附件1

中国藏学文献资源数据中心专项课题指南

一、课题概述

本课题由中国藏学研究中心发布，旨在汇聚各方科研力量，聚焦藏学信息处理关键环节，形成可直接服务于藏学研究、文化保护、社会应用的技术成果，为藏学研究信息化进程提供核心支撑，推动藏学研究在新时代背景下的创新性发展。

具体研究目标包括：​

1.开发高精度的藏文印刷体OCR系统，识别准确率达到国内领先水平。

2.构建覆盖卫藏、安多、康巴三种方言的藏语语音识别和语音合成系统，满足针对不同方言区的应用需求。

3.研发高效的藏文机器翻译和计算机辅助翻译工具，提升藏汉互译的准确性和效率。

4.研究探索以思维链为主的推理模式对大模型藏文能力影响的原理和应用。

5.研究和完善藏文多编码转码实现，确保藏文信息在不同系统和平台之间的无障碍交换。

二、资助单位

中国藏学研究中心

三、课题周期

本次课题为1年期课题，从委托业务合同签订之日起计算。

四、各方向课题要求

**（一）藏文印刷体OCR技术研究​**

**1.研究内容**

针对藏文印刷体印刷字符的光学识别技术进行深入研究，制作高质量训练集、验证集，形成模型和应用工具（系统）。

**2.经费规模**

50万元。

**3.技术指标**

（1）识别准确率

印刷体藏文识别准确率应达到97%及以上。

对于褪色、污损等低质量图像，识别准确率应不低于90%。

（2）识别速度

在分辨率300dpi条件下，单张A4页面识别时间不超过5秒。

（3）多语言支持

支持藏汉英多语言识别。

藏汉混排文本识别准确率需达到90%及以上。

藏汉英混排文本识别准确率需达到85%及以上。

**4.交付成果形式**

（1）藏文OCR识别模型与数据集

提供训练好的藏文OCR识别模型和权重文件​。

提供高质量藏文OCR数据集，包含典型场景图片和对应文本标注。

（2）软件或工具集成

支持本地离线部署和使用，所使用的技术和软件自主可控。

支持主流图片和文档格式，支持队列任务和批量处理，如JPG、BMP、TIF、PNG、PDF等。

可实现在线接入使用，支持二次开发和集成。

（3）文档与报告

提供系统设计文档、模型训练和使用文档、部署应用和开发集成文档，提供简明使用教程。

提交一份详细的研究报告，全面阐述课题研究的背景、目标、方法、过程、结果以及结论等内容。

**（二）藏语（卫藏、安多、康巴方言书面语与口语）语音识别（ASR）技术研究​**

**1.研究内容**

针对藏语语音，包括卫藏、安多、康巴方言书面语与口语的语音识别技术进行深入研究，制作高质量训练集、验证集，形成模型和应用工具（系统）。

**2.经费规模**

50万元。

**3.技术指标**

（1）识别准确率

理想场景下，清晰、无噪音、标准的单人语音识别字准确率不低于95%。

低噪声环境下的语音识别字准确率不低于90%。

（2）识别速度

流式实时识别的延迟不超过500毫秒。

非流式离线识别速度不小于5倍实时速度，即每秒识别不少于5秒语料。

**4.交付成果形式**

（1）藏语ASR识别模型与数据集

提供训练好的藏语ASR识别模型和权重文件​。

提供高质量藏语ASR数据集，包含典型场景语音和对应文本标注。

（2）软件或工具集成

支持本地离线部署和使用，所使用的技术和软件自主可控。

支持流式和非流式输入，非流式输入支持队列任务和批量处理，支持主流音频格式，如WAV、MP3、PCM等。

可实现在线接入使用，支持二次开发和集成。

（3）文档与报告

提供系统设计文档、模型训练和使用文档、部署应用和开发集成文档，提供简明使用教程。

提交一份详细的研究报告，全面阐述课题研究的背景、目标、方法、过程、结果以及结论等内容。

**（三）藏语（卫藏、安多、康巴方言）语音合成（TTS）技术研究​**

**1.研究内容**

针对藏语语音，包括卫藏、安多、康巴方言的语音合成技术进行深入研究，制作高质量训练集、验证集，形成模型和应用工具（系统）。

**2.经费规模**

50万元。

**3.技术指标**

（1）合成质量

语音合成的音色接近真实人声，具备自然流畅清晰的语音效果。

可支持音色选择，至少提供男、女2种音色。

语音合成准确，识别校验的词错误率不超过8%。

（2）合成速度

单次语音合成的延迟不超过500毫秒。

推理速度RTF（推理耗时/合成音频耗时）不大于5。

**4.交付成果形式**

（1）藏语TTS合成模型与数据集

提供训练好的藏语TTS合成模型和权重文件。

提供高质量藏语TTS数据集，包含典型场景文本和对应音频数据。

（2）软件或工具集成

支持本地离线部署和使用，所使用的技术和软件自主可控。

支持流式和非流式输出，非流式输出支持队列任务和批量处理，支持主流音频格式，如WAV、MP3、PCM等。

可实现在线接入使用，支持二次开发和集成。

（3）文档与报告

提供系统设计文档、模型训练和使用文档、部署应用和开发集成文档，提供简明使用教程。

提交一份详细的研究报告，全面阐述课题研究的背景、目标、方法、过程、结果以及结论等内容。

**（四）藏文机器翻译（MT）技术研究​**

**1.研究内容**

针对藏汉翻译技术进行深入研究，制作高质量训练集、验证集，形成模型和应用工具（系统）。

**2.经费规模**

50万元。

**3.技术指标**

（1）翻译质量

常见领域语料，如新闻、公文报告、日常对话等，藏汉翻译采用翻译多维质量标准（Multidimensional Quality Metrics，MQM）评分，得分不低于80分。

人工评价翻译结果，需确保准确、完整、无歧义。

（2）翻译速度

在满足模型运行基础硬件条件的平台部署，通常要求首字输出延迟不超过1秒，每分钟翻译字数不少于500字。

如采用大参数模型，首字输出延迟和翻译速度可根据模型参数量和推理方式对应增加。

**4.交付成果形式**

（1）藏语MT翻译模型与数据集

提供训练好的藏语MT翻译模型和权重文件。

提供高质量藏汉平行语料数据集，包含典型领域和应用场景。

（2）软件或工具集成

支持本地离线部署和使用，所使用的技术和软件自主可控。

支持流式和非流式输出，非流式输出支持队列任务和批量处理。

可实现在线接入使用，支持二次开发和集成。

提供WPS、Microsoft Word插件，可实现文档内直接翻译。

（3）文档与报告

提供系统设计文档、模型训练和使用文档、部署应用和开发集成文档，提供简明使用教程。

提交一份详细的研究报告，全面阐述课题研究的背景、目标、方法、过程、结果以及结论等内容。

**（五）藏文计算机辅助翻译（CAT）技术研究​**

**1.研究内容**

结合大语言模型和翻译技术进步，针对藏汉翻译有关计算机辅助翻译技术进行深入研究，提供适配藏文、适应新的内容生产方式的计算机辅助系统和相关工具。

**2.经费规模**

50万元。

**3.技术指标**

（1）翻译记忆库性能

支持TMX等国际标准，便于维护管理。

精确匹配时翻译记忆匹配准确率不小于99%。

模糊匹配、基于编辑距离相似度≥85%时，翻译记忆匹配召回率不小于85%。

（2）术语库性能

可准确识别文本中已规定术语并提供术语译文，识别准确率不小于99%。

可实现术语提取，从文本中发现术语并入库。

术语查询响应时间不超过1秒（术语库规模不小于2万条时）。

术语一致性检查准确率不小于95%。

（3）质量保证性能

格式错误、拼写错误、标记错误等常规错误检出率不低于90%。

可自定义错误类型，并基于模型实现自动检查。

**4.交付成果形式**

（1）软件或工具集成

提供功能完整的藏文CAT软件或工具系统。

提供翻译记忆库和术语库管理工具。

支持文档提取功能，可接入OCR工具或模型API，进行藏文、中文OCR识别。

支持预翻译功能，可接入翻译模型或工具API，进行藏汉文本预翻译。

（3）文档与报告

提供系统设计文档和使用文档、部署应用和开发集成文档，提供简明使用教程。

提交一份详细的研究报告，全面阐述课题研究的背景、目标、方法、过程、结果以及结论等内容。

**（六）****以****思维链为主的推理模式对大模型藏文能力影响研究**

**1.研究内容**

藏文作为一门低资源语言，研究探索同等数据基础上，以思维链为主的推理模式对大模型藏文能力影响的原理和应用，探索在同等规模语料数据基础上持续提升模型藏文能力和泛化能力。

**2.经费规模**

50万元。

**3.技术指标**

研究探索以思维链为主的推理模式对大语言模型藏文能力影响，在模型性能指标、训练效率、泛化能力方面的量化表现。

**4.交付成果形式**

（1）模型与数据集

提供可验证研究成果的具备藏文能力大语言模型的模型和权重文件。

提供思维链数据集及构建方法和代码，需包括高质量藏文思维链数据。

提供模型效果验证评估数据集。

模型可本地离线部署和应用验证。

（2）文档与报告

提供模型部署应用文档，模型效果评估报告。

提交8000字以上研究报告1篇，全面阐述课题研究的背景、目标、方法、过程、结果以及结论等内容。

提交2500-3000字行业建议报告1篇。

**（七）藏文多编码转换实现研究**

**1.研究内容**

研究和完善藏文多编码转码实现，确保藏文信息在不同系统和平台之间的无障碍交换，实现对多种编码形式藏文数据资源的高效利用。

**2.经费规模**

15万元。

**3.技术指标**

（1）编码兼容性

完全兼容国标和Unicode（0F00-0FFF范围）藏文编码标准。

可实现不少于6种主流编码类型的识别和转换。

（2）跨平台兼容性

转换后内容在Windows、macOS、Linux等主流操作系统上，及WPS、Microsoft Word等主流办公软件中显示一致性不小于95%。

（3）藏文拉丁字母转写适配

可实现自定转写方案的藏文拉丁字母转写。

**4.交付成果形式**

（1）软件或工具集成

提供功能完整的藏文多编码转换软件或工具系统。

支持主流文档格式文件浏览和文本提取，如DOC、DOCX、RTF、网页、XML等。

支持对粘贴的UNICODE编码文本或拉丁字母转写进行处理。

支持转写结果导出为主流文档格式文件，如DOC、DOCX、PDF等。

支持本地离线部署和使用，所使用的技术和软件自主可控。

可实现在线接入使用，支持二次开发和集成。

（2）文档与报告

提供系统设计文档和使用文档、部署应用和开发集成文档。

提交一份详细的研究报告，全面阐述课题研究的背景、目标、方法、过程、结果以及结论等内容。

五、交付成果权利归属

资助单位享有成果的非独占性使用权，有权将课题成果用于藏学研究、文化保护、公益宣传等非商业用途。如需将成果用于其他商业用途，需与课题承担单位协商并签订相关协议。

课题承担单位有权在本单位内部活动中使用课题研究成果，在成果推广应用过程中，应优先保障资助单位的权益，并及时向资助单位通报成果使用情况。

相关成果发布、应用过程中，需于显著位置注明课题资助信息。