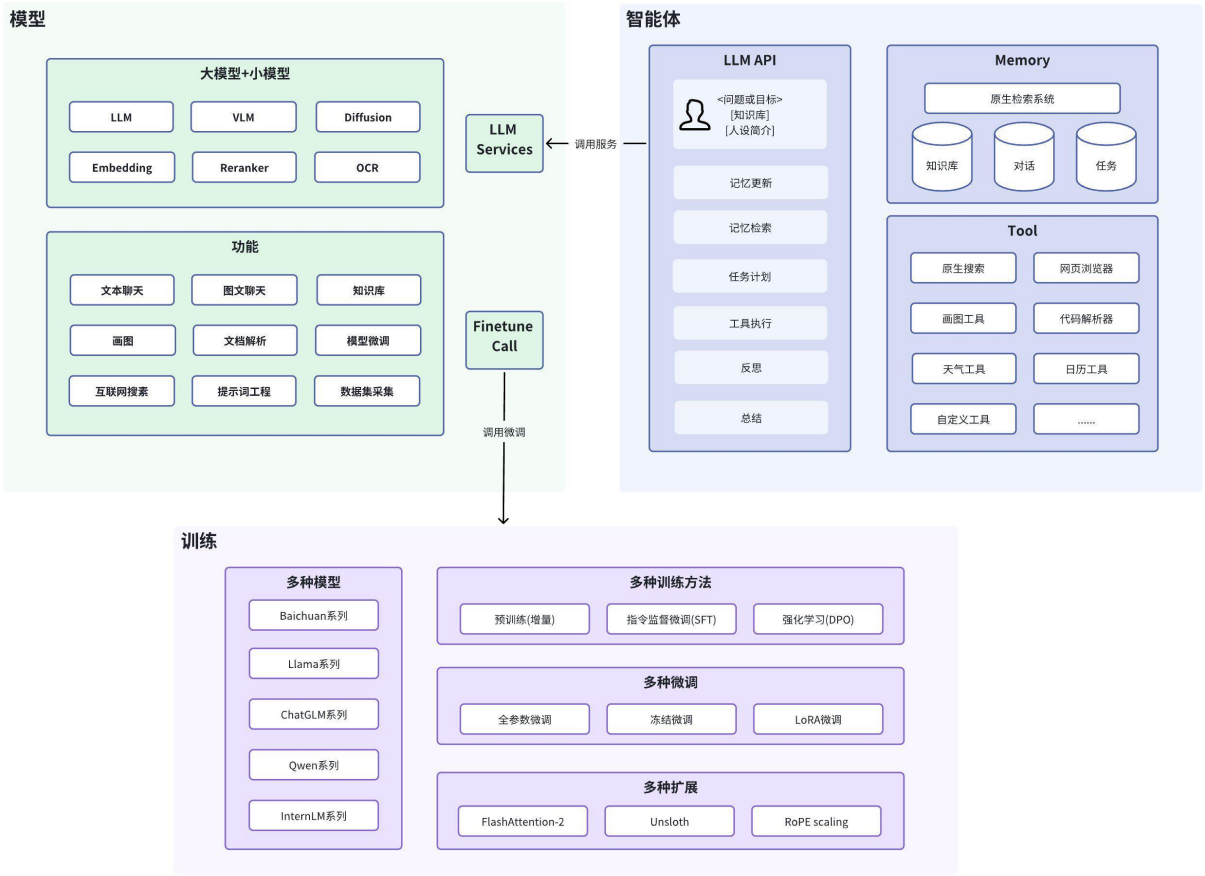
循环神经网络是一种深度学习模型，特别适用于处理序列数据，如时间序列数据、文本或任何具有时间或序列特性的数据。RNN 的特点是网络中存在循环连接，允许信息在网络中持续存在。

3.21. 思维链 chain of thought, CoT

思维链是一种认知心理学和认知科学中的概念，人类在解决问题或进行决策时，思维过程的连贯性和逻辑性。思维链强调的是从初始问题或信息出发，通过一系列的思考步骤，最终达到解决问题或做出决策的过程。

4. 大模型中间件系统架构

4.1. 大模型中间件系统功能架构



这个架构可以分为三个主要模块：模型、智能体 和 训练。

用户通过智能体模块与系统交互，输入数据被传递到模型模块进行处理。 模型模块返回处理结果给智能体模块，智能体模块再将结果反馈给用户。智能体模块通过 LLM Services 调用模型模块的服务。 模型模块提供基础功能和模型服务，同时通过 Finetune Call 接口，调用训练模块对模型模块中的模型进行微调和优化。这三个模块通过紧密协作，共同构建了一个高效、智能、可持续优化的系统架构。模型模块提供核心能力，智能体模块实现实际应用，训练模块确保模型不断进化。

以下是对每个功能模块的详细描述：

① 模型

模型模块是由大模型和小模型负责提供多种语言和视觉处理服务。这些服务是整个系统的基础，处理各

种输入数据并提供相应的输出。主要包含以下组件：

- LLM (Large Language Model)：主要用于处理和生成自然语言文本。它可以执行对话生成、文本补全、翻译等任务。
- VLM (Vision-Language Model)：处理图文混合任务，如图像描述生成、图像问答等。
- Diffusion：使用扩散模型进行图像生成或其他复杂数据生成任务。
- Embedding：将文本或其他数据转换为高维向量表示，以便进行相似度计算、聚类等任务。
- Reranker：对搜索结果或推荐列表进行重新排序，以提高结果的相关性和准确性。
- OCR (Optical Character Recognition)：从图像中提取文字信息，适用于文档扫描、图像中的文字识别等任务。

1） 必要功能模块

- 文本聊天：实现基于自然语言的对话功能。
- 图文聊天：支持包含图像和文本的对话。
- 知识库：管理和查询知识库中的数据。
- 模型微调：根据特定需求对模型进行微调。

2） 拓展功能模块

- 提示词工程：创建和优化提示词以提高模型性能。
- 数据集采集：收集和管理训练和测试数据集。
- 画图：使用模型生成图像。
- 互联网搜索：通过互联网检索信息。
- 文档解析：解析和理解文档内容。

② 智能体

智能体模块提供了与用户交互和管理模型的接口。它处理用户请求，管理模型记忆和任务，执行工具操作，并进行反思和总结。

1） 必要功能模块

思维链(CoT)

- 记忆更新：更新模型的记忆库，以便模型在未来的对话或任务中能够参考这些信息。
- 记忆检索：从记忆库中检索相关信息，帮助模型在对话或任务中提供准确的回答或执行正确的操作。

- 任务计划：创建和管理任务计划，确保任务按时执行并跟踪其进展。
- 工具执行：调用和执行各种工具，如搜索、解析、生成等。
- 反思：模型自我反思，改进和优化自身性能。
- 总结：对对话或任务进行总结，提供简洁的概述和重要信息。

记忆(Memory)

- 知识库：存储和管理知识库中的数据，提供查询和更新功能。
- 对话：记录和管理对话历史，帮助模型记住和参考过去的对话内容。
- 任务：管理和追踪任务信息，包括任务的状态、进展和结果。

2) 拓展功能模块

工具(Tool)

- 原生搜索：执行搜索功能，检索互联网或其他数据源的信息。
- 网页浏览器：提供网页浏览功能，帮助模型从网页中获取信息。
- 画图工具：生成和编辑图像。
- 代码解析器：理解和解析代码，帮助解决编程相关的问题。
- 天气工具：获取和展示天气信息。
- 日历工具：管理和查询日历事件。
- 自定义工具：提供扩展功能，用户可以根据需要添加自定义工具。

③ 训练

训练模块使得模型能够从训练数据中学习到一般化的模式和规律，从而提高对未见数据的泛化能力。这是确保模型在真实场景中表现良好。

1) 必要功能模块

多种模型：提供了多种不同系列的模型，同时可拓展更多模型，满足不同类型的任务需求：

- Llama 系列
- Baichuan 系列
- ChatGLM 系列
- Qwen 系列

- InternLM 系列

多种训练方法:支持多种训练和微调方法，以提升模型性能和适应性：

- 预训练（单集）：基于单一大型数据集进行预训练。
- 指令微调训练（SFT）：通过特定指令集进行微调训练。
- 强化学习（DPO）：使用强化学习方法进行模型优化。

多种微调:提供多种微调方法，使模型能够更好地适应特定任务：

- 全参数微调：对模型的所有参数进行微调。
- 提示词微调：通过提示词对模型进行微调。
- LoRA 微调：使用低秩适应（LoRA）方法进行微调。

2) 拓展功能模块

多种扩展:提供多种扩展技术，提高模型的性能和效率：

- FlashAttention-2：优化的注意力机制，提升计算效率。
- Unsloth：一种高效训练优化技术，高效快速地微调模型，同时内存消耗大大减少。
- RoPE scaling：位置编码扩展技术，提升模型处理长序列的能力。

4.2. 大模型中间件系统组件架构

该架构图展示了一个多组件的智能系统，分为数据检索、数据处理、智能体、模型和监控五大模块。

数据检索模块从向量数据库、内存和索引中检索数据，并通过查询器提供给数据处理模块。数据处理模块利用文档转换器和加载器处理从数据检索模块获取的数据，并将处理后的数据存储在数据库、知识库

或数据集中。智能体模块通过数据连接器与数据处理模块进行数据交互，供智能体模块使用。智能体模块通过模型 I/O 接口调用模型模块中的各种模型（如基础模型、聊天模型、多模态模型等）来完成特定任务。监控模块对各模块进行性能评测和运行监控。数据检索和数据处理模块提供数据基础，智能体模块执行具体任务，模型模块提供核心计算能力，监控模块确保系统性能和稳定性。

以下是对每个组件模块的详细描述：

① 数据检索

- 向量数据库：存储和检索高维向量数据，用于快速查询相似度。
- 内存：存储临时数据，提供快速读写访问。
- 索引：创建数据的索引结构，提高查询效率。
- 查询器：执行和优化数据查询操作。

② 数据处理

- 数据库：存储结构化数据，支持数据的高效存取。
- 知识库：存储和管理知识数据，支持知识查询和推理。
- 数据文件：管理和存储数据文件，支持数据的上传和下载。
- 数据集：组织和管理数据集，用于模型训练和评估。
- 文档转换器：将文档转换为结构化数据格式。
- 文档加载器：加载和解析文档数据。

③ 智能体

- 智能体：自主执行特定任务的人工智能实体。
- 多智能体协作：多个智能体协同工作，解决复杂任务。
- 链：连接多个任务或操作，形成工作流。
- 提示：为模型提供指令或背景信息，提高模型响应质量。
- 工具：辅助智能体完成特定任务的功能模块。
- 数据连接器：建立和管理与数据源的连接。
- 模型 I/O：处理模型输入和输出的数据格式和传输。

- 分词器：将文本分解成独立的词或短语。
- 回调器：在特定事件或条件下执行回调操作。
- 模式：定义智能体或系统的工作模式。

④ **模型**

- 基础模型：提供基本的模型架构和功能。
- 聊天模型：专门用于对话生成的模型。
- 多模态模型：同时处理多种数据类型（如文本、图像）的模型。
- 扩展模型：提供扩展功能的模型模块。
- 嵌入模型：将输入数据转换为向量表示的模型。
- 语音模型：处理和生成语音数据的模型。

⑤ **监控**

- 性能测评：评估系统和模型的性能。
- 运行监控：实时监控系统和模型的运行状态。

5. 大模型中间件模块组件规范

5.1. 向量数据库

① **功能描述**

向量数据库是专门设计用于存储和管理向量数据的数据库组件。它支持高效地存储大规模的向量数据集，并提供强大的检索和分析能力。

② **功能要求**

- 支持高效存储和管理各种维度的向量数据，包括稠密向量和稀疏向量。
- 支持处理大规模数据集，高并发的数据写入和读取操作。
- 支持提供多种索引结构选项，如基于树结构的索引、哈希表索引等，以支持快速的相似度搜索，范围查询和处理复杂的向量检索需求，包括精确匹配、K 近邻搜索等。

- 支持提供直观和强大的查询语言或 API，支持用户进行灵活的向量数据查询和分析。

5.2. 内存

① 功能描述

用于存储和管理对话历史。其主要功能是允许模型在多轮对话中保持上下文，从而能够理解和回应基于先前对话内容的查询。内存组件通过存储和检索对话中的信息，帮助模型实现更连贯和相关的交互。

② 功能要求

- 支持有效存储和更新对话中的上下文信息，包括用户提出的问题、模型的响应和相关环境变化。
- 支持提供快速的信息存储和检索机制，支持按照时间顺序或特定上下文索引进行数据访问。
- 支持实时更新对话历史和上下文信息，以反映最新的会话状态和用户意图变化。
- 支持提供自动化的历史记录清理和管理功能，以控制存储容量和保持系统性能。

5.3. 索引

① 功能描述

索引是一个专门为处理语言数据链而设计的索引系统。它利用先进的自然语言处理技术，如词嵌入、语义分析和机器学习算法，来创建和管理一个高效的语言数据索引。这个索引能够快速检索和分析大量的文本数据，支持多种语言，并能够理解上下文和语义关系，从而提供精确和相关的搜索结果。

② 功能要求

- 支持对文本进行有效的预处理和清洗，包括分词、词形还原、去除停用词等，以提升索引的质量和效率。
- 支持使用词嵌入技术将文本转换为向量表示，以捕捉词语之间的语义相似性。
- 支持处理多种语言的文本数据，并能够在不同语言之间进行翻译和语义映射。
- 支持提供快速的文本检索和分析能力，支持复杂的查询操作，并保证检索结果的准确性和相关性。
- 支持动态更新索引内容，并能够有效扩展以处理不断增长的文本数据量。

5.4. 查询器

① 功能描述

查询器是用于执行数据源查询操作的组件，能够根据给定的查询语句从数据源中检索数据。

② 功能要求

- 支持灵活和强大的查询语言，能够处理各种类型的查询需求，包括条件查询、聚合查询、排序和相识度匹配等。
- 支持连接和操作多种数据源，包括向量数据库、内存、索引、传统型数据库等，以及文件系统和 API 接口。
- 支持分布式查询和并行处理，以应对高并发查询的需求，确保系统的扩展性和性能。

5.5. 文档转换器

① 功能描述

文档转换器能够将文档从一种格式转换为另一种格式，例如将 PDF 文档转换为文本文件或 MarkDown 格式。

② 功能要求

- 支持多种文档格式之间的转换，包括但不限于：PDF、Microsoft Office 文档（如 Word、Excel、PowerPoint）、文本文件、MarkDown、HTML 等常见格式。
- 支持解析和理解各种格式的文档内容，包括文字、图片、表格等元素。
- 支持批量处理多个文档，能够高效地进行大规模文档转换操作。

5.6. 文档加载器

① 功能描述：

文档加载器负责从外部加载文档到系统中，例如从文件系统、网络或数据库中加载文档。

② 功能要求：

- 支持从多种数据源加载文档，包括本地文件系统、远程网络资源（如 URL）以及其他存储系统。
- 支持从源数据中准确地获取文档内容，并进行必要的数据抽取和转换，以便系统进一步处理和分析。
- 支持批量加载多个文档，并能够进行并行处理，以提高数据加载和处理的效率。

5.7. 智能体

① 功能描述

智能体是大模型中的核心执行单元，具备自主执行任务的能力。它能够理解并解析接收到的任务指令，结合当前环境的状态信息，进行智能决策和行动规划。智能体的设计旨在提高任务执行的效率和准确性，同时能够适应多变的环境条件，实现灵活的任务处理和问题解决。

② 功能要求

- 支持能够准确理解和解析接收到的任务指令，包括语义理解、意图识别和任务分类等。
- 支持具备环境感知能力，能够实时监测和获取当前环境的状态信息，如传感器数据、外部条件等。
- 支持基于任务指令和环境状态信息，能够进行智能决策和行动规划，制定有效的任务执行策略。
- 支持能够自主执行任务和行动。
- 支持并行处理多个任务，能够灵活地调度和管理任务执行顺序，以提高整体执行效率。
- 支持实现健壮的错误处理和异常处理机制，能够识别和应对执行过程中可能出现的问题和异常情况并反馈。

5.8. 链

① 功能描述

链是将多个中间件模块连接起来，以实现复杂功能的组件。它可以串联各种模块，形成数据处理或任务执行的流程。

② 功能要求

- 支持能够定义和控制数据处理或任务执行的流程，确保各个中间件模块按照预定顺序执行。

- 支持在模块之间传递数据，并能够进行数据格式转换、数据映射或数据整合，以确保数据流的连贯性和有效性。

5.9. 提示

① 功能描述

在为大型语言模型提供一个结构化的输入框架，以生成高质量、相关性强的文本输出。通过精心设计的提示词，模型能够更好地理解用户的意图和需求，从而产生更加精确和有用的回答或内容。

② 功能要求

- 支持多种类型的提示，包括关键词、短语、问题模板等，以满足不同场景和应用需求。
- 支持允许用户自定义提示词或输入文本的内容和格式，支持定制化设置，以适应特定领域或任务的需求。

5.10. 工具

① 功能描述

工具是一款集成了人工智能技术的多功能辅助工具库。工具库的每个工具实现不同的功能。

② 功能要求

- 支持通用工具,如互联网搜索、网页浏览器、画图工具、代码解析器、天气工具、日历工具等。
- 支持自定义工具,实现工具拓展。

5.11. 数据连接器

① 功能描述

数据连接器是大模型的中间件组件，用于建立和管理大模型与数据源之间的连接。

② 功能要求

- 支持处理连接中的故障或中断，支持恢复和重试机制，确保数据传输的可靠性和稳定性。

- 支持能够扩展以支持新的数据源和数据处理需求，同时灵活适应不同的数据集成场景。

5.12. 模型 I/O

① 功能描述

模型 I/O 组件负责管理模型与外部系统之间的数据传输，包括输入数据的接收和输出结果的传递。

② 功能要求

- 支持多种数据输入输出格式，例如 JSON、文本、图像、视频、音频等。
- 支持高性能数据传输，提高传输效率。
- 支持保护数据在传输过程中加密，保证数据安全。

5.13. 分词器

① 功能描述

分词器是一个文本处理工具，能够将文本按照一定规则分解成词语或词条，用于文本处理、分析和理解。

② 功能要求

- 支持能够处理多种语言的文本，包括但不限于英文、中文等。
- 支持不同的字符集和编码，如 UTF-8、GBK 等。
- 支持能够处理复杂的语言结构，如长句、复合句、主谓宾结构等，保持分词结果的准确性和连贯性。

5.14. 回调器

① 功能描述

回调器是用于处理异步事件的组件，能够在事件发生后执行预定义的回调函数，实现事件驱动的编程模式。

② 功能要求

- 支持能够向回调函数传递必要的参数，包括事件相关的数据、状态信息等。
- 支持灵活的参数配置和传递方式，以满足不同回调函数的参数需求。

5.15. 模式

① 功能描述

模式是指在应用开发中用于智能体相关任务的具备固定设计的流程库。

② 功能要求

- 支持多种任务模式，如文本处理、语言模型、语义分析、交互式对话等。
- 支持具备清晰模式解耦，模式之间不直接依赖，模式更容易组合使用。

5.16. 聊天模型

① 功能描述

聊天模型是一个自然语言处理的大模型，能够根据输入的对话历史和上下文信息生成自然流畅的回复文本。

② 功能要求

- 支持主流的开源和商用大模型，如文心一言、gpt4 、llama、qwen 等。
- 支持能够理解用户问题的语义，意图结合多轮对话中的上下文等信息，生成自然、流畅且语法正确的文本回复，并考虑到语气、情感和表达方式的多样性。
- 支持不仅限于问答场景，还能处理推理、解释、建议等多种语言交流需求，扩展其在实际应用中的应用场景。
- 支持能够快速响应用户的输入，保持高效率的交互速度，同时具备处理大量并发请求的能力。