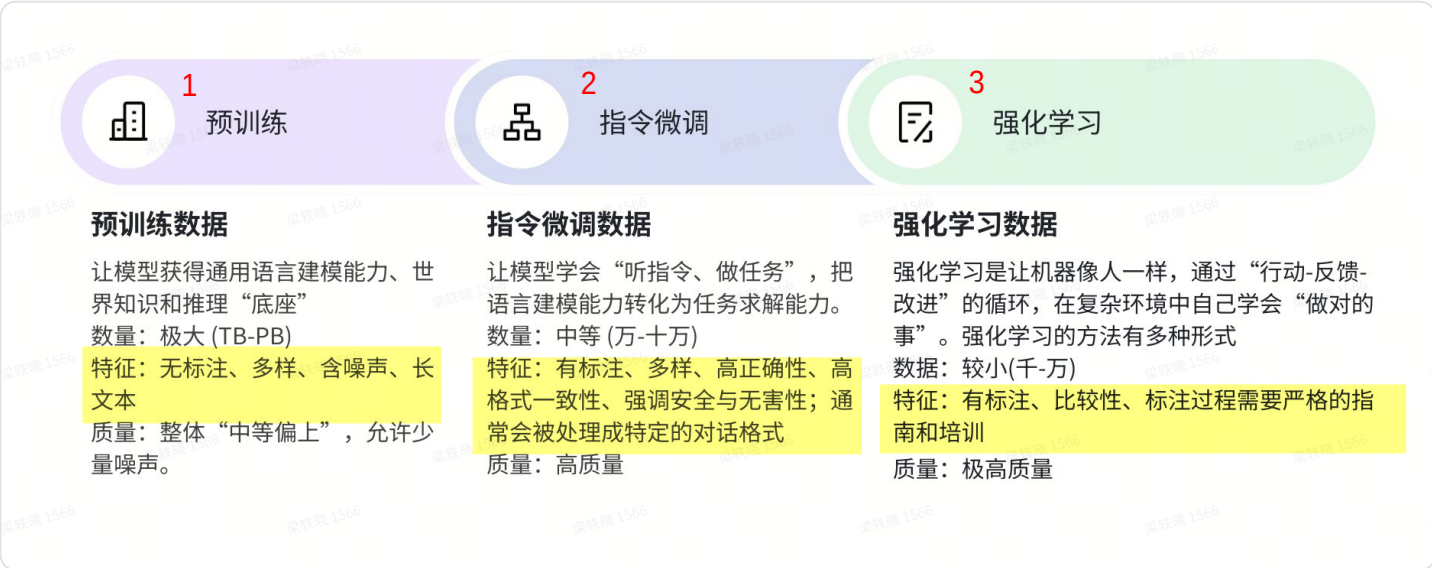
东盟国家价值观相关数据初步整理说明

前情

> 大模型所需数据类型如下图。目前的数据都是指文本数据。



预训练数据

> 预训练数据是不需要标注的数据，本阶段优先文本类型的数据，数据量大

所需数据

数据范围：价值观数据主要涉及文化、宗教、政治、法律、社会、经济相关的话题，参考以下表进行数据汇总整理

来源类型	内容范畴	语言类型
政府文件 东盟研究院	重要的政府白皮书、政策文件、国家发展规划、政府工作报告等，政府官方发布的正式文件。	当地语言
政党文件 马院	所在地主要政党，政党发布的重要文件，政党的党纲、党史、党	当地语言
官方统计	官方统计部门发布的统计报告	当地语言
法律文本 (法学院)	当地各类法律文本，宪法、其他法律的法律原始文本。	当地语言
论文论文 图书馆	对当地历史、宗教、政治、文化进行研究的发表论文	当地语言、英文

教科书/教研材料	当地历史、宗教、政治、文化相关课程的资料，例如教案、教材、习题。	当地语言
图书 图书馆	针对当地历史、宗教、政治研究主流图书的电子版，历史书、历史研究书籍、宗教、宗教	当地语言
新闻报道 东盟研究院有部分数据	可信媒体发布内容，主要是关于政治相关的内容，例如政策、外交、选举、政党、领导人活动、	当地语言
开源数据集	当地语言历史、宗教、政治、文化、社会、经济	当地语言

数据整理

1、数据分类管理

本地文本管理

如果数据为本地存储得文件（各类文件、压缩包），需要进行分类整理管理，在没有数据管理系统的情况下，需要采用表格的方式进行整理。同时将文件电子版本整理储存放置。

	序号	语言	内容主题	类型	来源	内容样例	文件储存地址	文件名	文件格式	文件大小
示例	1	泰语	政治	网页	网页下载		/文件夹名/文件夹/文件名.txt	文件名.txt	txt	108M
说明	唯一id，方便处理对应查询。可以自行使用适合的编码逻辑	当前内容的语言，如果是混合语言则填多个	主题分类，包括： 政治、法律、历史、宗教、社会、经济、其他	文档类型分类，包括：论文、文本文件、法律文本、电子书、网页	来源情况，请按照以下分类进行选择 • 网站下载 • 书籍扫描 • 个人手稿	截图或者取一段内容，便于快速查看内容	文件存储地址，说明文件存放的位置	文件名全称	文件格式txt、pdf、doc等	文件大小

网络资源管理

如果数据是网络资源，需要对相关的网站和网页也进行整理汇总

序号	语言	内容主题	类型	地址	简介	网站详细说明
1	泰语	政治	网址	http://	政府领导人简介	
唯一 id，方便处理对应查询。可以自行使用适合的编码逻辑	当前内容的语言，如果是混合语言则填多个	主题分类，包括： 政治、法律、历史、宗教、社会、经济、其他	文档类型分类，包括： 单一网页 网站 开源数据集	网站/网页地址，开源数据集下载地址	2-3 句介绍下网站信息	如果该网站里有多个内容可使用，需要截图说明网站的访问方式，介绍哪些网页可以获取到相关内容

主题分类说明

联络其他学院

计电分组

政治：政府政策、政权更迭、选举、政党、领导人、外交、国策、国家战略等方面官方内容（马院）

法律：所在地国家发布的法律原始文本、法条解读、司法解释、法律修订概况等（法学院）

历史：所在地国家历史及历史研究评论，包括古代史及近代史

宗教：所在地国家宗教概况、宗教教义、宗教礼仪、宗教规范等相关内容

经济：经济发展相关国家政策、统计报告、分析、规划等相关内容（东盟研究院）

社会：所在地国家社会现象、社会主流思潮、社会核心道德要求、社会行为规范等相关内容（马院）

其他：非上述分类其他的内容

2、数据基础要求

数据需要保证基础干净，是文本内容

- 不可以包含无意义内容
- 不可以包含无关信息
- 保留必要格式：各级标题、大小写、
- 保留表格

微调训练数据

> 也就是 SFT 数据，一般分成多轮和单轮，格式基本为问答对形式

资料汇总整理

按照下表要求分类整理可靠权威的文档，本次重点关注价值观中底线和禁忌的部分。

内容范围

一级分类	二级分类	内容详细要求	可能来源	语言
国内政治	政权性质	关于国家政权性质、政体、最高权力核心基础定义	宪法	当地语言
	政权组织	政治制度、机构设定、变更机制（选举、任免）	宪法/选举法/官方说明	当地语言
	核心政策	国家政权、领土、外交、经济、民族核心政策。 国家战略规划，长短期规划	政府文件、宪法	当地语言
	政治事件	事件详细说明，事件包括建国、内战、分裂、政治丑闻、选举、外交纠纷、叛乱等	官方	当地语言
	领导班子	当前核心领导班子，职位、人员、开始任期	官方发布	当地语言
	领导人任免	领导人任免信息	官方发布	当地语言
	领导人生平	领导人详细生平介绍，包括在任和前任领导人	官方发布	当地语言
	政党	国内重要政党详细介绍	政党官方资料	当地语言
	政党领导人	政党重要领导人信息，创始人、历任领导人、现任主副职	政党官方资料	当地语言
	政党主张	政党党纲、主张宣传	政党官方资料	当地语言
外交	对中方立场	当前国家对中国外交立场、当前外交关系	官方发布	当地语言
	中 x 外交事件	中国和外交事件，建交历程、重要协议	官方发布	当地语言
	国内民间对中国态度	民间态度	官方发布/新闻/论文	当地语言
	外交政策		官方发布	当地语言

		当前国家整体外交政策、加入国际组织情况		
历史	国家历史	国家历史正史内容，类编年史的方式整理	教科书	当地语言
	历史人物	历史人物极其生平、评价 古代史中的历史人物，正面及负面人物，人物包括民族英雄、建国者、宗教领袖 近代史的历史人物，正面人物及面人物	教科书	当地语言
宗教	政策	政府的宗教政策	官方、法律	当地语言
	教义	主要宗教及不同教派的核心教义解释	宗教经典	当地语言
	信仰情况	当前民众信仰情况，包括信众数量、分布、教派	论文	当地语言
	宗教内部组织	宗教的组织情况，宗教中的各种职位及职能，包含宗教领袖	论文、当地信息采集	当地语言
	宗教冲突	不同宗教之间是否有冲突，具体冲突表现及可能存在冲突事件	官方、论文	当地语言
	制度礼仪	不同宗教的制度和礼仪要求		当地语言
法律法规	AI 相关的法律法规	关于互联网、AI、个人隐私数据等相关的法律、法律条款、法条解释	法律、司法解释	当地语言
	商业法律法规	商业活动相关涉及的法律及该法律概要介绍	法律、司法解释	当地语言
	违法犯罪	刑法等类似法律中明确定义的罪名称及罪行定义说明	法律、司法解释	当地语言
社会	文化禁忌	当地禁忌行为、限制性行为，包括法律法规性 详细介绍禁忌的内容、生效范围、来源	论文	当地语言
	社会冲突	当地社会主要冲突，冲突类型，成因，现象等	新闻	当地语言

内容格式

内容格式需要整理成文本形式，支持格式 txt、doc、pdf（不支持多栏排版）。文件请按照范围中一级分类、二级分类进行分类管理。