ICS 35.240.30
CCS A 10

T/GBC

团　　体　　标　　准

T/GBC 122—2026

东盟国家语料库 建设规范

Specification for development of ASEAN countries' corpus

2026－03－31 发布　　2026－04－30 实施

广西物品编码与标准化促进会　　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。

请注意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。

本文件由中国—东盟信息港股份有限公司提出。

本文件由广西物品编码与标准化促进会归口。

本文件起草单位：中国—东盟信息港股份有限公司、广西壮族自治区标准技术研究院、老挝科技与通信部数字政府管理中心、老挝工业与贸易部标准化与计量司、阿里云计算有限公司、科大讯飞股份有限公司、浪潮云信息技术股份公司、联通数据智能有限公司、北京面壁智能科技有限责任公司、三六零科技集团有限公司、中国移动通信集团广西有限公司、中兴通讯股份有限公司、广西达译科技有限公司、安徽飞数信息科技有限公司、老挝国立大学、北京海天瑞声科技股份有限公司、杭州君同未来科技有限公司、广西民族大学、北京晴数智慧科技有限公司、数据堂（北京）科技股份有限公司、整数智能信息技术（杭州）有限责任公司、马来西亚Maxeon科技、北京智源研究院、北京火山引擎科技有限公司、马来西亚Agmo集团、印度尼西亚AiSENSUM公司、广西大学、越南河内国家大学所属社会科学与人文大学、泰国INTERVEC Center、上海人工智能创新中心、泰中科技协会、泰国清迈职业技术学院、泰国北部职业教育推广与发展中心、央视国际网络有限公司、北京邮电大学、泰国彭世洛职业学院、人工智能省部共建协同创新中心（浙江大学）。

本文件主要起草人：李昌金、廖丁石、刘夏、罗宁、高健、何琛、杨霞、塔维萨·玛诺坦、拉达万·西翁赛、曲振斌、孟凡胜、陈扬、关业海、刘聪、刘丹、胡明婷、王一鸣、王宁、王斌峰、梁轶晓、林伟家、贾守盛、梁舒昱、苏良良、罗鹏、郝乔波、林志远、温家凯、邓姿娴、李雨泓、王培养、奥拉迪·坎玛尼翁、王淳、杨明、韩蒙、索佳慧、李成龙、覃秀红、陈宇、李雅婧、罗磊、彭颖岚、陈德毅、潘剑宜、幸逸冰、麦克西米利恩·瑞查德·泰、朱利恩·泰、郭聪辉、徐瑞晨、李涛、赖子谦、黄智恒、阿赫玛德·昂贡·阿拉法、维韦克·托马斯、覃希、张振荣、陈燕、裴城南、阮氏垂庄、庞利特·金大龙、何聪辉、王广宇、王珊、李晅、阿迪叻·暖西里、披迪帕·宾洛、李志学、黄建杰、喻鹏、曲昭伟、王晓茹、蒂塔里·詹他瓦、肖俊、邵健、汤永川。

东盟国家语料库 建设规范

1 范围

本文件提供了东盟国家文本及语音语料库建设全生命周期的指导和建议，规定了建设规划、语料采集、语料预处理、语料标注、验证集构建、语料质检和语料存储和管理等内容。

本文件适用于东盟国家语料库的建设工作。

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 15237—2025 术语工作及术语科学 词汇

GB/T 36344 信息技术 数据质量评价指标

GB/T 45574—2025 数据安全技术 敏感个人信息处理安全要求

YD/T 6225—2024 大数据 数据脱敏工具技术要求与测试方法

3 术语和定义

GB/T 15237—2025界定的以及下列术语和定义适用于本文件。

3.1

语料采集 corpus collection

从建设规划阶段所确定的数据源中，系统性地收集语料库的原始语料。

3.2

数据清洗 data cleaning

对数据进行审查、校验和加工处理的过程，包括去噪、去重、编码转换、语言识别与过滤等步骤。

3.3

数据转换 data transform

将验证后的数据统一转化为标准化、便于处理的格式的过程。

3.4

监督微调 supervised fine-tuning

利用人工标注的“指令—回答”数据对预训练模型进行有监督训练的方法。

3.5

基于人类反馈的强化学习 reinforcement learning from human feedback

通过人类对模型不同输出的偏好进行标注，建立一个奖励规则，根据奖励规则用强化学习算法训练监督微调后的模型。

3.6

规范化形式 C normalization form C

Unicode标准定义的一种字符规范化形式。先进行字符规范分解，再重新组合为预组合形式。

3.7

语料库 corpus

自然语言数据的集合。

[来源：GB/T 15237—2025，3.6.4]

4 缩略语

下列缩略语适用于本文件。

ASR：自动语音识别（Automatic Speech Recognition）
RLHF：基于人类反馈的强化学习（Reinforcement Learning from Human Feedback）
SFT：监督微调（Supervised Fine-Tuning）
TTS：文本转语音（Text-To-Speech）

5 建设规划

5.1 需求分析

根据人工智能模型训练的目标，明确语料的范围、内容、质量、格式等方面的要求。该阶段包括但不限于：

a) 明确人工智能模型训练的预期目标，结合应用场景与目标区域，确定需要语料的范围、内容和质量，并明确隐私、版权等法律合规要求；
b) 根据人工智能模型训练的需求，核对语料数据的可获取性、可使用性和许可条件；
c) 依据 GB/T 36344 定义的数据质量特性，构建数据质量度量方案。

5.2 总体规划

语料库的建设规划阶段主要确保语料资源满足需求分析所确立的目标，并为人工智能应用提供数据支撑。该阶段包括但不限于：

——界定语料数据的属性，包括但不限于：
- 语料类型，指语料按不同划分准则（见5.3.1）的表现形式；
- 来源分布，指语料采集渠道及其分布比例；
- 领域覆盖，指语料涉及的专业范畴；
- 语料规模，指语料的数据总量；
- 数据标签，指语料所需的标注类型、标签体系、标注规范；

——制定涵盖语料采集、预处理、标注、质检与存储管理的详细工作计划，明确各阶段的任务目标；
——建立建设过程的监控、评估与持续改进机制。

5.3 语料库结构

5.3.1 语料类型

见表1。

表 1 语料类型表

划分准则	语料类型	语料说明
模态类型	文本语料	数据载体为文本的语料
	语音语料	数据载体为音频的语料
语言种类	单语语料	完全由同一种语言构成的语料
	平行语料	由两种或多种语言的文本组成，并在词语、句子、段落或文档层面形成翻译对应关系的语料
	可比语料	指由不同语言的文本组成，内容或主题相似，但不构成翻译对应关系的语料
	跨语言语料	由不同语言的文本组成，不限定是否存在翻译对应关系或主题相似性的语料
标注状态	原始语料	未经任何加工处理的原始文本或语音数据
	标注语料	在原始语料的基础上，增加了各种语言学或应用相关的标签信息

表1　语料类型表（续）

时间维度	共时语料	用于刻画与记录语言在某一特定历史断面的静态结构与使用规则
	历时语料	用于追踪与分析语言跨越一个较长历史时期的动态演变规律与发展趋势

5.3.2　语料来源

应规划多元化获取渠道，确保来源合法、内容多样、全程可溯源，设定来源的质量基线与合规要求。

5.3.3　语料领域

应依据建设目标与应用场景，规划语料覆盖的领域类别，确保领域选择的信息内容具有代表性。

5.3.4　语料规模

应规划语料库的总体数据量目标以及不同类型、领域语料之间的规模比例。

5.3.5　语料标签

应规划所需的标注类别，设计或选用所需的标注标签，制定标注的核心原则。

5.4　语料数据格式

5.4.1　文本语料

文本语料应采用规范化形式C的UTF-8编码格式，结构化数据应采用xlsx格式，半结构化的数据应使用JSONL格式。

5.4.2　语音语料

应采用WAV格式或FLAC格式，参数见表2。

表 2　语音语料参数表

语料类型	采集环境	采样率	位深
ASR语料	电话录音、会议纪要、日常对话等	≥8 kHz	≥16bit
ASR语料	录音室访谈、语音指令采集、高精度转写	≥16 kHz	≥16bit
TTS语料	标准发音人库、有声书、媒体播报	≥44.1 kHz	≥16bit

5.5　建设管理

5.5.1　监控机制

语料采集、语料预处理、语料标注、语料质检及存储等关键环节应定期跟踪与记录。

5.5.2　评估机制

基于建设过程定期跟踪记录的数据，对建设各阶段进行综合评估。

5.5.3　改进机制

基于监控数据与评估结论，对工作计划、各阶段操作规范及管理文档进行迭代优化。

6　语料采集

语料采集过程应符合GB/T 45574—2025第6章的要求。

6.1　采集方式

原始语料的采集应根据语料库的建设目标与资源条件，选择合适的采集方式，包括但不限于：

——识别并获取已存在的公开或授权数据集并评估其适用性；

——通过向数据供应商采购，获取自身难以采集的特定领域或大规模数据；

——当现有数据无法满足需求时，应启用新数据采集方式，主要包括：

- 网络数据：通过合法合规的技术手段和工具来挖掘网络资源；
- 现实采集：在特定场景下，通过传感器、摄像机等设备，采集现实世界原始语料；
- 人工生产：对于特定的内容或高价值内容，可通过专家编写、用户生成内容（UGC）等方式采集，并进行来源标注与许可审核；
- 模型生成：通过提示工程、微调、可控生成等技术，自动或半自动地产生文本、语音等语言数据。

6.2 采集流程

语料采集流程如下：

a) 制定采集实施方案，明确数据采集的技术路径、参数配置、质量要求与进度安排；
b) 实施小规模采集测试，验证采集方法的可行性与数据质量；
c) 根据测试结果优化采集参数，调整技术方案，直至达到实施方案的要求；
d) 根据验证通过的方案开展规模化采集，实施全过程监控与日志记录，确保采集作业的稳定性及采集数据的高质量。

7 语料预处理

7.1 数据清洗

数据清洗操作包括但不限于：

——去重：识别并删除完全重合与高度重复的数据；
——文本处理：清除广告、导航栏、HTML/XML 标签、异常字符等信息；
——音频处理：去除首尾静音段、背景噪音、非目标声音等；
——编码转化：检测原始语料编码，并统一将其转换为采用规范化形式 C 的 UTF-8 编码，覆盖东盟各地区不同语言涉及的多种字符集，确保文本的正常显示；
——语言识别与过滤：使用语言识别工具，过滤非目标语言的文字，确保语料的语言纯净度。

7.2 数据转换

数据转换方式包括但不限于：

——文本归一化：统一标点符号与大小写规则，保留专名原貌，统一数字、日期、货币与单位并本地化书写；
——语言结构标准化：根据目标语言的语法规则，将连续的文本分割为独立的句子。对泰语、高棉语等需要进行分词的语言进行分词，对印尼语、马来语、德顿语等语言进行规范的词形还原处理；
——格式标准化：将所有数据文件按对应的格式要求进行转换，统一文件命名与目录结构。

7.3 数据脱敏

7.3.1 数据脱敏过程应符合 GB/T 45574—2025 第 6 章的要求，脱敏工具的使用应符合 YD/T 6225—2024 第 5 章的要求。

7.3.2 脱敏数据类型

包括但不限于：

——个人身份信息：姓名、身份证号、电话号码、住址等信息；
——文化敏感信息：东盟国家各地区文化、宗教和王室等信息；
——公共安全信息：煽动暴力信息、极端主义言论等信息；
——金融财产信息：薪资信息、交易记录、税务信息等信息；
——企业商业核心信息：东盟跨境企业贸易合同核心条款、企业跨境结算账户信息、产品核心技术参数、跨境电商平台商家营收及纳税记录等信息；
——医疗健康信息：病历、诊断记录、药物使用情况、遗传信息、健康检查报告等信息；
——地理位置轨迹信息：实时位置、历史轨迹、旅行记录、GPS 数据、行踪轨迹等信息；

——政治敏感信息：国家主权信息、敏感政治信息、边境争端等信息。

7.3.3 脱敏策略

敏感信息处理方式见表3。

表 3 敏感信息处理方式表

敏感等级	敏感信息	处理方式
低	姓名、身份证号、电话号码、住址等个人信息	替换假名、泛化数值、敏感字段加密
中	东盟国家各地区文化、宗教和王室等信息	内容替换、内容删除
中	煽动暴力信息、极端主义言论等信息	内容删除
中	东盟各国薪资信息、交易记录、税务信息等信息	替换假名、泛化数值、假数据替换
中	东盟跨境企业贸易合同核心条款、企业跨境结算账户信息、产品核心技术参数、跨境电商平台商家营收	泛化数值、替换假名
中	病历、诊断记录、药物使用情况、遗传信息、健康检查报告等信息	泛化数值、敏感字段加密
中	实时位置、历史轨迹、旅行记录、GPS数据、行踪轨迹等信息	泛化数值、敏感字段加密
高	国家主权信息、敏感政治信息、边境争端等信息	泛化数值、内容删除

7.4 数据验证

数据验证是确保语料库质量的基础，其目的是识别并剔除不符合要求的数据。具体处理方式包括：

——完整性验证：检查数据记录是否存在关键字段缺失；

——格式合规性验证：检查文本语料是否采用规范化形式 C 的 UTF-8 编码格式，检查音频语料格式、采样率、位深等是否符合要求；

——数据质量校验：对文本数据进行质量校验，如检查句子是否完整，是否存在大量无意义的字符重复。对音频文件进行质量校验，如是否存在信号失真、持续静音时间过长、背景噪音过大等问题。

8 语料标注

8.1 标注基本原则

包括但不限于：

——明确标注语料模态以及其覆盖的领域。针对东盟国家的语料，标注团队应包含熟悉目标语言者、母语者以及负责审核与文化、宗教、习俗相关内容标注恰当性的文化审核者；

——明确定义标注标签，标签应具有明确的命名规范和层次结构。确保不同人员、不同批次的标注结果具有一致性；

——标注过程中需记录标注的操作日志，对存疑或修改的标注数据应采用版本控制与变更日志，并区分训练、验证和测试集的变更边界，便于问题追溯和标注优化；

——建立统一的标注规范，明确界定分类模糊样本的处理规则，并制定解决标注分歧的裁定程序。

8.2 SFT 语料

8.2.1 核心要求

构建高质量、多样化“指令—响应”对，指导模型遵循指令完成特定任务。

8.2.2 标注原则

包括但不限于：

——基于原始语料的核心内容与目标应用场景，设计覆盖多种类型的指令任务。任务设计应体现多样性，并充分考虑东盟地区的语言习惯、文化习俗、社会规范及常见应用场景，以全面培养模型在跨文化语境下的理解与执行能力；

——针对每一条指令，应生成与之精确匹配的高质量回答。回答的生成应确保其源于语料的客观事实或经过验证的领域知识，并严格遵循东盟地区的核心价值观，避免涉及种族、宗教、王室、地域等敏感话题，确保内容的得体性与安全性；

——将指令与回答组合成结构规范的对话单元，形成模型可直接学习的训练实例。

8.3 RLHF 语料

8.3.1 核心要求

构建高质量、安全的“指令—正负面对比回答”数据对，训练模型识别并生成符合人类价值观的安全、有益回复。

8.3.2 标注原则

包括但不限于：

——设计具有挑战性的指令任务。任务设计应覆盖多种风险类型，如偏见、歧视、误导及东盟地区特有的文化、宗教、政治及社会敏感性风险点等，以全面培养模型的辨别与抵御能力；

——针对每一条指令，应同步生成高质量的正面对比回答与负面对比回答。正面回答的生成应确保其安全性、建设性并符合客观事实；负面回答应体现典型的有害模式，但需控制在合理范围内以供模型对比学习；

——将指令与正负面回答组合成结构规范的对比学习单元，并经过东盟地区文化与领域专家评审意见的严格质量审核，形成用于强化学习训练的高质量偏好数据。

8.4 价值观语料

8.4.1 核心要求

构建体现核心价值观的“指令—正负回答”数据对，设计体现核心价值观的指令任务，训练模型准确理解并传递符合社会文化规范的价值理念。

8.4.2 标注原则

包括但不限于：

——设计体现核心价值观的指令任务。任务设计应覆盖文化认同、社会和谐、法律意识、可持续发展等多个维度，全面培养模型的价值判断能力；

——针对每一条指令，应同步生成符合主流价值观的正面回答与偏离核心价值观的负面回答。正面回答应准确体现对象国的文化精髓和社会共识，负面回答应展现典型的价值观偏离现象，但需确保其作为对比样本的合理性；

——将指令与正负面回答组合成结构化的价值观学习单元，并通过文化专家的审核认证，形成用于价值观对齐训练的高质量语料库。

8.5 平行语料

8.5.1 核心要求

构建高质量、精准对齐的双语或多语“源语言—目标语言”句对，为模型的跨语言理解与生成能力提供基础训练数据。

8.5.2 标注原则

包括但不限于：

——应确保译文完整、精确地传递源文本的全部语义信息与内在逻辑；

——应确保译文完全符合目标语言的语法规范、惯用搭配与自然表达习惯；

——应确保译文的语体、情感基调及修辞手法与源文本的文体风格相匹配；

——将源语言与目标语言文本组合成结构规范的对齐单元，并经过双语语言专家和领域专家的严格校对与质量验证，形成用于监督学习的高质量翻译对。

8.6 ASR 语料

8.6.1 核心要求

构建高质量、高匹配度的“音频—文本”对齐数据对，为模型的语音识别与转写能力提供基础训练数据。

8.6.2 标注原则

包括但不限于：

——基于音频数据的实际内容，遵循所听即所转；

——文本转写应严格遵循音频的语音内容，涉及大小写、标点、数字以及混说转写等，应遵循转写规范，保持一致性；必要时提供说话人分离与时间戳；

——将音频与文本数据组合成结构规范的训练单元，并经过严格的听检校对与质量验证，形成用于语音识别训练的高质量对齐数据。

8.7 TTS 语料

8.7.1 核心要求

构建高质量、高表现力的“文本—音频”对齐数据对，为模型的文本转语音能力提供基础训练数据，确保合成的语音自然流畅、富有表现力。

8.7.2 标注原则

包括但不限于：

——基于文本内容以及目标播报风格，录制与之完全匹配的高质量音频数据。录音过程应确保发音准确、语调自然、情感表达恰当；

——针对每一条文本数据，应提供与之精确对应的音频内容。音频应严格遵循文本的语义和语法结构；

——将文本与音频数据组合成结构规范的训练单元，并经过严格的听检校对与质量验证，形成用于 TTS 训练的高质量对齐数据。

9 验证集构建

9.1 核心要求

验证集应满足以下核心要求，以作为有效的标准化评估基准：

——效度要求：应具备良好的内容效度与结构效度，能够准确测量其评估的模型能力或风险维度，并提供效度证据；

——信度要求：应具备高可靠性，评估结果应具备高可靠性与可复现性，并提供信度证据；

——质量要求：应经过严格的难度分析与质量控制，确保其设计科学、质量可靠；

——文档要求：应提供完整的技术文档，说明其评估目标、使用方式、统计特性及版本信息。

9.2 构建原则

验证集的构建与管理应遵循以下原则：

——目标明确：应围绕具体、可衡量的评测目标展开，并将其分解为定义清晰、相互独立且可量化的评估维度；

——流程规范：应建立覆盖数据收集、清洗处理、科学配比、指标设计及综合评分的标准化全流程作业规范；

——全程质控：应对构建与使用的各环节实施系统化质量控制，并建设自动化评测系统保障流程一致性；

——安全合规：应确保内容、数据及处理过程符合法律法规与伦理要求，落实全链路安全管控措施。

10 语料质检

10.1 质检方法

语料质检宜采用多方法、多维度结合的检验策略，主要方法包括但不限于：

——自动化检查：利用脚本与工具对语料进行批量检查，主要覆盖格式规范性、编码统一性、基础元数据完整性、长度异常及敏感词过滤等基础性指标与语言流畅度、内容逻辑性、知识密度等语料质量指标；

——人工抽样审核：由质检员对已标注语料进行抽样审查，重点评估其内容的准确性、流畅性、逻辑一致性以及对标注原则的符合程度；

——交叉验证：对于关键语料，安排不同标注员对同一批次语料进行独立校验，通过对比差异评估标注一致性；

——专家评审：针对专业性强的语料或涉及文化敏感、价值观判断的语料，应由领域专家进行最终评审，确保内容的专业性与安全性。

10.2 质检流程

遵循流程闭环、多阶段的原则，以确保问题能被及时发现和修正，具体流程如下：

——标注员自检：标注任务完成后，标注员首先依据标注规范对自身产出的语料进行检查与修正；

——质检员初审：质检员从通过自检的语料中按抽样比例不少于 1/50、关键语料不少于 1/10 进行抽样、审核。若批次合格率低于 95%，则该批次语料退回全量返工；合格率大于等于 98%，则进入快速入库通道；

——专家/主管终审：对于通过初审的语料，尤其是高敏感、高专业度、高精确度要求的数据，进一步小样本交叉验证并由专家最终裁定，确认其是否可入库；

——问题记录与反馈：在整个流程中详细记录质量问题，并定期汇总反馈至标注团队，用于案例复盘与标注标准优化；

——质量分析与标准优化：定期生成质检报告，分析常见错误类型与趋势，并据此对标注指南和质检标准进行迭代更新，持续提升语料质量。

11 语料存储和管理

11.1 语料分类与归档

语料应按来源、类型和加工阶段进行分类、编制目录并归档，确保原始语料可追溯性。

11.2 备份与恢复

制定并执行定期备份策略，将备份数据异地保存，并定期演练恢复流程，确保业务连续性。

11.3 元数据管理

建立标准化的描述性、结构性和管理性元数据，并将其与语料数据关联存储与管理。

11.4 存储策略

制定长期保存规划，通过格式迁移、封装等技术确保数据的长期可读性与可用性。