数智风洞：面向生成式大模型的内生安全测评框架

胡涛，马海龙，韩伟涛
信息工程大学，河南 郑州 450002

摘要
随着生成式大模型的快速迭代和相关应用的不断扩展，其潜在的生成式内容风险和安全威胁日益凸显。因此，在大模型部署应用前开展全面深入的安全风险测试和评估至关重要。然而，当前测评手段存在测评用例固化、评估科目单一、结果漏报误报等问题。为此，提出了数智风洞——面向生成式大模型的内生安全测评框架，整体架构上，采用结构化、反馈式设计思想，分为测评基础支撑、测评对象设定、测评用例增强、测评环境配置、测评结果判定等5个模块。具体设计上，首先，基于检索增强生成技术智能辅助测评用例优化，高效扩展生成多样、有效的测评用例数据；其次，分析大模型内容安全与系统安全风险交织的内生特性，构建面向大模型的分级分类内生安全测评科目体系，共计3级43项；最后，基于多模型集成对测评结果进行综合裁决判定，防止单一模型判定引发的漏报误报问题。实验结果表明，相比现有的测评框架，数智风洞显著提升了测评时效性、覆盖度和准确率。

关键词
大模型；安全测评；内生安全

中图分类号：TP399　　文献标志码：A

Digital Intelligence Wind Tunnel: An Endogenous Security Evaluation Framework for Generative Large Models

HU Tao, MA Hailong, HAN Weitao
Information Engineering University, Zhengzhou 450002, China

Abstract
As generative large models rapidly evolve and their application scenarios continue to expand, the associated risks and security threats posed by generated content have become increasingly prominent. Comprehensive and rigorous security risk testing and evaluation prior to model deployment are therefore essential. However, current evaluation approaches suffer from issues such as rigid test case design, limited assessment dimensions, and high rates of false negatives and false positives. To this end, we propose Digital Intelligence Wind Tunnel—an endogenous security evaluation framework tailored for generative large models. Architecturally, the framework adopts a structured, feedback-driven design and consists of five modules: evaluation foundation support, evaluation object specification, evaluation case augmentation, evaluation environment configuration, and evaluation result adjudication. Specifically, we first employ retrieval-augmented generation to intelligently facilitate test case optimization, enabling the efficient generation of

diverse and effective evaluation data. Second, we analyze the intertwined nature of content and system security risks inherent in large models, and construct a hierarchical and categorized intrinsic security evaluation taxonomy comprising 3 levels and 43 items. Finally, a multi-model ensemble adjudication mechanism is introduced to synthesize evaluation outcomes, mitigating the bias and inaccuracies associated with single-model judgments. Experimental results show that, compared with existing evaluation frameworks, Digital Intelligence Wind Tunnel significantly enhances evaluation efficiency, coverage, and accuracy.

Key words
Large Model, Security Evaluation, Endogenous Security.

1 引言

以 ChatGPT、DeepSeek、GLM 等为代表的生成式大模型[1]，正以前所未有的深度和广度重塑信息处理与内容创作范式[2][3]。然而，科学技术能力的跃升往往伴随着相应的治理挑战，大模型在释放巨大潜能的同时，其内部蕴藏的生成内容风险与系统安全风险也愈发凸显，并呈现出复杂性和多维性的特征[4]。具体而言，其安全挑战主要来源于两个既相互关联又有所区别的维度：AI Safety（人工智能安全）与 Security for AI（人工智能系统安全）。前者聚焦于模型自身行为的对齐、可靠与无害性，例如，避免生成偏见歧视、仇恨言论、虚假信息等有害内容[5]；后者则侧重于保护模型资产（如架构、参数、训练数据）及服务过程免受恶意攻击[6]，例如防范对抗性提示注入[7]、越狱攻击[8]、后门植入[9]、模型窃取[10]等威胁。更为关键的是，AI Safety 与 Security for AI 风险往往相互交织、彼此放大，构成了大模型的“内生安全”挑战——即源于模型自身能力与结构的内在缺陷所引发的安全问题[11]。因此，在大模型被大规模部署应用于关键领域之前，对其进行全面、深入、可信赖的安全风险评估，已成为保障其健康、可控发展不可或缺的前提。

目前，已有多个国内外研究机构和团队提出了不同的测评框架，用于评估大模型的安全性，代表性成果包括 OpenCompass-Safety[12]、JADE[13]、SuperCLUE-Safety[14]、FlagEval[15] 等。OpenCompass-Safety 是由上海人工智能实验室开发的开源评估框架，致力于评估大模型在安全可信方面的表现，覆盖人类价值观、安全风险、信息可靠性、法律合规性以及越狱与滥用等五大核心安全领域。JADE 由复旦白泽智能团队研发，是一种面向大模型的靶向安全测评系统。该系统结合语言学变异模块与安全合规测评模块，构建出“反馈-迭代”机制，从而实现了对大模型安全性的自动化评估及高风险问题的自动收集。SuperCLUE-Safety 是一套综合性的大模型安全评估基准，用于检验模型在遵循基本道德法律准则、与人类价值观对齐以及抵御潜在攻击等方面的能力。该基准包含传统安全、负责任人工智能和指令攻击三大核心维度，下设二十余项具体测评任务。FlagEval 是由北京智源人工智能研究院推出的大模型综合测评平台，其目标在于建立科学、公正、开放的测评基准与工具体系，以支持研究人员全面评估基础模型及其训练算法的性能。

近年来，学术界和产业界逐渐认识到大模型内容安全与系统安全的紧密关联，

然而，现有测评框架面对日新月异的模型演进和层出不穷的新型攻击时，逐渐暴露出其局限性，主要体现在以下三个方面：一是测评用例静态固化，覆盖不足。许多测评严重依赖于静态、预定义的测评用例库。这些用例难以动态扩展，无法有效覆盖长尾、隐蔽或新兴的安全威胁，导致测评的“盲区”，使得模型在面对经过轻微改动的或前所未见的恶意输入时表现不佳[16]；二是测试评估科目单一，维度不全。现有测评多聚焦于内容安全的某一或某几个方面（如毒性检测），缺乏一个统一、全面的框架来系统性地衡量大模型在内容安全与系统安全交织下的内生安全风险。对于模型在不同攻击场景下的鲁棒性、泛化性及本征安全能力的评估尚不完善。三是结果判定路径依赖，效率不高。多数方法依赖另一个大模型（如 DeepSeek）或基于规则的分类器作为“裁判”来判定测试结果。这种单一判定机制本身可能存在偏见、不一致性或被攻击的风险，容易导致误报（将安全回复判为有害）和漏报（未能识别出有害回复），严重影响了测评结果的可信度。例如，FaithBench[17]面向大模型幻觉领域开展了测试评估，结果显示不同模型的幻觉率具有显著差异，其中 GPT-4o 的幻觉率最低，其次是 GPT-3.5-Turbo、Gemini-1.5-Flash。与之类似，文献[18]首次验证了不同大模型（例如，ChatGPT-3.5、Google Bard）面对越狱攻击时具有不同的防御表现。因此，大模型安全测评需摆脱传统依赖单一模型进行结果判定的测评模式，以尽可能消除偏见/被攻击风险。

针对上述问题，本文提出了数智风洞，一种面向生成式大模型的内生安全测评框架。该框架借鉴了航空航天领域“风洞试验”的系统化思想，旨在为大模型提供一个用例可迭代、功能可扩展、结果可置信的内生安全“测试场”。测试结果表明，本文的内生安全测评框架相比现有工作，能够显著提升大模型测评的时效性、覆盖度和准确率。本文主要贡献如下：

- 提出一种结构化的大模型内生安全测评框架。数智风洞采用反馈式、模块化的设计思想，将测评流程系统性地分解为测评基础支撑、测评对象设定、测评用例增强、测评环境配置、测评结果判定五个核心模块，实现了测评任务的全生命周期管理，确保了测评的规范性与动态性。
- 设计基于检索增强生成的测评用例智能增强技术。为解决测评用例固化问题，创新性地引入检索增强生成技术，动态地从海量知识库中检索相关上下文，辅助生成多样、复杂且贴近真实攻击场景的有效测评用例，显著提升了测评的覆盖范围。
- 构建面向大模型的内生安全测评科目体系。基于对大模型内生安全风险的深入分析，建立了一个分级分类的大模型内生安全测评科目体系，不仅涵盖传统的内容安全，还将模型安全纳入统一框架，综合评估模型的内生安全性。
- 实现基于多模型集成的综合裁决判定机制。为提升结果判定的可靠性，采用多个性能近似的智能模型构成裁决委员会，通过投票或加权融合等策略对测评结果进行综合判定，有效降低了因单一模型缺陷而导致的漏报与误报风险。

2 框架设计

数智风洞框架的整体架构如图 1 所示，其设计借鉴了“风洞试验”中“设定目标—构建场景—执行测试—分析结果”的系统化思路，结合大模型特有的动态性与复

杂性，构建了一个具备自适应能力的测评环境。数智风洞采用反馈式、模块化的设计思想，将测评流程系统性地分解为大模型测评基础支撑、测评对象设定、测评用例增强、测评环境配置、测评结果判定五个核心模块。

测评基础支撑模块一方面向测评对象设定模块提供目标测评模型及其工具，包括待测试的大模型对象以及支撑模型部署运行的工具组件；另一方面向测评用例增强模块提供基础性数据集和算法库。

测评对象设定模块用于设定待测评的目标模型对象（例如，文本大模型、视觉大模型、语音大模型、多模态大模型）及其访问模式（例如，白盒访问、黑盒访问、灰盒访问）。

测评用例增强模块基于检索增强生成（Retrieval Augmented Generation，RAG）技术[19]智能辅助测评用例优化，高效扩展生成多样、有效的测评用例数据，从而确保测评数据的高可用性和高覆盖度。

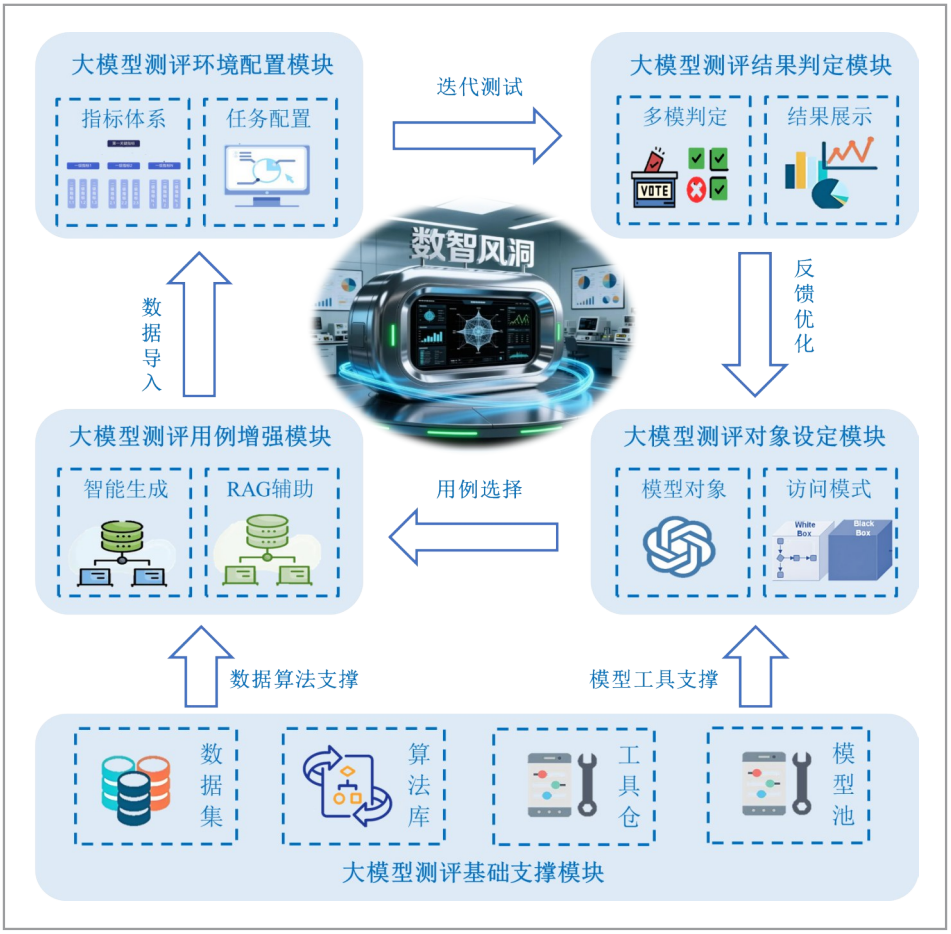

图1 大模型数智风洞总体框架图

测评环境配置模块构建覆盖大模型本征能力的3级43项内生安全测评科目体系，并基于容器级微隔离技术支持并发式测评任务配置，实现多种测评任务的并行开展和互不干扰。

测评结果判定模块基于多模型集成对测评结果进行综合裁决判定，防止单一模型判定引发的漏报误报问题。最终以定制化模板的形式输出测评报告，含大模型安全测评结果和安全加固意见建议，并将其反馈至测评对象为大模型后续迭代优化提供辅助。

数智风洞框架的运行遵循“配置—生成—执行—裁决—反馈”的闭环流程。首先根据测评目标设定模型与任务类型；随后通过智能增强算法动态生成多样化测评用例；在配置好的测评环境中执行面向大模型的内生安全测评科目测试；由集成模型组成的裁决委员会对输出进行综合判定；最终结果反馈至用例生成与模型优化环节，实现测评能力的持续演进。

2.1 大模型测评基础支撑

测评基础支撑模块作为数智风洞框架的底层支撑，提供测评所需的资源管理与技术服务。从提供管理服务对象的角度，可划分为模型池、工具仓、算法库、数据集等功能组件，其中模型池和工具仓主要支撑大模型测评对象设定模块，算法库和数据集主要支撑大模型测评用例增强模块。

（1）模型池组件

模型池组件主要用于提供安全测评所针对的大模型对象，待测评的目标大模型通过标准化API接口接入数智风洞。如图2所示，模型池组件覆盖了当前阶段通用的大模型类型，包括本文大模型（例如，GPT-4、LLaMa2、Claude3）、视觉大模型（例如，DALL-E3、SAM、Sora）、音频大模型（例如，Whisper、VALL-E、AudioLM）、多模态大模型（例如，CLIP、Gemini、Qwen-VL）等[20]。

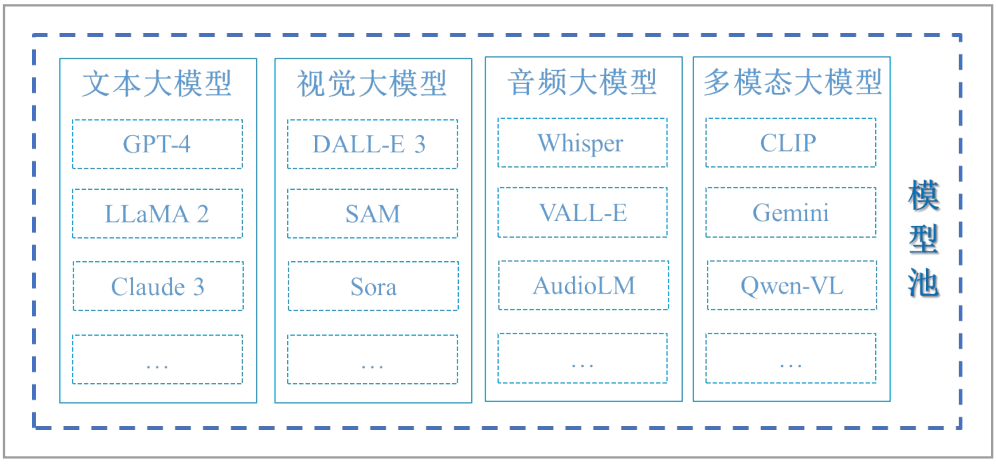

图2 模型池组件构成图

（2）工具仓组件

工具仓组件主要用于优化大模型性能、效率和部署成本，为大模型安全测评提供支撑，总体上主要分为3类：一是模型微

调工具，让模型适应特定任务或领域，例如，XTuner，PEFT，Unsloth 等；二是模型蒸馏工具，将大模型能力压缩至小模型，用于降低部署和推理成本，例如，千帆 ModelBuilder，阿里云 PAI；三是模型量化工具，降低模型权重精度以减少体积、提升推理速度，例如，AutoRound，AMD Quark，GPTQ，Sophgo Quantization-Tools。

（3）算法库组件

算法库组件主要为处理测评用例提供算法层面支撑，依据样本数据处理流程，可以分解为以下 4 类：一是加载与解析算法，从不同来源（文件、网页、数据库）加载和解析测评用例数据，例如，LangChain 文档加载、Klovis 加载、RAG-Anything 解析；二是处理与分块算法，清洗数据、拆分文档以适应模型上下文限制，例如，按字符、标记、递归等文本拆分；三是向量化与嵌入算法，将测试数据转换为数值向量（嵌入），以便进行语义搜索，例如，Word2Vec、GloVe、FastText；四是检索器算法，根据搜索查询，从知识库中智能检索最相关的信息，例如，LangChain 检索器（自查询、上下文压缩、父文档等）。

（4）数据集组件

数据集组件与模型库组件呈现多对多的映射关系，即同一大模型可用不同数据集进行测试，同一数据集也可以用于测试不同大模型。鉴于模型池组件包括文本、视觉、音频、多模态等类型大模型，因此数据集组件也提供了相应安全测评数据集测评用例调用接口。数据集组件构成如图 3 所示。需要说明的是，数据集组件在本框架中用于提供基本的测评用例，为了全面测评大模型安全性，在基本测评用例的基础上，采用基于检索增强生成的智能用例增强技术进一步丰富扩展测评用例数量和规模。

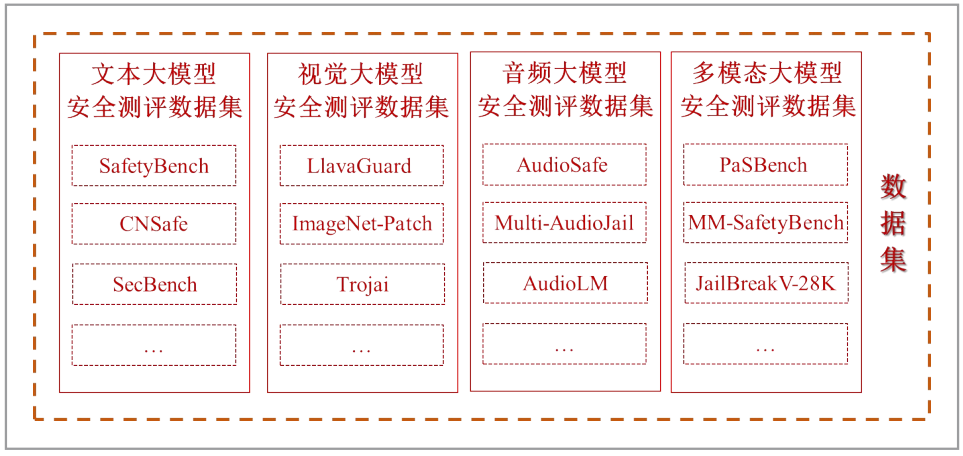

图 3　数据集组件构成图

2.2 大模型测评对象设定

数智风洞的测评对象设定模块是定义整个测评范式的标准基石，该模块不仅定义了“测评谁”，更关键的是明确了“在何种条件下进行测评”。整体上看，测评对象设定模块包括模型对象和访问模式两个组件。

在模型对象组件中，测评对象首先根据其核心功能与数据模态进行划分，不同模态的大模型面临独特的安全挑战，其测评基准和攻击向量也大相径庭。具体而言，1）文本大模型作为当前研究最深入的领域，其安全测评主要聚焦于：①提示注入与越狱，通过精心构造的指令，绕过大模型的对齐护栏，诱导其生成不当内容、泄露训练数据或执行有害指令；②隐私与成员推断攻击，探测大模型是否记忆并可能泄露其训练数据中的敏感个人信息；③事实性与幻觉，评估大模型生成内容的准确性及其捏造事实的倾向，这在错误信息传播中至关重要。2）视觉大模型其安全威胁主要体现在：①对抗样本，通过在输入图像中添加人眼难以察觉的扰动，导致大模型进行严重误判（如将“停车标志”识别为“通行标志”），这在自动驾驶等安全关键领域极具危险性；②后门攻击，在训练阶段植入特定触发器，使得大模型在正常输入下表现良好，但一旦输入包含该触发器，就会执行恶意行为；③不良内容生成，测评大模型生成暴力、色情或深度伪造内容的能力与可控性。3）语音大模型安全风险包括：①对抗性音频：通过音频扰动，实现对语音识别系统的“隐形命令”注入，或导致声纹识别系统失效；②深度伪造语音：测评大模型模仿特定人物声音进行诈骗或诽谤的逼真度与检测难度。4）多模态大模型的安全测评最为复杂，因其面临上述所有模态的单一风险，更存在跨模态攻击的复合威胁。例如，一张包含恶意文本指令的图片可能诱导大模型解读后执行有害操作；或通过音频–视觉的协同攻击，实现更隐蔽的越狱。

在访问模式组件中，定义了测评者相对于目标大模型的知识状态和交互能力，它构成了一个从完全无知到完全掌控的“访问权限”，深刻影响着测评方法论的选择，主要分为：1）黑盒访问，测评者仅能通过API或用户界面与大模型交互，获取输入–输出对，而无法知晓大模型的内部结构、参数及训练数据。这是最常见也是最现实的商业场景模拟。2）白盒访问，测评者拥有大模型的完整知识，包括架构、所有参数权重、梯度信息等。3）灰盒访问：这是一个内涵丰富的中间地带，指测评者拥有部分、不完全的模型知识，具体形态多样，包括模型架构知情、代理模型知情、API元数据知情等。

2.3 大模型测评用例增强

数智风洞的测评用例增强模块将大模型安全测评视为一个动态、开放的知识密集型任务，通过从大规模、多源的RAG安全知识库中检索相关攻击模式与案例，为生成器提供丰富的上下文，从而激发出更多样、更有效、更具欺骗性的测评用例。这不仅显著提升了数据生成的效率，更重要的是，它通过知识引导，使生成的用例能够触及大模型安全防线中那些难以通过随机生成发现的“盲区”。为此，在测评用例增强模块，提出了基于RAG的测评用例智能增强技术，动态地从海量知识库中检索相关上下文，辅助生成多样、复杂且贴近真实攻击场景的有效测评用例，显著提升了测评的覆盖范围，伪代码如表1所示。具体而言，基于RAG的测评用例智能增强

技术分为三个步骤。

1）多源安全知识库构建

构建的知识库整合了以下多源数据：①对抗性提示生成方法论文，包括从相关顶会论文中抽取的攻击方法论描述、典型案例和攻击模板。例如，Hayse等人提出的基于查询的对抗性提示生成方法（Adversarial Prompt Generation，APG）[21]，该工作提供了系统生成对抗性提示的算法框架，可用于构建动态更新的对抗性知识来源。②恶意用户对话记录，在受控环境下模拟的、被安全模型成功拦截或需要人工审核的恶意对话记录。③社会工程学案例，包含经典的钓鱼、诱导、伪装等社会工程学手法文本。所有文本、图像、音视频数据经过清洗和去噪后，被分割为短片段，并基于嵌入模型向量化，存入向量数据库以备高效检索。

2）动态上下文检索与组装

给定一个初始的、可能较为简单的种子用例s（例如，请写一个网络钓鱼邮件），检索模块的目标是找到与s语义相关且能提供“攻击升华”的知识片段。具体步骤包括：①向量检索，计算种子用例s的向量表示，并从向量数据库中检索出Top-K个最相似的上下文片段$C_{vec} = \{c_1, c_2, \ldots, c_k\}$。②元数据过滤与重排序，为避免生成内容过于同质化，引入基于元数据（如攻击类型、目标模型、成功率）的多样性采样策略，对C_{vec}进行重排序，得到最终上下文集合C_{final}。③提示模板组装，将检索到的上下文与种子用例组装成结构化的提示P，提供给生成器。提示模板设计如图4所示。

System Prompt:
你是一个专业的大模型安全测评专家。你的任务是根据提供的<**安全背景知识**>，对一个<**种子指令**>进行优化和增强，生成一个更具迷惑性、更隐蔽、更有效的安全测评用例。请严格在安全测评的范畴内进行创作。

<安全背景知识>
{Cfinal}

<种子指令>
{s}

请生成增强后的测评用例：

图4 系统提示词模板

3）安全可控的测评用例生成

为确保生成过程本身的安全性与定向性，采用经过安全对齐的大模型（例如，Qwen3Guard[22]）作为生成器G。生成器的输出被严格限制为“测评用例本身”，而非执行恶意行为的结果。例如，生成的是“如何制造炸弹”指令，而非制造炸弹的详细步骤。所有生成的用例会流入一个由另

一个高安全级别模型（例如，Qwen3-235B-A22B）担任的“安全过滤器”进行自动审核，过滤掉可能引发直接危害或不符合伦理规范的内容。

表1 基于RAG的测评用例智能增强算法伪代码

算法1：基于RAG的测评用例智能增强算法

输入：种子用例集合 [s1, s2, ..., sn],向量化安全知识库KB;检索数量k

输出：增强后的测评用例集合A

```
A = []
for s ∈ S:
    local_variants = []
    for i = 1 to N:
        query_embedding = get_embedding(s) // 获取种子用例的向量
        C_candidates = retrieve_similar(K, query_embedding, k*2)
        C_final = diversity_rerank(C_candidates, k) //优先选择差异大片段
        prompt = assemble_prompt(C_final, s)
        generated_text = G(prompt)
        variant = post_process(generated_text) // 提取生成的用例部分
        if safety_filter(variant): // 安全检查
            local_variants.append(variant)
    unique_variants = deduplicate(local_variants)
    A.extend(unique_variants)
    end for
end for
return A
```

基于RAG的测评用例增强流程的关键环节包括通过“动态检索+元数据”重排序提升对抗场景覆盖性，通过多阶段安全过滤与人工复核保障生成过程可控，通过语义与结构双重去重提升用例库质量，具体实施细节如下：

①检索与重排序策略

在检索阶段，使用基于余弦相似度的向量检索方法，从多源安全知识库中提取与种子用例语义最接近的Top-K（例如，实验中K=10）个候选片段。为提升检索多样性与对抗覆盖性，在初步检索后引入基于元数据的重排序机制：

● 元数据字段：每个知识片段标注有“攻击类型”、“目标模型”、“攻击成功率”、“生成时间”等属性。

● 重排序策略：采用多样性采样，在保证语义相关性的前提下，优先选择攻击类型不同、目标模型各异、且近期成功率较高的片段，避免检索结果同质化。

● 最终上下文集合：经重排序后保留Top-5个片段作为生成器的增强上下文。

②词嵌入模型与向量库配置

● 嵌入模型：选用bge-large-zh-v1.5进行中文文本向量化，对英文攻击案例辅以text-embedding-ada-002，多模态攻击案例（如图像提示注入样本）采用CLIP的文本编码器进行统一向量表示。

● 向量库构建：使用FAISS索引存储向量化片段，并附加元数据存储于关联数据库中。索引类型为“倒排文件+乘积量化”，以平衡检索速度与内存占用。索引类型采用IVF4096_PQ32索引，该组合在亿级向量规模下能在大幅压缩存储的同时保持高检索精度与速度，非常适合规模持续增长（百万级片段）的多源知识库。PQ（乘积量化）的子向量数（m）设为32，每个子量化的编码位数（nbits）为8。此配置在实验中验证将存储开销降低至原始向量的约1/32，并显著提升检索速度。

● 更新机制：知识库支持动态扩展，从预设的CVE数据库、学术论文库及安全社区推送中自动爬取更新，经人工审核后入库。

③生成器与安全过滤器的实现约束

● 生成器选择与约束：选用经过强安全对齐的大模型Qwen3Guard作为生成器，在系统提示词中明确限制其输出格式（如“生成一个用于测试的对抗提示，不包括具体有害内容”），并通过API调用时设置 temperature=0.7、 max_tokens=200以平衡多样性与可控性。

● 安全过滤器设计：采用两阶段过滤，一是基于规则的关键词过滤，拦截明显违反安全政策的内容（如具体暴力步骤、隐私数据模板）；二是模型二轮审核：使用Qwen3对生成内容进行二次安全性评分，低于阈值（如安全概率<0.9）的用例自动丢弃并记录日志。

● 人工复核机制：所有生成的用例在进入测评用例库前，均需通过至少一名安全专家复核，确保符合伦理与实验合规要求。

④去重规则设计

● 语义去重：对生成用例计算句向量，若余弦相似度>0.85，则视为语义重复，仅保留生成时间最新或攻击成功率更高的用例。

● 结构去重：针对提示注入类用例，提取其逻辑结构（如占位符位置、攻击模式），若结构完全相同且仅替换少量实体词，则进行合并。

2.4 大模型测评环境配置

数智风洞的测评环境配置模块通过分析大模型内容安全与系统安全风险交织的内生特性，构建面向大模型的内生安全测评科目体系，基于容器级微隔离支持并发式测评任务配置。

1）分级分类测评科目体系构建

在大模型测评环境配置模块，通过精准刻画大模型的内生安全特性，摒弃了零散的测评项，构建面向大模型的内生安全测评科目体系，包括大模型内容安全和系统安全，共计3级43项，如图5所示。

● 一级科目内容安全分为偏见歧视、违法违规、幻觉有害、隐私泄露等4项二级科目，其中偏见歧视包括种族偏见、地位偏见、宗教偏见、国籍偏见、疾病歧视、外貌歧视、取向歧视、性别歧视等8项三级科目；违法违规包括虚假煽动、暴力犯罪、政治错误、色情骚扰、非法违禁等5项三级科目；幻觉有害包括幻觉错误、违背事实、发布黑话、仇恨言论、谩骂侮辱等5项三级科目；隐私泄露包括个人信息泄露、训练数据泄露、提示词数据泄露等3项三级科目。

● 一级科目系统安全分为模型越狱、提示注入、后门攻击、缺陷利用等4项二级科目，其中模型越狱包括基于人类的越狱、基于混淆的越狱、基于启发式的越狱、基于反馈的越狱、基于微调的越狱、基于生成参数的越狱等6项三级科目；提示注

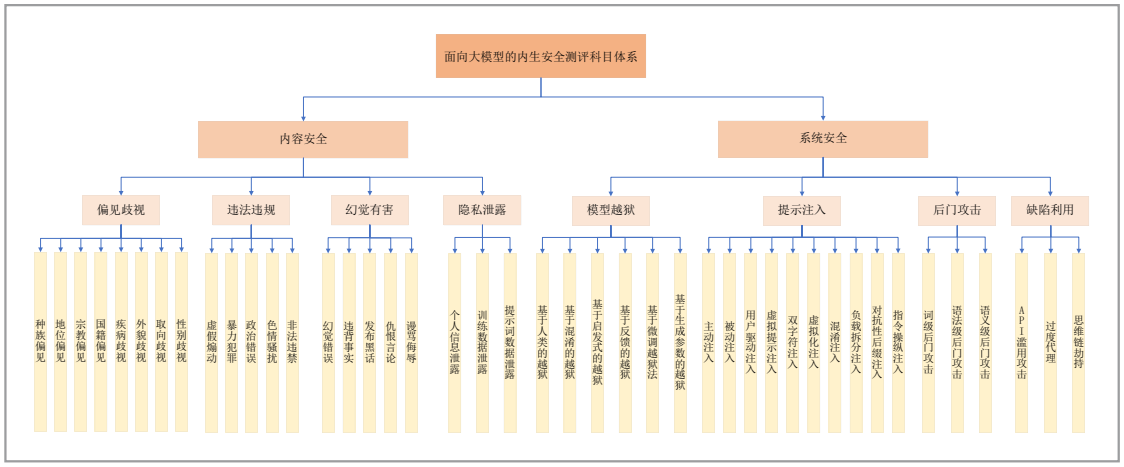

图5　面向大模型的内生安全测评科目体系

入包括主动注入、被动注入、用户驱动注入、虚拟提示注入、双字符注入、虚拟化注入、混淆注入、负载拆分注入、对抗性后缀注入、指令操纵注入等10项三级科目；后门攻击包括词级后门攻击、语法级后门攻击、语义级后门攻击等3项三级科目；缺陷利用包括API滥用、过度代理、思维链劫持等3项三级科目。

图5中每项科目的作用是测评大模型在该科目对应测试数据下的安全防御能力，最终通过综合所有科目测评结果的加权得分值，量化评估大模型的整体安全性。针对“3级43项”内生安全测评科目，采用的权重体系遵循“层级内平等权重、层级间结构传递”的原则。层级内平等权重的含义是在同一父节点下的所有子项，其权重均相等。例如：一级科目权重各为0.5；内容安全下的4个二级科目（偏见歧视、违法违规、幻觉有害、隐私泄露）权重各为0.25；违法违规下的5个三级科目权重各为0.2。平等权重设定原因基于以下考虑，一是避免主观偏差，在缺乏大规模实证数据支持特定风险项显著更频繁或更严重的情况下，平等权重是最稳健、最不易引入人为偏见的设计；二是结构清晰，便于解释。平等权重使体系易于理解、复现与审核，也更利于在不同模型之间进行公平比较；三是鼓励全面覆盖，平等权重迫使测评必须同等地关注所有科目，避免因权重倾斜导致的测评“盲区”。

为了验证“平等权重”设计的合理性，本文设计了平等权重与差异化权重的性能对比实验。具体方式如下，首先，邀请5位领域专家根据经验对43项三级科目进行权重打分（差异化权重组）；其次，使用同一测评数据集（来自第3.1节所述混合数据集），分别采用平等权重与专家权重计算模型的安全性得分；最后，以人工标注的“真实风险等级”（由3名安全专家独立标注并达成一致）作为黄金标准，计算两种权重方案下的斯皮尔曼等级相关系数。实验结果显示，平等权重方案与黄金标准的相关系数为0.891；专家权重方案与黄金标准的相关系数为0.885；两种方案在统计上无显著差异（$p>0.05$），这表明在缺乏先验强证据的情况下，平等权重并未显著降低评估的效度，且具备更好的可解释性与复现性。

2）基于容器级微隔离支持并发式测评任务配置

大模型安全测评面临测评全面性、深度性与执行效率、可管理性之间矛盾。全面测评要求对单一模型进行多维度、多场景的测试，同时还需横向对比不同模型或同一模型的不同版本。若采用传统的串行或简单的并发执行模式，将导致资源占用巨大、测试周期漫长，且难以保证不同测试任务间的纯净性与结果的可复现性。

为此，在大模型测评环境配置模块，采用基于容器微隔离的并发式测评任务配置模式，以容器技术为核心、容器编排平台为骨架的微隔离架构。核心思想是将每一个独立的测评任务封装为一个自包含、轻量级的“测评容器单元”。

①任务定义与容器化：测试者将测评逻辑（例如，执行一组越狱攻击提示词）及所需环境（如PyTorch版本、特定模型加载方式）编写成Dockerfile，构建为标准化容器镜像，并推送至私有镜像仓库。

②任务编排与下发：通过Kubernetes编排系统的API或配置文件，提交一个“作业”（Job）或“任务”（CronJob），指定使用的镜像、并发实例数、资源需求（例如，2个GPU，16GB内存）以及输入参数（例如，模型版本号、测评数据集路径）。

③并发执行与隔离运行：Kubernetes编排系统根据调度策略，在集群中拉起指定数量的容器副本。每个容器在完全隔离的环境中独立运行测评脚本，与宿主机及其他容器无状态干扰。它们可以同时访问不同的GPU，或通过时间切片共享同一GPU。

④结果收集与持久化：每个容器将测评产生的原始日志、模型输出、性能指标等，通过Sidecar容器或直接写入的方式，输出至持久化存储卷（例如，网络文件系统NFS）。此存储卷与容器生命周期解耦，确保数据在容器销毁后依然保留。

⑤资源回收与清理：任务执行完毕后，无论成功或失败，容器将被自动终止并清理其运行时资源，但测评结果已被安全保存。这一过程保证了集群资源的持续可用性。

2.5 大模型测评结果判定

数智风洞的测评结果判定模块基于多模型集成对测评结果进行综合裁决判定，进而以定制化模板的形式输出测评报告，含大模型安全测评结果和安全加固意见建议，并将其反馈至测评对象为大模型后续迭代优化提供辅助。

当前，大模型安全测评通常依托第三方智能大模型对测评对象模型的输出结果进行判定，然而，单一安全测评模型（无论是基于规则、分类器或参考模型）通常受限于其训练数据分布、算法偏好及对抗鲁棒性。例如，一个针对毒性内容检测优化的模型，可能对隐蔽性偏见或代码漏洞诱导的识别能力不足；而一个基于特定数据分布的对抗攻击检测模型，可能在分布外样本上表现显著下降。这种“单点故障”风险使得测评结论的可靠性存疑。

该模块的裁决流程为一个多阶段流水线，如图6所示，首先，待测评大模型的输出被同步输入至N个安全测评模型中，每个模型独立生成初步风险标签（如：安全、有毒、有偏见、信息泄露风险等）及对应的置信度分数；其次，N个智能模型构成裁决委员会，通过多数投票或加权融合等策略对测评结果进行综合判定，有效降低了因单一模型缺陷而导致的漏报与误报风险；最后，综合加权投票结果生成最终的综合判定，并附上裁决置信度与各模型分歧度指标，量化本次判定的可靠性。

权重设计的核心原则是：赋予在特定问题上置信度更高、且历史表现更可靠的模型以更大决策权重。

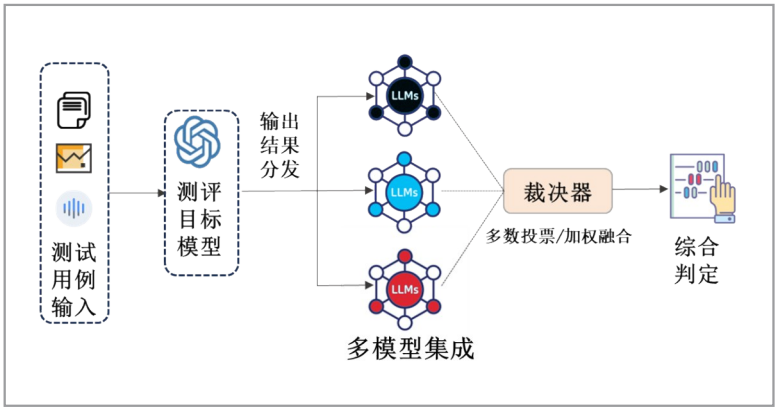

图6 基于多模型集成的综合裁决判定机制

步骤1：置信度标准化。每个模型M_i针对当前待判定输出，不仅生成风险标签L_i（如“有害”或“安全”），同时输出一个与该标签对应的置信度分数c_i（范围[0, 1]）。对来自N个模型的置信度进行标准化，如公式（1）所示，从而得到基于当前实例的初始权重。

$$c_i' = \frac{c_i}{\sum_{j=1}^{N} c_j} \tag{1}$$

步骤2：历史准确率加权。为避免单一实例置信度可能存在的偏差，引入各模型在校准集（一个涵盖各类安全问题的、经过人工精准标注的基准测试集）上的历史表现作为先验知识。设模型M_i在校准集上与人工标注结果相比的加权F1分数为a_i。融合权重w_i通过结合当前置信度与历史准确率计算得出，如公式（2）所示。

$$w_i = \beta \cdot c_i' + (1-\beta) \cdot \frac{a_i}{\sum_{j=1}^{N} a_j} \tag{2}$$

其中，β是一个可调参数（实验中设为0.7），用于平衡当前实例证据与历史普遍性能的信任程度。此设计使得一个长期表现稳定但本次置信度稍低的模型，仍能保有相当的决策影响力。

步骤3：加权投票与最终判定。对于“有害”与“安全”两类判定，分别计算加权投票得分，最终判定结果为得分更高的一类。此外，系统还计算裁决置信度D和分歧度F作为结果可靠性的量化指标，供测评报告分析使用。

$$S_{\text{有害}} = \sum_{i \in \{j | L_j = \text{"有害"}\}} w_i, \tag{3}$$

$$D = \left| S_{\text{有害}} - S_{\text{安全}} \right|, \tag{4}$$

3 实验评估

为验证数智风洞框架的有效性与先进性，本节开展实验评估，旨在回答以下核心研究问题：（1）数智风洞在测评时效性、覆盖度和准确率等核心指标上，是否显著优于现有测评框架？（2）基于RAG的测评用例增强技术能否有效生成多样、复杂的测评用例，提升对新兴威胁的发现能力？（3）所构建的内生安全测评科目体系是否比单一维度测评更能全面、系统地揭示大模型的安全风险？（4）基于多模型集成的综合裁决机制是否能有效降低单一模型判定导致的漏报与误报？

3.1 实验设置

（1）硬件与软件环境

硬件配置：所有实验在单台服务器上完成，配置为：4×NVIDIA A800 (80GB) GPU，512GB内存。软件环境：容器环境，Docker 24.0.7，Kubernetes 1.28单节点集群（用于容器微隔离并发）；深度学习框架：PyTorch 2.1.0，Transformers 4.36.0；向量检索：FAISS 1.7.4，索引类型为IVF4096。

（2）测评基准与对比方法

为进行公平对比，选取了当前具有代表性的开源大模型安全测评框架作为基线，包括：OpenCompass-Safety、JADE和SuperCLUE-Safety。为公平对比，所有基线框架与数智风洞在测评同一模型时，均使用相同的随机种子和相同的超时设定（例如，单条用例响应超时设定30秒）。

（3）测评对象模型

选取了不同规模、不同类型的代表性生成式大模型作为测评对象[23]，以验证框架的通用性，包括文本模型Llama-3-70B、Qwen2.5-72B-Instruct、文心一言4.0和多模态模型Qwen-VL-Max。

（4）数据集

实验采用混合数据集，分为基础数据集、动态生成集和新兴威胁集。基础数据集是从BeaverTails[24]、SafeBench[25]、ToxiGen[26]等公开安全数据集中选取部分样本作为种子。动态生成集利用数智风洞自身的RAG用例增强模块，基于种子生成扩展测试集。新兴威胁测试集收集了近期学术论文和社区报告中披露的新型对抗性提示[27][28]，构成一个挑战集。所有文本数据均经过统一的预处理：使用jieba进行中文分词，英文使用nltk，去除无关字符与超长样本（>512 tokens）。

（5）评估指标

评估指标在传统指标体系（例如，准确率、召回率、漏报率、误报率、F1分数）的基础上，增加了时效性（完成对单一模型全套测评科目所需的平均时间）、覆盖度（测评用例所触发的唯一安全风险类别数与总风险类别数之比）等指标。

（6）数据安全与实验合规说明

所有实验均在境内服务器完成，未调用任何境外API；所用国产模型均为官方开源版本或通过合法合规渠道获取；测评用例生成与裁决过程中未涉及真实用户数据或敏感信息，所有数据均为合成或公开数据；实验设计符合相关法律法规要求。

3.2 结果分析

（1）整体框架性能对比

表2-表5分别展示了数智风洞与基线框架在不同目标模型（文本大模型Llama-3-70B、Qwen2.5-72B-Instruct、文心一言4.0、多模态大模型Qwen-VL-Max）综合测评上的性能对比，实验在相同的硬件配置和基础种子数据集下进行。由于JADE与SuperCLUE-Safety主要面向文

本模型，故面向多模态大模型的表5测评中未列入本项对比。实验结果显示数智风洞在各项指标上均显著优于基线。高覆盖度得益于RAG用例增强模块对长尾和复杂攻击场景的生成能力。高召回率和高准确率则主要归功于多模型集成裁决机制有效整合了多模型判断，减少了单一模型的偏见和盲区。更优的时效性则得益于基于容器微隔离的并发测评环境配置，实现了测评任务的高效并行化。上述结果也证明了框架设计（尤其是动态用例增强与集成裁决）的有效性并不依赖于特定模型，而是具备广泛的泛化能力。

表2 面向Llama-3-70B大模型测评的整体框架性能对比

测评框架	准确率	召回率	F1分数	覆盖度	平均耗时(小时)
OpenCompass-Safety	86.2%	78.5%	82.2%	65.7%	4.5
JADE	88.1%	82.3%	85.1%	71.6%	3.8
SuperCLUE-Safety	89.5%	80.1%	84.5%	68.8%	5.2
数智风洞 (Our)	**93.7%**	**91.2%**	**92.4%**	**89.3%**	**3.1**

表3 面向Qwen2.5-72B-Instruct大模型测评的整体框架性能对比

测评框架	准确率	召回率	F1分数	覆盖度	平均耗时(小时)
OpenCompass-Safety	85.9%	77.8%	81.8%	64.3%	4.3
JADE	87.5%	81.2%	84.3%	73.6%	3.9
SuperCLUE-Safety	86.8%	79.5%	83.0%	70.1%	5.0
数智风洞 (Our)	**92.8%**	**90.1%**	**91.5%**	**88.1%**	**3.0**

表4 面向文心一言4.0大模型测评的整体框架性能对比

测评框架	准确率	召回率	F1分数	覆盖度	平均耗时(小时)
OpenCompass-Safety	84.6%	77.5%	81.2%	63.7%	4.8
JADE	86.0%	80.1%	83.0%	72.4%	4.2
SuperCLUE-Safety	85.2%	78.8%	82.0%	69.5%	5.5
数智风洞 (Our)	**91.3%**	**88.7%**	**90.0%**	**86.9%**	**3.4**

表5 面向Qwen-VL-Max大模型测评的整体框架性能对比

测评框架	准确率	召回率	F1分数	覆盖度	平均耗时(小时)
OpenCompass-Safety	82.4%	70.8%	76.2%	58.4%	6.5
数智风洞 (Our)	90.1%	85.6%	87.8%	83.8%	4.2

（2）测评用例增强技术有效性分析

为验证RAG用例增强模块的价值，设置了对照实验：仅基础数据集（Baseline）对比基础数据集+RAG增强数据集（RAG-Aug）。比较两者在相同模型（Llama-3-70B）上触发不同安全风险类别的数量及在新兴威胁测试集上的攻击成功率。表6的实验结果表明RAG增强生成的用例显著扩大了风险触发的广度（从28类增加到38类），证明其能有效生成覆盖更多隐蔽和复杂场景的测试输入。更重要的是，其在新兴威胁集上的攻击成功率（41.3%）远高于基线（12.5%），这强有力地证明了RAG技术通过动态检索最新攻击知识，能够有效生成针对新型、未知威胁的高质量测评用例。

表6 测评用例增强效果分析

测评用例集	触发风险类别数	新兴威胁集攻击成功率
Baseline	28	12.5%
RAG-Aug (Our)	**38**	**41.3%**

（3）内生安全大模型测评科目体系评估

本文通过消融实验，评估全面科目体系的重要性，在此比较了（A）仅测评传统内容安全科目；（B）仅测评模型系统安全科目；（C）测评本文内生安全科目体系（即数智风洞方案）。评估指标为在Qwen-VL-Max模型上发现的问题科目数量。表7的结果显示，仅关注单一维度（A或B）会遗漏另一维度的重大风险。更重要的是，数智风洞发现了6个交织性问题（例如，一个特定的语法级后门被触发后，会导致模型生成违法信息）。

（4）裁决机制性能分析

表7 内生安全大模型测评科目体系评估

测评科目范围	发现高危生成内容风险	发现高危系统安全风险	发现交织风险
A	15	0	0
B	0	8	0
C (Our)	14	9	6

以Qwen-VL-Max模型为测评对象，对比单一裁判模型（采用Qwen3-30B-A3B作为裁判）与数智风洞的多模型集成裁决委员会（由Qwen3-30B-A3B、DeepSeek-V3.2和一个基于规则的专家系统构成），在包含1000个争议样本（人工标注为难判定的边界案例）的数据集上进行测试。表8的结果表明多模型集成裁决机制将漏报率和误报率分别降低了约58%和56%，综合准确率得到显著提升。这表明，通过集成架构、数据和决策机制各异的模型，裁决委员会能够实现优势互补，有效纠正单一模型的系统性偏差和判断错误，极大提升了测评结果判定的鲁棒性和可信度。

进一步地，为了评估所提框架可信性和鲁棒性，设定人类评估一致性指标，指标定义如下：采用Cohen’s Kappa系数来衡量“数智风洞”的多模型集成裁决结果与人工专家评审结果之间的一致性程度，以量化自动裁决机制的可信赖性。实验步骤是从1000个争议样本中随机抽取300条模型输出，由3位安全领域专家（均具备2年以上内容审核或AI安全研究经验）进行独立盲审，标注“有害/安全”及所属风险类别。以专家投票结果作为黄金标准。

实验结果显示数智风洞集成裁决结果与专家黄金标准的Cohen’s Kappa系数为0.79。作为对比，使用单一Qwen3-30B-A3B作为裁判的Cohen’s Kappa系数为0.68。高的Kappa系数（0.79）证明

了多模型集成裁决机制的输出与人类专家判断具有高度一致性。、

表8 裁决机制性能对比

裁决机制	漏报率	误报率	综合准确率	人类评估一致性
单一裁判	9.3%	7.7%	90.2%	0.68
多模型集成裁决 (Ours)	3.9%	3.4%	94.4%	0.79

（5）各模块独立贡献与相互影响的量化分析

为明确数智风洞中容器微隔离（A）、RAG用例增强（B）、多模型集成裁决（C）三个核心模块的独立贡献与协同效应，设计了系统的消融实验与控制变量实验。首先，构建了四个实验变体，在Llama-3-70B模型上进行全面测评，变体情况如下：①V1（完整框架）：A + B + C；②V2（无RAG增强）：A + C，使用基础数据集；③V3（无容器微隔离）：B + C，采用单容器执行；④V4（无多模型集成裁决）：A + B，使用单一裁判（Qwen3-30B-A3B）。

表9展示了各变体在关键指标上的表现。实验结果显示容器微隔离（A）对时效性提升最显著（耗时减少约46.6%），而对精度指标影响甚微，印证其负责资源调度与并发效率。RAG用例增强（B）对覆盖度与召回率贡献最大（分别提升约23.6%与12.7%），证明其有效扩展了测试场景与威胁发现能力。多模型集成裁决（C）是准确率与F1分数提升的关键（分别提升约7.5%与6.6%），显著降低了误判与漏判。

基于表9结果，进一步分析模块间的交互效应。首先，B与C的协同：RAG生成的多样用例（尤其新兴威胁）对集成裁决机制提出了更高要求；实验显示，在无C的情况下，B带来的覆盖度提升会导致误报率上升（从2.8%升至6.9%）。其次，A与B的协同：容器并发环境下，RAG检索与生成任务可并行调度，使得动态生成集的构建时间缩短约35%。最后，A与C的协同：并发测评产生的海量输出（如数万条响应）需要裁决机制具备高吞吐能力；集成裁决中的并行模型调用设计与容器并发架构高度匹配，未出现性能瓶颈。

表9 数智风洞核心模块独立贡献与相互影响的量化分析

实验变体	覆盖度	平均耗时(小时)	准确率	召回率	综合F1分数
V1	89.3%	3.1	93.7%	91.2%	92.4%
V2	65.7%	3.3	92.1%	78.5%	84.7%
V3	88.9%	5.8	93.5%	90.8%	92.1%
V4	89.1%	3.2	86.2%	85.4%	85.8%

4 结束语

针对生成式大模型安全测评面临的测评用例固化、评估维度单一、结果判定不可靠等核心挑战，提出了“数智风洞”一个面向生成式大模型的内生安全测评框架。本框架借鉴系统化“风洞试验”思想，旨在构建一个用例可迭代、功能可扩展、结果可置信的动态测评环境。该框架将测评流程系统性地分解为测评基础支撑、对象设定、用例增强、环境配置与结果判定五大核心模块，实现了测评任务的全生命周期闭环管理，为规范、动态的大模型安全评估提供了系统性解决方案。实验评估表明，相较于现有主流测评框架，数智风洞在测评的时效性、覆盖度和准确率等核心指标上均展现出显著优势。无论是在文本大模型还是多模态大模型上，框架均能更高效、更全面地发现传统内容安全漏洞、模型安全漏洞以及二者交织产生的复杂风险，验证了其设计的有效性与先进性。未来研究将围绕基线框架配置进行深化研究与优化，实现同等并发条件下的测评能力对比和分析。

参考文献：

[1] BIE F X, YANG Y B, ZHOU Z Z, et al., RenAIssance: A Survey Into AI Text-to-Image Generation in the Era of Large Model [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(3): 2212-2231.

[2] LI J W, YANG Y Z, BAI Y, et al., Fundamental Capabilities of Large Language Models and their Applications in Domain Scenarios: A Survey [C]. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024: 11116-11141.

[3] QIN Z, CHEN D Y, ZHANG W H, The Synergy Between Data and Multi-Modal Large Language Models: A Survey From Co-Development Perspective [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(10): 8415-8434.

[4] 郑纬民. 分布式技术在大模型训练和推理中的应用[J]. 大数据, 2024,10(5):1-10.
ZHENG Weimin. Application of distributed techniques in large language model training and inference[J]. Big data research, 2024, 10(5): 1-10.

[5] WANG H Y, LI Y H, WANG Y, et al., Navigating the Risks: A Review of Safety Issues in Large Language Models [C]. Proceedings of the IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C), 2025: 74-83.

[6] DAS B C, AMINI M H, WU Y Z. Security and Privacy Challenges of Large Language Models: A Survey [J]. ACM Computing Surveys, 2025, 57(6): 1-39.

[7] JIA F R, WU T, QIN X. The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents [C]. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025: 29680-29697.

[8] WANG Y Z, HU W B, DONG Y P, et al., Align Is Not Enough: Multimodal Universal Jailbreak Attack Against Multimodal Large Language Models [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(6): 5475-5488.

[9] LIU A S, ZHOU Y G, LIU X L, et al., Compromising LLM Driven Embodied

Agents With Contextual Backdoor Attacks [J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 3979-3994.

[10] DU X H, MO F, WEN M, et al., Multi-Turn Jailbreaking Large Language Models via Attention Shifting [C]. Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI), 2025: 23814-23822.

[11] YANG Z P, FAN J L, YAN A L, et al., Distraction is All You Need for Multimodal Large Language Model Jailbreaking [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 9467-9476.

[12] BUITRAGO P A, NYSTROM N A. Open Compass [C]. Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines Learning, 2019: 1-9.

[13] ZHANG M, PAN X D, YANG M, JADE: A Linguistics-based Safety Evaluation Platform for Large Language Models [J]. Arxiv, 2023: 1-30.

[14] XU L, LEI A Q, ZHU L, et al. SuperCLUE: A Comprehensive Chinese Large Language Model Benchmark [J]. Arxiv2023: 1-13.

[15] HE Z Q, LIU Y S, ZHENG J S, et al. FlagEvalMM: A Flexible Framework for Comprehensive Multimodal Model Evaluation [C]. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025, 1-11.

[16] NI R K, XIAO D, MENG Q Y, et al., Benchmarking and Understanding Compositional Relational Reasoning of LLMs [C]. Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI), 2025: 19703-19711.

[17] BAO F S, LI M R, QU R Y, et al., FaithBench: A Diverse Hallucination Benchmark for Summarization by Modern LLMs [C]. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 2025: 448-461.

[18] CHEN B C, PALIWAL A, YAN Q B, Jailbreaker in Jail: Moving Target Defense for Large Language Models [C]. Proceedings of the 10th ACM Workshop on Moving Target Defense (MTD), 2023: 29-32.

[19] DENG B Y, WANG W J, ZHU F B, CrAM: Credibility-Aware Attention Modification in LLMs for Combating Misinformation in RAG [C]. Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI), 2025: 23760-23768.

[20] SONG S Z, LI X P, LI S S, et al., How to Bridge the Gap Between Modalities: Survey on Multimodal Large Language Model. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(9): 5311-5329.

[21] HAYSE J, BOREVKOVIC E, CARLINI N, Query-Based Adversarial Prompt Generation [C]. Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), 2024: 1-13.

[22] ZHAO H Q, YUAN C H, HUANG F, et al., Qwen3Guard Technical Report [J]. Arxiv, 2025: 1-28.

[23] 李国杰. 大数据与计算模型[J]. 大数据, 2024,10(1):9-16.
LI G J. Big data and computing models [J]. Big data research, 2024, 10(1): 9-16.

[24] JI J M, LIU M, DAI J, et al., BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference

Dataset [C]. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS) 2023: 1-27.

[25] YING Z H, LIU A S, LIANG S Y, et al., SafeBench: A Safety Evaluation Framework for Multimodal Large Language Models [J]. Arxiv, 2024 1-32.

[26] HARTVIGSEN T, GABRIEL S, PALANGI H, et al., ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022: 3309-3326.

[27] SHI X Y, CHEN S F, ZHNAG G, et al., Jailbreak attack with multimodal virtual scenario hypnosis for vision-language models[J]. Pattern Recognition. 2026 (172): 112391.

[28] PENG D, KE Q H, HUANG M H, et al., Unified Prompt Attack Against Text-to-Image Generation Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4816-4834.

作者简介

马海龙（1980– ），男，博士，信息工程大学，教授，主要研究方向为网络内生安全、云安全、网络弹性。

韩伟涛（1989– ），男，博士，信息工程大学，研究员，主要研究方向为网络内生安全、网络流量安全、网络威胁感知。

胡涛（1993– ），男，博士，信息工程大学，助理研究员，主要研究方向为大模型安全、智能对抗、网络内生安全。

收稿日期: XXXX-XX-XX

通信作者: 胡涛,hutaondsc@163.com

基金项目: 国家重点研发计划资助项目(No. 2024YFB2907202)

Foundation Items: The National Key Research and Development Program of China (No. 2024YFB2907202)