ICS 25.160.01
CCS J 33

CWA

团　　体　　标　　准

T/CWAN 0175—2026

焊接大语言模型的评价方法

Evaluation method for welding large language model

2026-01-07 发布　　2026-02-01 实施

中国焊接协会　发布

目　　录

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。

本文件由中国焊接协会提出并归口。

本文件起草单位：北京理工大学（珠海）、山东大学、北京博清科技有限公司、中焊科技发展（哈尔滨）有限公司、天津市特种设备监督检验技术研究院、重庆三峡学院、南昌航空大学、上海中巽科技股份有限公司、江苏北人智能制造科技股份有限公司、天津大学、唐山松下产业机器有限公司、哈尔滨职业技术大学、北部湾大学、广西柳工机械股份有限公司、哈尔滨华德学院、中国兵器工业集团航空弹药研究院有限公司、南昌职业大学、福建省特种设备检验研究院、黑龙江科技大学、无锡合泰教育咨询有限责任公司、坤智大数据科技（哈尔滨）有限公司、威海职业学院、黑龙江工程学院。

本文件主要起草人：于兴华、孙震、冯消冰、武鹏博、马青军、尹立孟、陈玉华、王铭秋、林涛、刘晨曦、刘金龙、刘建国、邓军林、侯国清、朱斌海、范东辉、张大林、孙明辉、黄小宇、王永东、于修和、郝亮、李爱民、隋英杰、刘洋、于春洋、牛董山钰、方乃文。

焊接大语言模型评价方法

1 范围

本文件规定了焊接大语言模型（welding large language model，Weld LLM）的评价原则、评价维度与指标体系、数据集构建方法、评测流程、评分与等级划分方法、检验记录要求，并提供了评价报告格式。

本文件适用于以自然语言处理为核心的焊接大语言模型的评价，包括通用型大语言模型在焊接领域的能力评估，以及面向焊接领域开发的专用大语言模型性能测试与验证。

本文件不适用于对非自然语言处理类的焊接人工智能系统（如基于规则的专家系统、单纯的图像识别算法）的独立评价，但可为此类系统与 LLM 融合后的整体性能评估提供参考。

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 3323 焊缝无损检测 射线检测

GB/T 3375 焊接术语

GB/T 5185 焊接及相关工艺方法代号

GB/T 6417 金属熔化焊接头缺欠分类及说明

GB/T 11345 焊缝无损检测超声检测技术、检测等级和评定

GB/T 12467 金属材料熔焊质量要求

GB/T 19418 钢的弧焊接头缺陷质量分级指南

GB/T 20737 无损检测通用术语和定义

GB/T 34628 焊缝无损检测 金属材料应用通则

GB/T 45288 人工智能大模型

3 术语和定义

GB/T 3375 和 GB/T 45288 界定的以及下列术语和定义适用于本文件。

3.1

焊接大模型 welding large-scale model

可用于焊接工艺优化、焊接缺陷检测、焊接过程控制等基于大规模数据和人工智能技术构建的具有大规模参数和复杂结构的模型。

3.2

单模态 mono-modal

和焊接相关的文本、图像、音频的任意一种数据类型。

3.3

多模态 multi-modal

和焊接相关的图文、图音、文音、或图文音的任意多种数据类型。

3.4

焊接数据 welding data

与焊接过程相关的基础知识、多模态数据、焊接质量检测结果等。焊接多模态数据包括但不限于焊缝图像/视频、焊接电流、电压、焊速、焊枪摆幅、左侧停留时间、右侧停留时间、左摆速、右摆速、材料参数、热成像、熔池图像/视频、声发射信号等，环境温、湿度等。

3.5

数据标注 data labeling

对焊接数据（如图像、视频、传感器信号）进行人工或半自动化标记，赋予其语义标签（如焊接质量合格、缺陷类型等）。

3.6

预训练 pre-training

在高质量焊接领域知识数据上对模型进行初步训练，以提取基础特征，后续通过微调适配焊接场景。

3.7

微调 fine-tuning

在预训练模型基础上，使用焊接领域优质小样本标注数据，提升模型在特定任务上的性能。

3.8

词元 token

词元是指文本处理的最小语义单元，是模型输入、输出及计算的基本单位，通常是经过分词算法分割后的可处理的离散单元，通常情况下中文 1token≈2 汉字，英文 1token≈4 字母。

3.9

焊接大语言模型 welding large language model

在大规模通用语料及焊接专业语料上训练，并针对焊接领域任务进行优化或微调的自然语言处理模型，能够以自然语言形式回答、推理、分析或生成焊接相关内容。

3.10

焊接基准数据集 weld QA-benchmark

面向焊接领域的多维能力评估基准数据集，涵盖知识面、推理能力、实践应用能力等维度，包含多项选择题。

3.11

标准符合度 standard compliance

模型在多项选择题中回答涉及焊接行业标准、法规、规范等内容时，选择结果与相关标准一致的程度。

3.12

专家共识突破指数 expert consensus breakthrough index, ECBI

衡量模型在多选题中选择超出专家共识的创新性方案的有效性。

3.13

工艺创新有效性系数 process innovation validity, PIV

模型提出的工艺创新方案，经仿真或实验验证后有效的比例。

3.14

无损检测盲测准确率 non-destructive testing blind accuracy, NDT-BA

在不提供检测图谱，仅提供材料成分、焊接热输入、拘束度及环境参数的前提下，模型准确预测潜在缺陷类型、位置及形态的能力。

3.15

强干扰项 hard distractor

在多项选择题中设计的干扰选项。该选项在理论公式或一般教科书层面看似正确，但在特定复杂工况（如高海拔、极端温湿度、特殊材料组合）下为不可行方案，用于区分模型是否具备工程实践经验。

3.16

防检索鲁棒性 anti-retrieval robustness, ARR

题目在主流搜索引擎中无法直接命中答案的比例。

4 评价方法概述

4.1 总体原则

焊接大语言模型的评价应遵循以下原则：

1) 科学性：评价体系应建立在焊接行业知识体系与 AI 测评技术结合的基础上，确保指标设计符合焊接工程实践和 AI 评价规范；

2) 客观性：测试过程应避免人工主观干预，确保结果可复现，并采用标准化评分机制；

3) 全面性：通过多项选择题全面覆盖焊接知识，掌握推理能力、实际应用能力及创新能力；

4) 对抗性：测试集必须具备对抗性，确保无法通过搜索引擎或基础 RAG（检索增强生成）获取直接答案。

4.2 评价维度

建议的评价维度如下（可根据实际需求调整）：

1) 知识面：10%；

2) 推理能力：45%；

3) 实践能力：45%。

4.3 拓展创新指标

除基础维度外，应关注模型的创新能力与预测能力，包括：

1) ECBI：用于评估模型在处理“反直觉”工程问题时的表现，计算公式为

$$ECBI = \frac{\text{专家认可的创新性答案数}}{\text{模型提出的创新性答案总数}} \times 100\%$$

2) PIV：度量模型在多项选择题中选择焊接工艺创新方案的有效性，经仿真或实验验证的比例，计算公为

$$PIV = \frac{\text{经验证有效的创新方案数}}{\text{模型提出的创新方案总数}} \times 100\%$$

3) ARR： 题目在主流搜索引擎中无法直接命中答案的比例（要求 100%）。

4) NDT-BA：用于评估模型对“工艺-缺陷”物理因果链的掌握程度，计算公式为

$$NDT - BA = \frac{\text{预测结果与实际 NDT 结果一致的题目数}}{\text{总题目数}} \times 100\%$$

4.4 测试类型

1) 静态问答测试；

2) 场景分析测试；

3) 交互式决策测试；

4) 实时工艺干预测试；

5) 长周期稳定性监测。

5 数据及构建技术要求

5.1 数据来源及采集原则

数据集应来源于权威、公开、可验证的资料及专家原创内容，确保覆盖焊接领域的主要知识模块，并符合数据隐私和知识产权要求。来源包括但不限于：

a) 标准与规范：国家标准、行业标准、团体标准及企业标准等，重点采集与工艺参数、缺陷评定、材料要求等相关的条款，并转化为可测试的多项选择题形式。

b) 专家案例：经行业认可的焊接专家提供的工艺案例、缺陷分析报告、事故调查记录等，提炼其中的情景与关键特征，形成多项选择题。

c) 考试与教材：焊接工程师资格考试题库、焊工国家职业资格考核题库、教材中的典型题目与习题。

采集时应确保数据不涉及商业机密和个人隐私信息，在必要时进行匿名化处理，并验证数据准确性和多样性以避免偏差。

5.2 题型设计与比例要求

本标准采用多项选择题（MCQ）作为唯一题型，以增加难度、避免随机选择的干扰，并便于标准化评分。题型设计应聚焦于多选题形式，干扰项设计需科学，以考察模型的细微知识区分能力。所有考核维度（知识面、推理能力、实践能力、创新能力）均通过多选题实现，例如：

知识面：选择正确的基础概念或标准条款；

推理能力：选择多步骤逻辑推导的正确结论；

实践能力：选择最佳工艺优化方案；

扩展指标：通过设计需创新思维的选择项评估 ECBI、PIV、ARR、NDT-BA。

多项选择题（MCQ）：100%，便于标准化评分。

5.3 难度分级

题库应按难度分为四级（Lv1、Lv2、Lv3、Lv4），并保持均衡分布。难度确定方法基于专家共识，对标人类教育和职业水平。Lv1 考察基础概念；Lv2 考察多步骤推理能力；Lv3 考察综合分析复杂情景能力；Lv4 考察开放性问题和创新思维，但均通过多选题形式实现（如设计多个干扰项），难度分布如下：

Lv1（基础）：约 15%，考察焊接基础知识与简单概念；

Lv2（进阶）：约 25%，涉及多步骤推理或组合知识应用；

Lv3（专业）：约 30%，深度推理，需要跨学科知识（冶金学+力学+传感技术等）综合推理才能解答；

Lv4（专家）：约 30%，超高难度问题，涉及行业公认难题或无定论领域，要求模型给出经实验验证有效的创新解。

难度分级应经专家共识确定，并定期审视以维持科学性。Lv3 和 Lv4 级别的题目数据源需要无法用主流搜索引擎搜索到，例如：

* 未公开的失效分析报告： 企业内部的事故调查、裂纹分析复盘。

* 实验室原始数据：包含负面结果（失败试验）的焊接工艺评定记录（procedure qualification record, PQR）。

* 极端工况实测数据： 高原、深海、强辐射等特殊环境下的焊接工艺验证数据。

5.4 对抗性数据清洗

每一道入选题目必须经过“搜索引擎测试”。将题干关键词输入主流搜索引擎，若在前三页结果中能找到直接答案或高度相似案例，该题目必须剔除或重写。

5.5 标签与元数据管理

每道题应附加以下标签：

a) 知识模块标签：如“材料学”、“焊接方法”、“质量检验”、“标准规范”等；

b) 认知类型标签：如“计算题类”、“诊断题类”、“规划题类”、“解释题类”等；

c) 标准/规程引用标签：标明引用的标准号及条款（如 GB/T 6417, 相关条款）；

d) 标签应以结构化数据格式存储（JSON 或 CSV），便于自动筛选与统计，并支持版本控制。

5.6 质量控制与验证机制

Lv1 和 Lv2 出题人员应具有焊接工程师或同等资质的行业专家；

Lv3 出题人员应具有材料加工工程博士学位，或 10 年以上专项研发经验；

Lv4 题目需采用“双专家制”，由高校教授（理论）与企业总工艺师（实践）联合命题。

a) 双专家验证：每道题须由两名材料加工工程博士学位，或 10 年以上专项研发经验人员审核，确保表述清晰、无歧义，且正确答案准确无误；

b) 动态淘汰机制：题库实行年度动态更新机制。结合当前主流搜索引擎及其常规检索能力，对题目进行可检索性评估。凡是能够通过简单关键词搜索即可在公开网络资源中直接获得明确标准答案的题目，判定为信息可得性过高、认知门槛过低，应从题库中予以剔除或转出核心题库;

c) 审计记录：每道题的来源、修改记录、审核意见应保存，以备追溯，并符合数据治理规范。

6 评价实施细则

6.1 功能要求

6.1.1 通用基础能力

6.1.1.1 测试环境与准备

a) 测试应在可控、隔离的环境中进行，禁止模型访问外部互联网或非授权数据库，以确保公平性和安全性；

b) 测试建议使用统一的硬件条件如表 1 所示，以模拟实际部署环境，确保所有模型在相同水平下进行答题。

表 1 测试项目与建议配置要求

项目	配置要求
CPU	≥Intel i9 / AMD Ryzen 9（核心≥16）
GPU	≥ NVIDIA RTX 4090 / A100 / 同等级显卡（显存≥24GB）
内存	≥128GB
存储	≥1TB NVMe SSD
操作系统	Ubuntu 22.04 LTS 或等效 Linux 系统
网络	离线/局域网隔离环境

6.2 测试流程

测试流程基于多项选择题实现，模型在指定硬件条件下直接答题，考核知识面、推理、实践和创新能力，以答题最终得分为准进行评级。

6.2.1 静态问答测试

内容：独立的多项选择题集，涵盖知识面与基本推理；

限制：每道题目答题时间限时 2 分钟，超时认定为答错，每题独立作答；

目的：测量模型的知识覆盖度与基础推理能力。

6.2.2 场景分析测试

内容：含多条件的情景描述的多项选择题；

要求：模型需选择整合所有信息的正确结论；

目的：考察模型在复杂场景下的多步骤推理能力。

6.2.3 交互式决策测试

内容：模拟决策场景的多项选择题，根据描述反馈选择调整方案；

示例：根据焊接过程中缺陷迹象描述，选择动态调整工艺参数的选项；

目的：评估模型在决策优化中的选择能力。

6.2.4 实时工艺干预测试

内容：模拟焊接参数监控的多项选择题；

要求：选择发现偏离正常迹象时的干预指令；

评分：根据选择的正确性进行量化。

6.3 评分与加权

a) 各维度分值权重相同：

对于多项选择题（MCQ）的评分规则如下：全对得满分；少选扣分；多选不得分；干扰项设计需确保区分模型的精确知识掌握。

b) 总分即为最终的得分。

6.4 等级划分

等级划分基于与人类基线对比（如焊工等级考试），达到相应水平后，模型可安全应用于对应任务，体现评价标准的实用价值，模型等级确定与能力对标如表 2 所示。

表 2 模型等级确定与能力对标

等级	分数	能力
Lv1	40－59	能回答基础知识，缺乏复杂推理与优化能力，能处理简单焊接查询，如基本概念解释。
Lv2	60－74	能应对常见问题，有一定推理与优化能力，能分析常规工艺问题，如缺陷初步诊断。
Lv3	75－89	能综合分析、优化工艺，接近单个专家水平，能优化复杂工艺参数。

Lv4	90－100	具备创新能力，表现接近或超越专家团队，能提出创新方案。

7 检验记录要求

7.1 记录内容

检验记录应完整、准确、可追溯，至少应包括以下内容：

a) 模型信息

模型名称、版本号及发布日期；

模型架构类型（如 Transformer、混合架构等）；

训练数据概述（包括专业语料比例等）；

微调或领域适配方法说明（如指令微调、LoRA、知识蒸馏等）。

b) 测试环境

硬件配置（CPU/GPU 型号、内存、存储等）；

操作系统及版本；

测试软件平台；

网络环境说明（如离线/隔离测试）。

c) 测试数据集信息

数据集版本号与发布日期；

题目总数及难度分布（Lv1~Lv4）；

题型比例（全部为 MCQ）；

数据来源类别比例（标准条款、案例、考试、仿真等）。

d) 测试过程信息

测试时间及持续时长；

测试步骤记录（静态问答、场景分析、交互式决策、实时干预等的执行情况）；

异常情况记录（包括中途系统故障、输入数据异常、模型拒答等）。

e) 评分记录

各维度得分（知识面、推理能力、实践能力、扩展指标等）；

单题得分与正确答案对照；

总分及等级判定。

7.2 记录保存与格式

记录应以电子文档形式保存，并定期备份；

建议使用结构化数据格式（如 JSON、CSV）保存评分结果，便于后续统计与分析；

记录保存期限不应少于 3 年，以便复查与认证，并符合数据保护法规。

附录 A

（资料性）

检验记录模板

检验记录模版如表A.1所示。

表A.1 检验记录模版

类别	项目	记录内容要求	数据类型	必填	示例/建议格式
记录基础	记录编号 record_id	唯一可追溯 ID	string	是	IR-2026-0001
模型信息	模型名称 model_name	模型名称	string	是	Weld-LLM
	版本号 model_version	版本号	string	是	v1.2.0
	发布日期 model_release_date	版本发布日期	date	是	2026-01-01
	架构类型 model_arch_type	Transformer/混合等	string/enum	是	Transformer
	训练数据概述 training_data_summary	训练数据总体描述	string	是	通用+焊接专业语料
	专业语料比例 domain_corpus_ratio	专业语料占比	number/%	是	0.35 或 35%
	领域适配方法 adaptation_method	指令微调/LoRA/蒸馏等	string	是	指令微调+LoRA
	适配方法说明 adaptation_detail	关键参数/做法说明	string		LoRA r=16, α=32…
测试环境	硬件配置 hardware_spec	CPU/GPU/内存/存储	string	是	CPU:i9-12900K; GPU:4090; RAM:64GB; SSD:2TB
	操作系统 os_version	OS 及版本	string	是	Ubuntu 22.04
	测试软件平台 test_platform	评测框架/脚本/版本	string	是	EvalRunner v1.2 + Python 3.11
	网络环境 network_env	离线/隔离/联网	string/enum	是	离线/隔离测试
数据集信息	数据集版本 dataset_version	数据集版本号+发布日期	string	是	GPQA-Weld v1.3 (2025-12-20)
	题目总数 total_questions	总题量	int	是	400
	难度分布 difficulty_dist	Lv1~Lv4 分布	JSON 字符串 / string	是	{"Lv1":120,"Lv2":150,"Lv3":90,"Lv4":40}
	题型比例 question_type_ratio	全部为 MCQ	string	是	MCQ=100%
测试过程	测试时间 test_time	开始/结束/时长	string/datetime	是	2026-01-06 09:10–11:05 (115min)
	步骤执行记录 test_steps	静态问答/场景分析/交互决策/实时干预执行情况	string	是	静态问答=完成；场景分析=完成；交互决策=跳过；实时干预=完成

	异常情况 abnormal_log	故障/输入异常/拒答等	string	是	无；或 Q-032 拒答；10:21 脚本重启
评分记录	维度得分 dimension_scores	知识面/推理/实践/扩展等	JSON 字符串 / string	是	{"知识面":78.5,"推理":72.0,"实践":69.0,"扩展":60.0}
	总分 total_score	总分	number	是	71.2
	等级 grade	等级判定	string/enum	是	Lv2