语　言　文　字　规　范

GF 0031—2026

人工智能　语料库　基础术语

Artificial intelligence — Corpus — Basic terminology

2026-03-02 发布　　2026-07-01 实施

中华人民共和国教育部
国家语言文字工作委员会　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。

本文件由教育部语言文字应用研究所提出。

本文件由教育部语言文字信息管理司归口。

本文件由国家语言文字工作委员会语言文字规范标准审定委员会审定。

本文件由教育部、国家语言文字工作委员会发布。

本文件起草单位：教育部语言文字应用研究所（国家语委普通话与文字应用培训测试中心）、北京语言大学、语文出版社有限公司、中国社会科学院语言研究所、北京海天瑞声科技股份有限公司、全国科学技术名词审定委员会事务中心、北京外国语大学、南京师范大学、暨南大学、数据堂（北京）科技股份有限公司、科大讯飞股份有限公司、中移动信息技术有限公司、北京师范大学、华中师范大学、中国传媒大学、人民教育出版社有限公司、上海外国语大学、江苏师范大学、中国社会科学院大学、标贝（青岛）科技有限公司、北京希尔贝壳科技有限公司、安徽极致思维智能科技有限公司、哈尔滨师范大学、上海库帕思科技有限公司、北京国际算力服务有限公司、安徽声云智能科技有限公司、中央民族大学、内蒙古大学、北京航空航天大学、北京邮电大学、北京软件和信息服务业协会、东北大学、浙江外国语学院。

本文件主要起草人：刘朋建、李慧、刘培俊、刘露翌、饶高琦、富丽、陈茜、郝玉峰、张永伟、许家金、李斌、刘华、于春迟、王敏、王永强、汪张龙、王铁琨、王立军、荀恩东、张劲松、肖永红、王莉宁、沈威、刘鼎甲、杜振雷、程书秋、胡韧奋、张弛、苏祺、丁石庆、王翠叶、文秋芳、程荣、张天伟、贾媛、吴坤、陈先云、何婷婷、田由甲、冯晓莉、高迎明、陈霖、熊文新、郝瑜鑫、穆向禹、卜辉、王贵荣、陈明、达胡白乙拉、梁茂成、尹鸾飞、王会珍、黄海清、乔思渊、韩国仕、徐昕、李琳、李文曦。

引　言

随着人工智能技术的发展，语料库在智能语音、自然语言处理和多模态理解等领域中发挥基础支撑作用，相关术语不断丰富。

本文件围绕人工智能领域语料库的基础术语，对共性、基础性概念进行整理和规范表述，并依据语料库在人工智能中的属性特征、建设过程和应用方式，对相关术语进行分类说明，形成基础术语框架。

本文件所涉及的术语包括：

——属性术语，用于界定人工智能语料库的通用概念和基础属性，对语料库的结构单位、内部特征及外部分类等方面的术语进行表述；

——建设术语，用于界定人工智能语料库建设过程中涉及的相关术语，围绕语料库的全生命周期，对设计、采集、处理、标注、质量控制和管理等阶段的常用术语进行表述；

——应用术语，用于界定人工智能语料库在实际应用中涉及的相关术语，对语料库的载体建立、共享流通和使用方式等方面的术语进行表述。

人工智能　语料库　基础术语

1　范围

本文件界定了语料库在属性、建设和应用方面的常用术语和定义。

本文件适用于人工智能领域语料库的建设、使用、管理、测评和研究等工作。

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 37988—2019　信息安全技术　数据安全能力成熟度模型

GB/T 40035—2021　双语平行语料加工服务基本要求

3　术语和定义

3.1　语料库属性术语

3.1.1

语料　corpus

语言材料或其他模态数据，通常以文本、语音、图像、音频或视频等模态形式存在。

3.1.2

语料库　corpus

按照特定目的，经过系统性采集、加工后形成的大规模、可机读的语言及其他模态数据集合。

3.1.3

库容　corpus size

语料库中包含语料的量，通常以条数、字数、词数、字节数或小时数等衡量。

3.1.4

词元　token

原始文本中可识别和建模的最小基本单元。其粒度可以是单词、子词、字符及多模态基本符码等。

注：将文本切分为词元的过程称为“词元化”。

3.1.5

元数据　metadata

关于语料内容、结构、来源、质量、状况及其他特性的描述性数据。

［来源：GB/T 40035—2021，3.7，有修改］

3.1.6

自然语料　natural corpus

直接来源于真实世界语言交际的语料。

3.1.7

合成语料　synthetic corpus

通过人工规则、传统算法或人工智能模型自动生成的语料，通常用作自然语料的替代或补充。

3.1.8

生语料　raw corpus

未经人工或系统性加工处理的原始语料。

3.1.9

标注语料　annotated corpus

对生语料进行人工或算法加工，添加了标注信息后形成的语料。

注：别名为“熟语料”，与“生语料”相对。

3.1.10

结构化语料　structured corpus

经过标准化处理与深度标注，并具备预定义数据模型和固定格式的语料。

注：其内容通常存储在数据库表或特定字段中，整合了标签与元数据，机器无须深度语义解析即可高效处理。其特点是模式固定、高度组织化。

3.1.11

半结构化语料　semi-structured corpus

一种介于结构化和非结构化之间的语料，通过内嵌的标签、标记或层级结构来组织内容，其数据模式具备灵活性。

注：它虽不具备严格的全域模式，但内含的标记为其提供了一定的机器可读性，允许字段缺失、重复或嵌套。常见形式包括 JSON、XML 等格式的数据。

3.1.12

非结构化语料　unstructured corpus

未经过预定义数据模型格式化，以原始自然形态存在，缺乏统一字段或标签的语料。

注：机器在利用前应对其进行采集、清洗及深度语义解析。此类语料覆盖范围极广，包括纯文本、图像、音频、视频等，是最原始和普遍的数据形态。

3.1.13

预训练语料　pre-training corpus

用于模型基础训练的大规模、未标注的语料，涵盖广泛的主题和领域，为模型提供通用的语言知识基础。

3.1.14

微调语料　fine-tuning corpus

在预训练语言模型基础上，用于优化模型性能的结构化标注语料，包含特定任务相关指令和对应答案，引导模型理解和完成具体任务。

3.1.15

评测语料　evaluation corpus

用于在模型开发完成后，评估其各项性能指标的、经过精心构建和标注的语料。

注：该语料须与模型的预训练和微调语料等严格隔离，以确保评估结果的可靠性和泛化性。

3.1.16

通用语料库　general corpus

旨在代表语言整体使用情况而构建的大规模语料库，其内容覆盖广泛的语言现象，具有跨领域的代表性。

3.1.17

领域语料库　domain-specific corpus

针对特定领域、行业或应用场景构建的语料库，用于反映该领域的专用语言特点和规律。

注：涵盖的领域包括但不限于教育、金融、医学、法律、科技等。

3.1.18

单模态语料库　monomodal corpus

由单一模态的语料及其相关标注构成的语料库。

3.1.19

多模态语料库　multimodal corpus

包含两种或多种不同模态的语料，并通过语义对齐、时序关联等技术实现跨模态组织与关联的语料库。

注：其核心在于模态间的关联与对齐。例如，在“图像—文本”对中，文本须精确描述图像内容；在“视频—音频—文本”数据中，音频、字幕须与视频画面在时间轴上精确同步。

3.1.20

文本语料库　text corpus

由文本数据构成的语料库，可包含相关标注。

注：文本数据可来源于书面、转写或合成文本。

3.1.21

语音语料库　speech corpus

由人类语音数据构成的语料库，通常包含转录等标注，也可包含合成语音。

3.1.22

图像语料库　image corpus

由图像数据构成的语料库，可包含相关标注。

注：常见的标注包括分类标签、边界框和文本描述等。

3.1.23

音频语料库　audio corpus

由各种音频数据构成的语料库。

注：包含语音、音乐和环境声等。

3.1.24

视频语料库　video corpus

由时序视觉数据及其相关标注构成的语料库，通常伴有音频流，标注可包括动作标签、文本描述、时间戳等。

3.1.25

单语语料库　monolingual corpus

仅包含一种语言语料的语料库。

3.1.26

多语语料库　multilingual corpus

包含两种及以上语言语料的语料库。

3.1.27

平行语料库　parallel corpus

由两种或多种语言的对齐文本组成的语料库。

3.2 语料库建设术语

3.2.1

语料库设计 corpus design

依据语料库的总体目标与建设规划而进行的系统性顶层规划过程。

注：主要包括两个阶段，一是总体规划，即明确语料库的目标、范围、代表性、规模、元数据与标注规范以及伦理法律框架；二是技术方案制定，包括数据采集、预处理、标注、存储和检索的具体方法与工具。

3.2.2

语料采集 corpus collection

从不同来源获取原始语料，并将其整理为可机读形式的过程，此过程严格遵循预先制定的采集范围、方法和标准。

3.2.3

语料预处理 corpus pre-processing

语料采集后，对语料进行清洗、脱敏等处理，是语料标注和分析等工作的基础。

3.2.4

语料清洗 corpus cleaning

对原始语料进行去重、去噪、纠错和格式统一等处理，以提升语料质量的过程。

3.2.5

语料脱敏 corpus de-identification

去除语料数据中可确认个人或组织身份的信息与数据主体之间联系的过程，以降低语料在采集、处理和使用过程中的隐私或安全风险。

[来源：GB/T 40035—2021，3.14，有修改]

3.2.6

语料安全 corpus security

通过管理和技术措施，确保语料处于有效保护和合法利用的状态。

[来源：GB/T 37988—2019，3.1，有修改]

3.2.7

语料标注 corpus annotation

对语料或语料中的数据单元进行标记、分类或注释，使其转化为机器可读、可计算的结构化数据。

3.2.8

标注规范 annotation guideline

对语料标注过程中涉及的标注对象、标注标准、标注符号及操作流程等作出统一规定，以规范和指导标注行为。

3.2.9

标注方法 annotation method

实施语料标注的具体方式、流程或技术策略。

注：包括人工标注、自动标注或智能辅助标注等。

3.2.10

人工标注 manual annotation

由标注员对语言数据进行标注的过程。

3.2.11

自动标注　automatic annotation

利用计算机程序或算法自动对语言数据进行标注的过程。

3.2.12

智能辅助标注　AI-assisted annotation

利用人工智能模型辅助人类进行数据标注，以提升效率与质量的人机协同过程。

3.2.13

分词　tokenization

将连续的文本字符序列切分成独立的最小单位的过程。

3.2.14

转写　transcription

将音频或视频等非文本语料中的语音转化为文本的过程。

3.2.15

语料对齐　corpus alignment

将双语或多语语料在篇章、段落、句子或其他层级上建立对应关系，使其构成相互对照形式的处理过程。

注：在人工智能应用中，语料对齐也可扩展用于语音、图像等不同模态语料之间的对应处理。

［来源：GB/T 40035—2021，3.11，有修改］

3.2.16

质量控制　quality control

通过制定标准、一致性检查、纠偏优化等系统性措施对语料库的质量进行系统化监测的过程，以保障语料质量符合质量指标。

3.2.17

质量评估　quality assessment

在语料库建设完成或特定阶段，依据质量指标对其进行系统评价的过程。它主要关注语料库是否满足质量目标和使用需求，并在过程中进行持续监测和调整。

3.2.18

质量指标　quality metrics

用于系统评估语料库整体质量的具体标准和参数。它从多个维度对语料库进行量化评估，通常包括规范性、代表性、一致性、完整性等。

3.2.19

规范性　standardization

语料库在语言规范和技术标准层面符合既定要求的程度。

3.2.20

代表性　representativeness

语料库反映目标语言的总体特征、分布规律或特定领域语言模式的准确程度。

3.2.21

一致性　consistency

语料库在其标注标准、数据格式和执行流程上保持统一的程度。

3.2.22

完整性　completeness

语料库包含其特定目标所需全部数据、标注信息及元数据的充分程度。

3.2.23

抽样检查　sampling inspection

依照统计学方法从语料库中抽取代表性子集，检测其质量并推断整体语料质量的过程。

3.2.24

价值对齐　value alignment

确保人工智能系统的目标与人类的价值观和利益保持一致的过程。

3.2.25

安全合规　security and compliance

语料库在建设、使用和管理过程中，对法律法规、安全标准及道德规范的符合性。具体体现在数据来源与处理的合法性、数据存储与使用的安全性、用户隐私的有效保护以及对伦理风险的全面规避。

3.2.26

语料存储　corpus storage

将采集、加工后的语料以系统化、结构化的方式进行保存和管理的过程，以便后续的检索、分析和再利用。

3.2.27

语料封装　corpus packaging

将语料及其相关的元数据、标注工具、标注规范和语料说明等进行打包整合，形成一个相对独立的、便于存储、传输和使用的集合。

3.2.28

语料说明　corpus documentation

对语料库的相关信息进行详细描述的文档或内容，包括语料来源、采集方法、数据规模、数据格式、标注规范、使用范围、版权信息等，帮助使用者了解语料库的基本情况和使用要求。

3.3　语料库应用术语

3.3.1

语料库平台　corpus platform

以语料库为基础，集成语料存储、标注、检索、分析和管理等功能的综合性应用平台。

3.3.2

语料库发布　corpus release

将语料库对外公开的过程。

3.3.3

语料库授权　corpus licensing

语料库持有方（或权利方）通过许可协议，向使用者授予特定使用权限的行为。

3.3.4

语料库访问　corpus access

使用者通过平台或接口获取语料库中特定语料的过程。

3.3.5

语料库共享　corpus sharing

将语料库在一定范围内向其他使用者开放的过程，允许他人使用和再加工等，以促进语料库的充分利用。

参 考 文 献

［1］GB/T 41867—2022　信息技术　人工智能　术语

［2］GB/T 42755—2023　人工智能　面向机器学习的数据标注规程

［3］GB/T 15237—2025　术语工作及术语科学　词汇

［4］ZYF 001—2018　语料库通用技术规范

［5］国家数据局 . 数据领域常用名词解释（第一批）［EB/OL］.（2024-12-30）［2026-3-1］. https://www.nda.gov.cn/sjj/zwgk/zcfb/1230/20241230160715745237413_pc.html

［6］中华人民共和国数据安全法［Z］.2021

［7］语言学名词审定委员会 . 语言学名词［M］. 北京：商务印书馆 ,2011.

索　引

汉语拼音索引

Z

英文对应词索引

A

C

D

E

F

G

I

M

N

P

Q

R

S

T

U

V