AI+智能体会议系统

数据来源文档

团队：李小慧 郭皓怡 罗飞 许夏轩 钟文振

时间：2025/3/28

目录

1. 引言

1.1 编写目的

本文档旨在清晰界定AI+智能体会议系统的数据来源构成，包括核心模型、第三方依赖、训练数据及工具库的获取途径、技术特性与使用规范，确保项目开发的透明度、可追溯性和合规性。

1.2 定义与专业术语

大模型（LLM）：基于深度学习的超大规模语言模型（如GPT系列）。

智能体（Agent）：可自主执行任务的AI模块，通过链式流程组合多组件功能。

RAG（检索增强生成）：结合检索与生成技术提升语言模型输出的准确性。

嵌入模型（Embedding Model）：将文本映射为语义向量的算法（如BGE-M3）。

知识蒸馏（Knowledge Distillation）：将大模型能力压缩至轻量级模型的技术。

2. 数据来源

2.1 模型及训练数据来源

主模型与开源组件

语言生成模型

- **模型/工具：**distilgpt2
- **来源：**https://huggingface.co/distilgpt2
- **获取方式：**

直接从 Hugging Face 下载：

```
git lfs install
git clone https://huggingface.co/distilgpt2
```

使用 transformers 库自动下载（首次加载时）：

```
from transformers import AutoModelForCausalLM, AutoTokenizer
model = AutoModelForCausalLM.from_pretrained("distilgpt2")
tokenizer = AutoTokenizer.from_pretrained("distilgpt2")
```

- **核心用途**：用于会议总结的后续行动建议生成（ai_generate_options 函数）。
- **关键特性**：轻量化（82M 参数，2GB 显存）、推理速度比 GPT - 2 快 60%

文本嵌入模型

- **模型/工具**：BGE - M3
- **来源与获取方式**：北京智源研究院，https://cloud.siliconflow.cn/models/Pro/BAAI/bge - m3
- **核心用途**：跨语言语义检索与重排序
- **关键特性**：支持 100 + 语言，兼容稠密/稀疏检索，输入长度 8k token

对话模型

- **模型/工具**：DeepSeek - V3 - 0324
- **来源与获取方式**： https://www.deepseek.com
- **核心用途**：代码生成与自然语言交互
- **关键特性**：MoE 架构优化（6850 亿参数），推理速度提升，强编程/数学能力

安全智能体框架

- **模型/工具**：恒脑安全智能体
- **来源与获取方式**：安恒信息自研，https://gc.das - ai.com/agentCenter
- **核心用途**：任务自动化执行与安全闭环管理
- **关键特性**：基于安全垂域大模型，支持智能化威胁检测与响应

第三方依赖模型

SadTalker 核心模型

- **用途**：语音驱动人脸动画合成
- **来源**：

官方渠道：

Google Drive

GitHub Releases

百度云盘（密码：sadt）

脚本下载：

bash scripts/download_models.sh

- **关键技术**：3D 运动预测、人脸渲染

GFPGAN

- **用途**：人脸增强
- **来源**：https://github.com/TencentARC/GFPGAN
- **关键技术**：生成对抗网络修复模糊人脸

Wav2Lip

- **用途**：唇形同步
- **来源**：https://github.com/Rudrabha/Wav2Lip
- **关键技术**：音视频对齐的时序建模

Real - ESRGAN

- **用途**：图像/视频超分辨率
- **来源**：https://github.com/xinntao/Real - ESRGAN

- **关键技术：**非局部注意力增强视觉细节

训练数据

主训练数据

由公开数据集（如 LRS3、VoxCeleb2）构成，覆盖多语言、多场景对话数据。

增强数据

通过 Tencent AI Lab 和 Ant Group 内部资源补充高精度场景化语料。

本地化优化

针对中文会议场景，使用 Jieba 分词器处理自定义词库（如技术术语缩写）。

2.2 工具库来源

LangChain

用途

构建智能体任务链

安装方式

pip install langchain

依赖数据模型

需配置本地或云端 LLM 服务端点

Coqui TTS

用途

文本转语音

安装方式

pip install TTS + 在线下载指定语音模型

依赖数据模型

tts_models/multilingual/multi - dataset/your_tts

facexlib

用途

人脸检测与特征分析

安装方式

pip install facexlib

依赖数据模型

预训练人脸关键点检测模型（自动下载）3. 安全与合规说明

数据匿名化：训练数据经脱敏处理，移除个人身份信息（PII）。

授权验证：第三方模型使用遵循Apache 2.0/MIT等开源协议，商用场景需单独申请授权。

依赖性管理：通过requirements.txt与Shell脚本自动化下载路径校验。

2.3 素材来源

HyQE(假设问题嵌入)：

HyQE（Hypothetical Query Embeddings，假设问题嵌入）是一种创新的上下文排序方法，旨在改进检索增强生成（RAG）系统。这种方法由Weichao Zhou等人开发，并在2024年EMNLP会议上发表。它的核心思想是利用大型语言模型(LLMs)从上下文中生成假设性问题，为这些问题建立索引，然后通过比较用户查询与这些假设性问题来对上下文进行排序。

- 来源：https://arxiv.org/abs/2410.15262
- 特性：反向推理机制、查询效率优化、多样化检索
- 核心用途：基于HyQE的高级检索系统