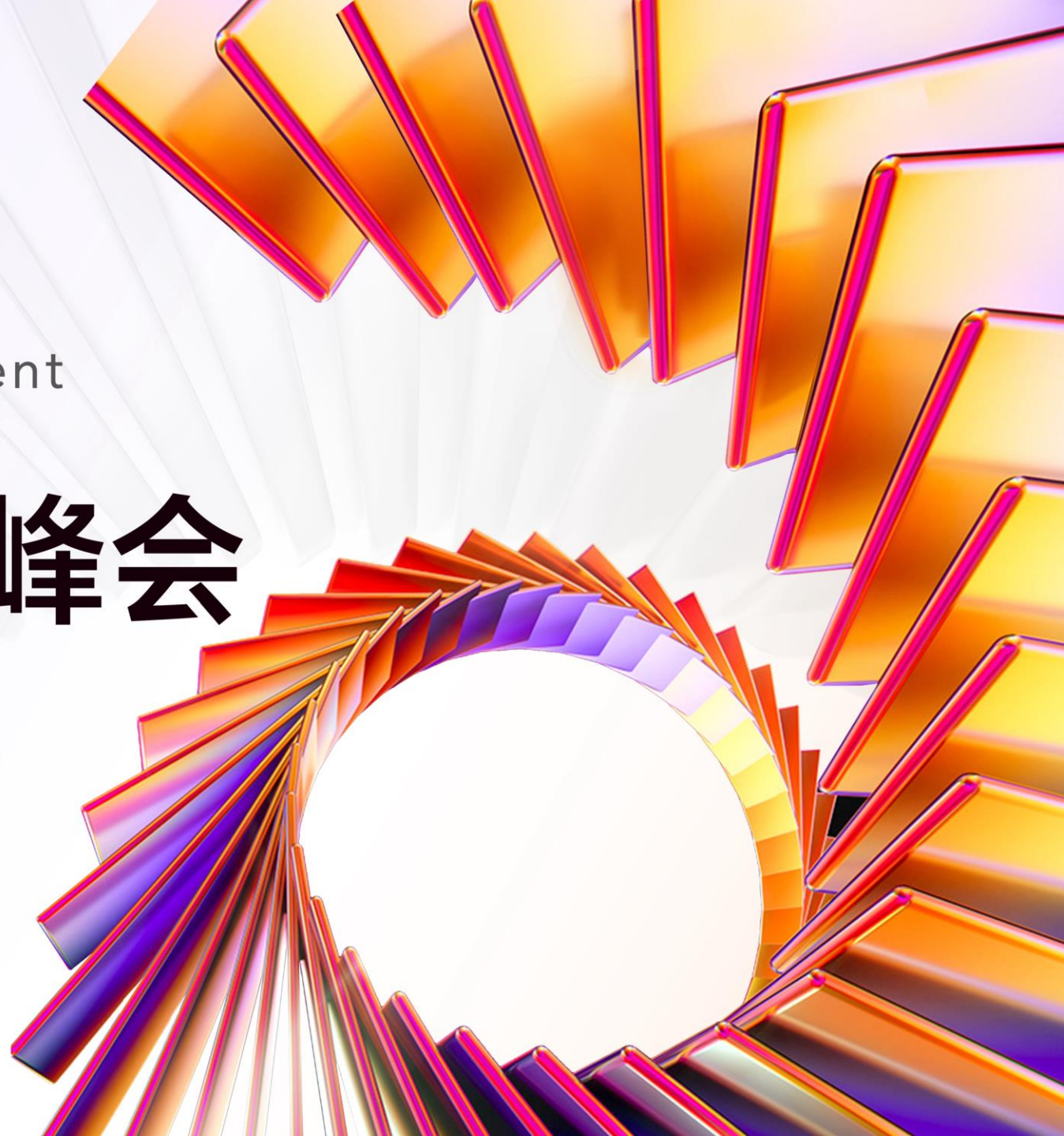
AiDD 7th 2025

第7届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

8月8-9日 | 北京站



AiDD 8th 2025

第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

查看会议详情

下一站预告 11/14-15 | 深圳站 12/19-20 | 上海站

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

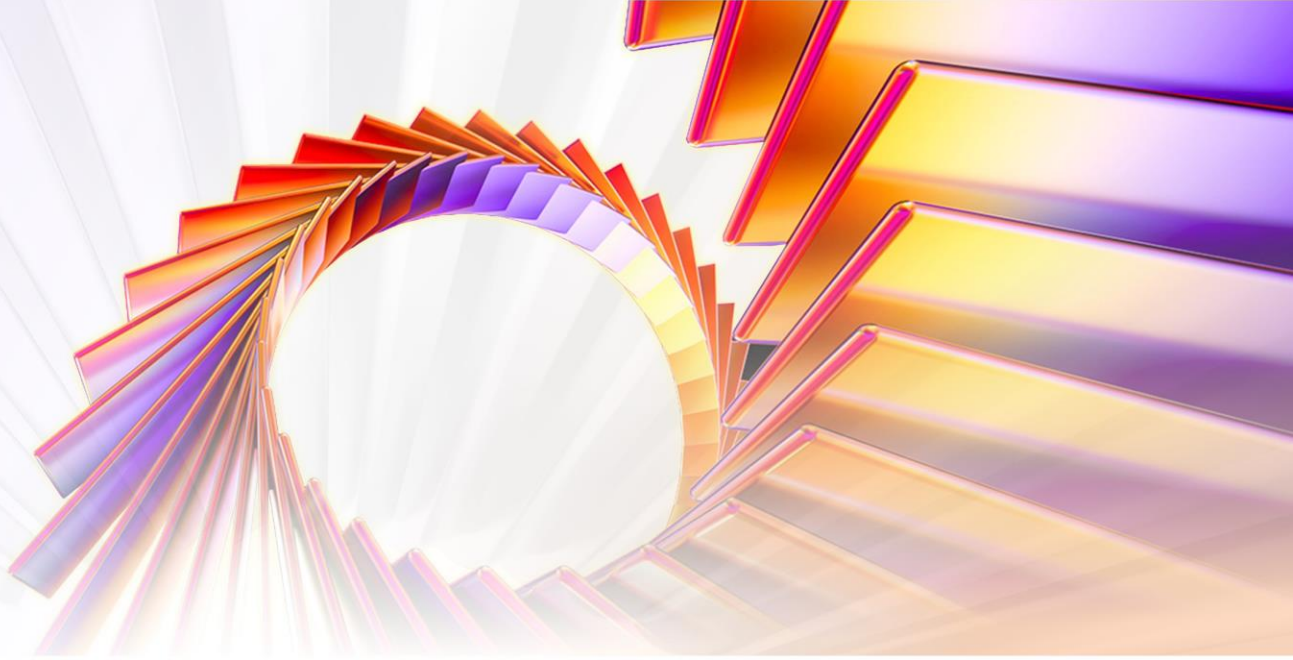
上下文工程

AiDD 7th 2025 | 8月8-9日 | 北京站

第7届 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



AI 模型评测新范式和关键技术

王旭东 | 蚂蚁集团 高级技术专家

王旭东

蚂蚁集团 高级技术专家

清华大学硕士，2019年加入支付宝，目前担任高级技术专家，技术风险部 AI 质量工程团队负责人。团队负责定义 AI 质量标准，管理 AI 特有风险，通过专业方法与平台工程，构建从数据到模型和 AI 应用的全生命周期保障，确保 AI 系统可靠、安全、高效地交付业务价值。

目录
CONTENTS

PART 01

AI 模型评测新范式概述

背景

- 2022 年 11 月底，ChatGPT 发布

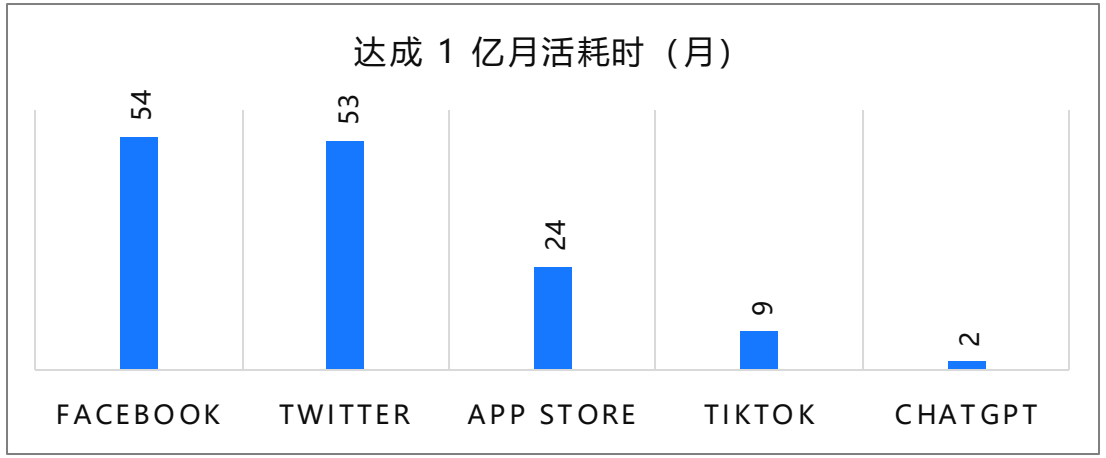
- **相信**：大模型会成为新的技术潮流

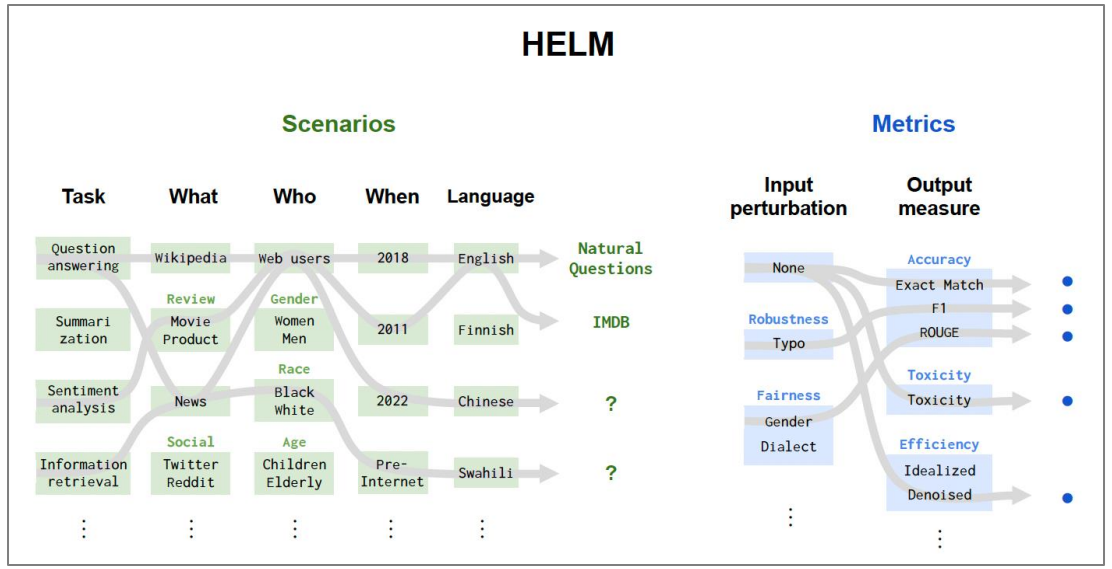
- **预判**：**以大模型为基础的 AI 研发，一定也需要强大的评测能力作为支撑**

1. 验证模型能力
2. 确定模型优化方向

- **行动**：

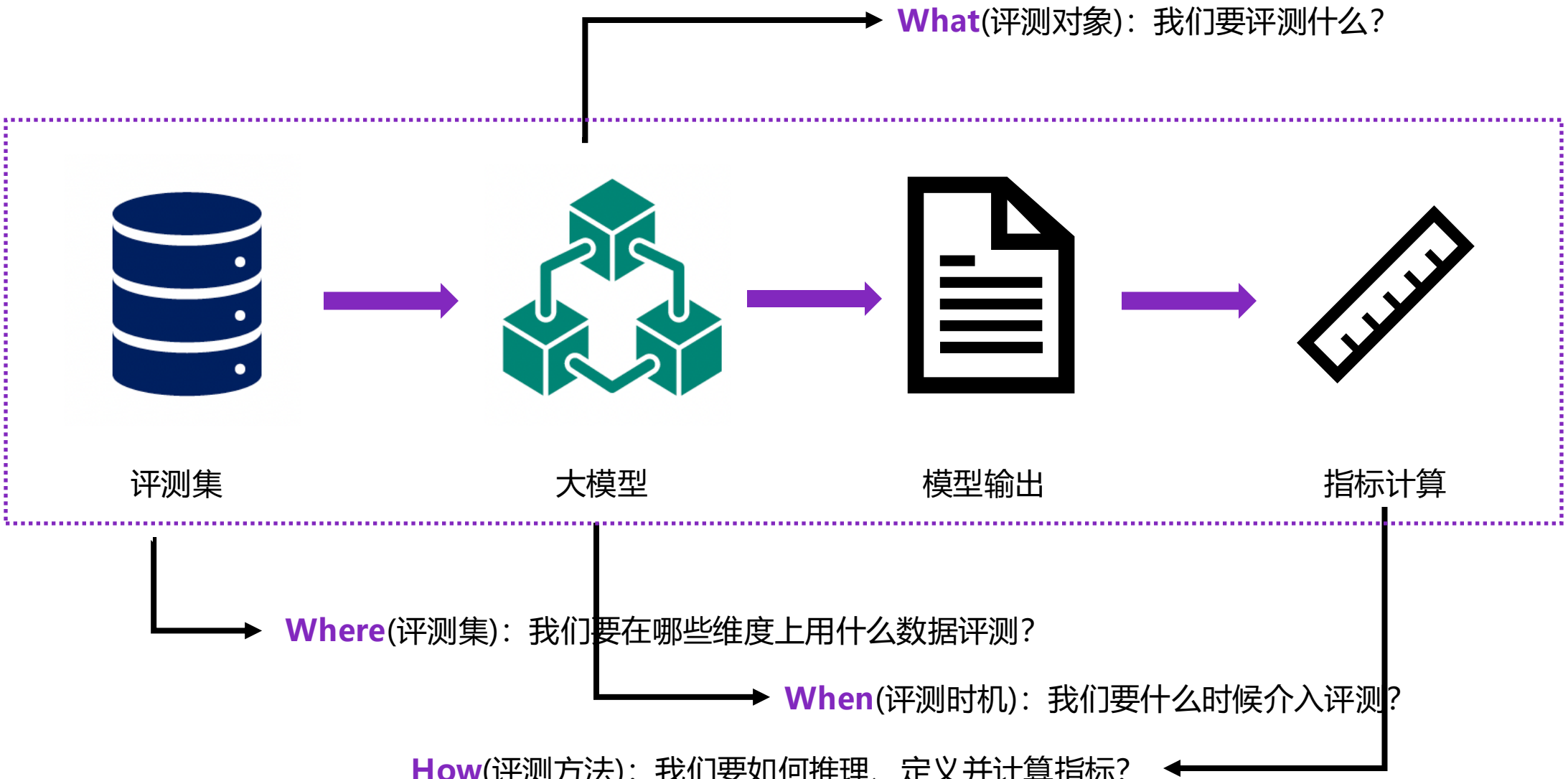
1. 深入调研大模型评测技术（从 HELM 这篇文章出发）
2. 立项，通用的 AI 模型评测能力





Percy Liang et al., 2022, Holistic Evaluation of Language Models

AI 模型评测的基本要素



AI 模型评测新范式

范式转移的核心理念：从“确定性”到“概率性”

维度	传统软件质量	AI 模型评测 (新范式)
系统行为	确定性：输入 A，永远得到输出 B	概率性：**输入 A，可能得到 B、C 或 D**，关心的是得到“好”答案的概率有多高
对正确的定义	二元的：功能有明确的对或错	统计的：**没有绝对的“正确”**

新范式的主要特征：

1. **评估目标的多维化：**远超“功能正确”，复杂的评估维度
2. **评估方法的根本性变革：**从“用例驱动”到“数据驱动”
3. **质量生命周期的延伸：**从“事后”到“全程”（训练全生命周期）

评测目标多维化

传统质量主要关注功能、性能、安全、兼容性等。

AI 模型评测在此基础上，引入了全新的、更复杂的评估维度：

1. **性能/准确性：**这是基础，在一系列复杂的评测基准上评价性能指标。
2. **安全性：**模型是否可能被用于恶意目的？是否会生成有害、违法或有毒的内容？是否容易受到数据投毒等攻击？
3. **幻觉：**对于大语言模型等生成式模型，它是否会“一本正经地胡说八道”，捏造事实？
4. **鲁棒性：**模型在面对非理想输入时的表现。例如，输入有噪声、有拼写错误、甚至是经过精心设计的对抗性攻击时，模型的性能是否会急剧下降？
5. 公平性与偏见：模型是否对不同群体（如性别、种族、地域）表现出一致的性能？是否存在歧视性行为？这是传统质量很少触及的伦理维度。
6. 可解释性：我们能理解模型为什么做出某个特定的决策吗？这对于金融、医疗等高风险领域至关重要。

评测数据驱动

传统质量的核心是测试用例，由质量工程师根据需求文档精心设计。

而 AI 模型评测的核心是评估数据集。

1. 评测基准：包含大量高质量、有代表性的标注数据，作为衡量模型性能的“标尺”。
2. 人类参与的评估：对于创造性、主观性很强的任务（如文案生成、对话质量），机器指标是不够的，由人类来打分和判断。
3. 对抗性测试：不再是测试常规场景，而是主动寻找模型的“盲区”和弱点，通过生成对抗样本来攻击模型。

评测贯穿模型开发全生命周期

模型评测不是模型训练结束后的“期末考试”，而是贯穿整个开发周期的“随堂测验”、“模拟考”

数据准备阶段：确保输入给模型的数据是高质量、多样化、无偏见且安全的。数据质量评测、数据分布评测、数据安全评测，从源头上保证模型的质量。

预训练阶段：监控训练进程，验证模型是否在正确学习，并选择最佳的模型版本。检查点（Checkpoint）评测、超参数调优评测，也叫做边训边评，评测结果可以帮助工程师判断当前训练策略是否有效。

后训练阶段：让模型学会人类的偏好，变得更“有用、诚实、无害”。奖励模型的构建与评测、对齐效果评测，这个阶段，评测本身就是训练的核心驱动力。

部署后：监控模型在真实世界中的表现，发现未知问题，并为下一代模型提供改进方向。

PART 02

AI 模型评测关键技术

Benchmark 建设的核心难题

Benchmark 建设的核心难题：Benchmark 为何难以与用户体感对齐，为何难以衡量 AGI?

- Benchmark 的本质：任何 Benchmark 都是对现实世界复杂问题的一种降维和抽象。
  - 它必须是可量化的、可重复的、标准化的，就像一把“尺子”。
  - 为了做到这一点，它必然会牺牲掉现实世界中大量的上下文、隐性知识和动态变化。
- 用户体感的本质：用户体感是高维的、整体的、充满个性化和情感因素的。
  - 用户不在乎模型在 MMLU 上得了多少分，而在乎“它能不能帮我写好这封邮件”、“它理不理解我刚才话里的反讽”、“它的回答是不是让我觉得舒服和被尊重”。
- AGI 的本质：AGI 的一个关键特征是泛化能力和适应性。
  - 它应该能处理前所未见的、定义模糊的任务，并在动态环境中持续学习和调整。
  - 而静态 Benchmark 本质上是“已知的未知”（Known Unknowns），无法衡量模型应对“未知的未知”（Unknown Unknowns）的能力。

用一把静态的、低维的尺子去衡量一个动态的、高维的、追求无限泛化能力的目标（AGI & 用户体感），本身就存在结构性的矛盾。

静态 or 动态 Benchmark

静态和动态 Benchmark 是一个评估体系的两个支柱，缺一不可。

静态 Benchmark：**“压舱石”，它像“高考”**，是筛选基础知识和基本逻辑能力的有效工具

优点：

✓ 可复现性与公平性：为不同模型提供了一个公平比较的平台。

✓ 诊断短板：可以针对性地测试模型在某个特定能力（如数学推理、代码生成、知识问答）上的强弱。

✓ 推动基础研究：清晰的指标可以引导学术界和工业界在特定方向上攻关。

缺点：

☒ 过拟合：模型可能会“刷分”，学会 Benchmark 的套路，而非真正的能力。

☒ 时效性差：知识和能力要求在快速变化，静态 Benchmark 很快会过时。

☒ 与应用脱节：高分不等于在实际应用中有用。

静态 or 动态 Benchmark

动态 Benchmark：**“试金石”，它像“社会实践”或“工作实习”**，检验的是解决实际问题的综合能力

优点：

✓ 真实反映用户需求：直接与用户体感和商业价值挂钩。

✓ 捕捉“涌现”问题：能发现模型在受控环境中暴露不出的新问题、偏见和安全漏洞（Red Teaming 就是一种形式）。

✓ 驱动模型“反脆弱”：充满噪声和变化的数据强迫模型变得更鲁棒、更具适应性。

缺点：

☒ 评估成本高、信噪比低：需要大量人工标注或复杂的 A/B 测试系统，且用户反馈充满主观性和噪声。

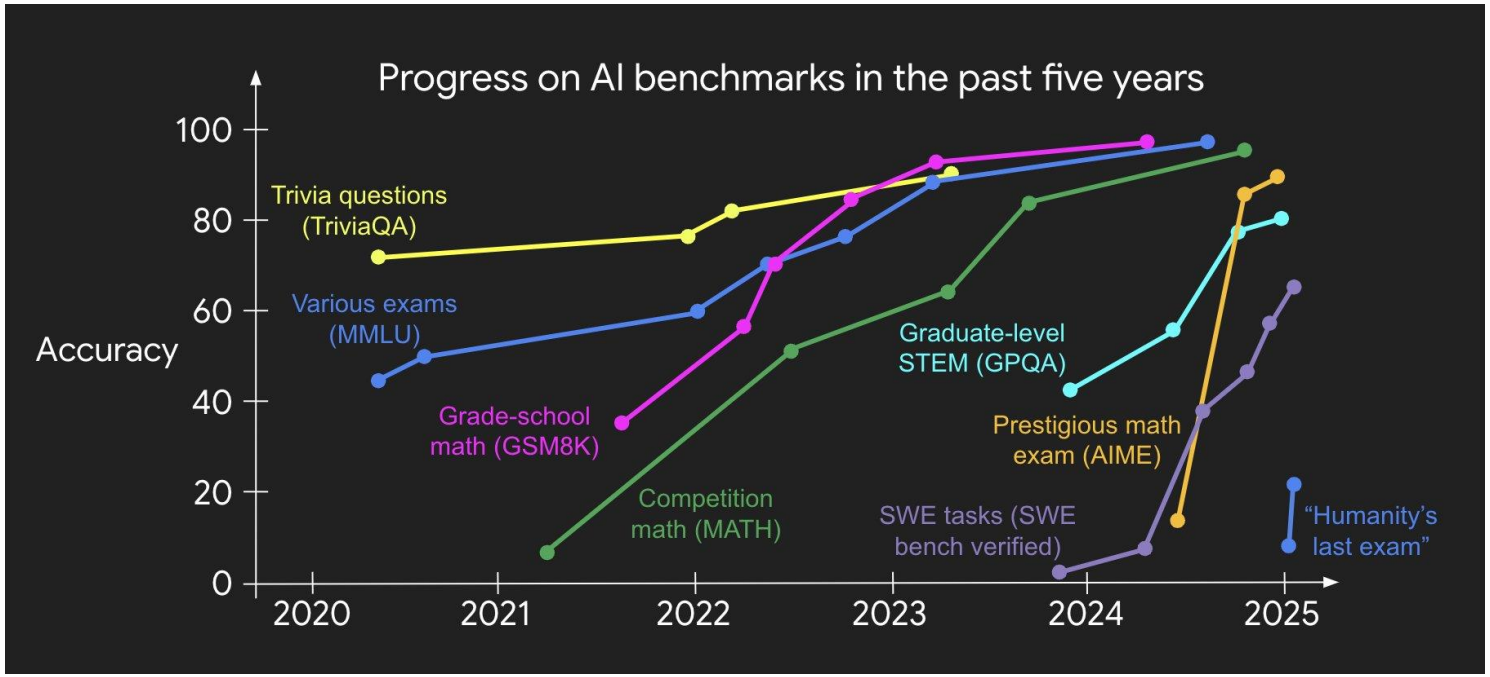
☒ 可复现性差：动态数据流难以精确复现，使得模型间的“apples-to-apples”比较变得困难。

☒ 指标定义模糊：如何量化“用户满意度”、“创造性”等指标本身就是难题。

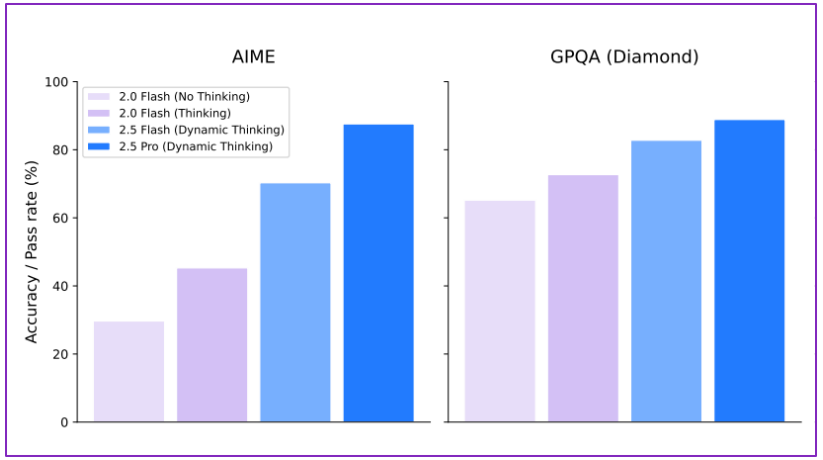
评测集需要随着模型能力提升持续迭代

GSM8K 的分数近年迅速攀升到 90 分以上，**很多模型不再在报告中发布 GSM8K 的分数**

取代的是例如 AIME24&25 这样的评测集



来源：Shunyu Yao 的 Blog, https://ysymyth.github.io/The-Second-Half/



来源：Gemini-2.5，https://arxiv.org/abs/2507.06261

Benchmark 如何演进

在基础模型研发阶段：更关注静态 Benchmark。

- 这个阶段的目标是构建模型的通用基础能力。
- 我们需要标准化的尺子来衡量模型在数学、逻辑、知识等核心维度上是否取得了突破。
- 没有这个基础，谈论应用是不切实际的。

在产品应用迭代阶段：更关注动态 Benchmark。

- 这个阶段的目标是解决用户的具体问题，创造价值。
- 模型在 Benchmark 上哪怕低几个点，只要能在特定场景下更好地满足用户需求，它就是更成功的“产品”。
- A/B 测试、用户留存率、满意度调查等指标是金标准。

Benchmark 如何演进

在长远视角下：两者形成一个螺旋上升的闭环，动态数据驱动静态 Benchmark 的进化

1. **启动：**用现有的静态 Benchmark 训练和评估模型，达到一个能力基线。
2. **发现：**将模型投入真实环境中，通过动态数据发现其能力的“短板”和静态 Benchmark 的“盲区”。
3. **抽象：**提炼和设计出新的、更高质量的静态 Benchmark。例如，发现模型只会做题，不会规划复杂任务，就催生了 Agent 类 Benchmark。
4. **迭代：**用新的 Benchmark 驱动下一代模型的研发。
5. **循环：**周而复始，这个体系不断进化，尺子越来越精良，模型能力越来越强。

什么是真正高质量的 Benchmark

1. **广度与深度**：不仅覆盖多领域知识，更要考察多步推理、规划、创造等深层能力。
2. **抗“应试”性**：题目设计巧妙，难以通过搜索或简单的模式匹配来“作弊”。最好是过程性评估，而非仅仅看最终答案。例如，评估代码不仅看运行结果，也看代码质量和解题思路。
3. **动态与演化性**：Benchmark 本身应该是一个“动态”的系统。它可以定期从真实世界数据中采样新问题，或者由人类专家、甚至其他 AI 持续地生成新的问题。
4. **诊断性与可解释性**：不仅给出分数，更能揭示模型在哪些模块上存在缺陷。它应该能回答“模型为什么错”，而不仅仅是“模型错了”。
5. **对齐人类价值观**：必须包含对安全性、公平性、偏见、伦理等方面的严格测试。例如，Safety Benchmarks。
6. **衡量泛化而非记忆**：确保评测数据在模型的训练集中是“零样本”或“少样本”的，真正考验其泛化能力。
7. **交互式与环境感知**：未来的 Benchmark 必然会走向交互式，在一个模拟或真实的环境中，评估模型完成复杂任务的能力，而不仅是“一问一答”。Agent 的评测就是这个方向的体现。

理想的评测体系是金字塔结构

- 塔基：广泛的、自动化的静态 Benchmark，用于大规模、高频次的基础能力评估。
- 塔中：人机结合的、动态的评测（如 Arena 模式），引入人类偏好，评估更主观的体验。
- 塔尖：在真实世界应用中的长期表现，如用户留存、任务成功率、商业价值等。

动态评测在多轮对话上的应用案例

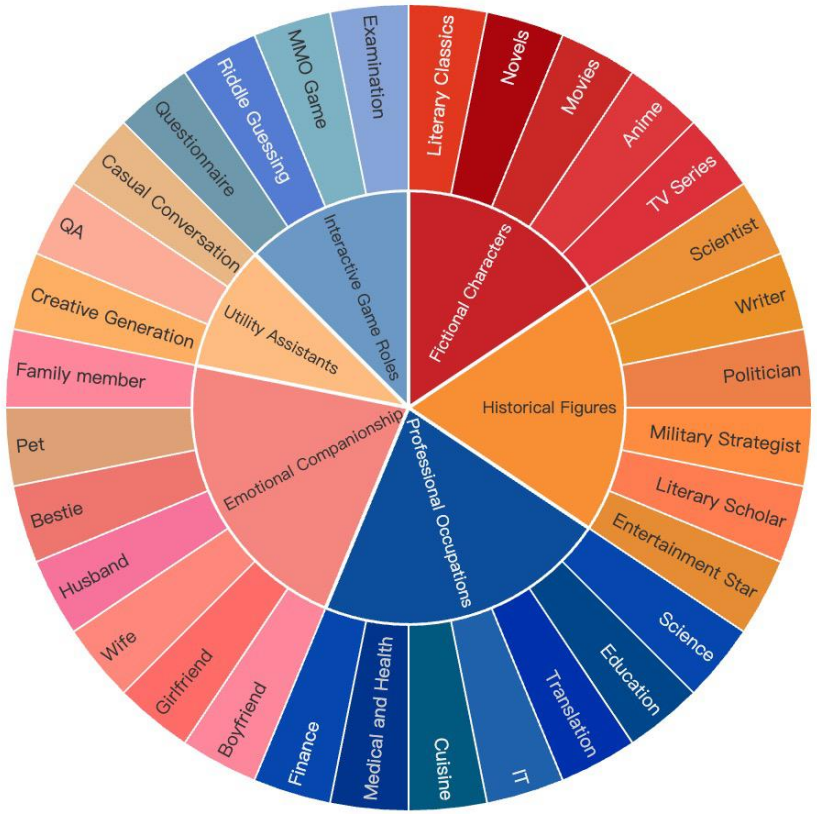
1. **LLMs Get Lost In Multi-Turn Conversation**：有文献指出，top open- and closed-weight LLMs 在多轮对话上的表现，相比单轮，在 6 个测试任务上平均会下降 39%
2. 静态的多轮对话评测，往往只能评测 2-3 轮，**后续轮次会因为模型真实 answer 变得不够连贯**

DMT-RoleBench

6 类角色：小说人物、情感陪伴、历史人物、实用助手、游戏角色、专业职位

7 类评测意图（Intent），约束生成的问题范围

3 类 System Prompt 设定

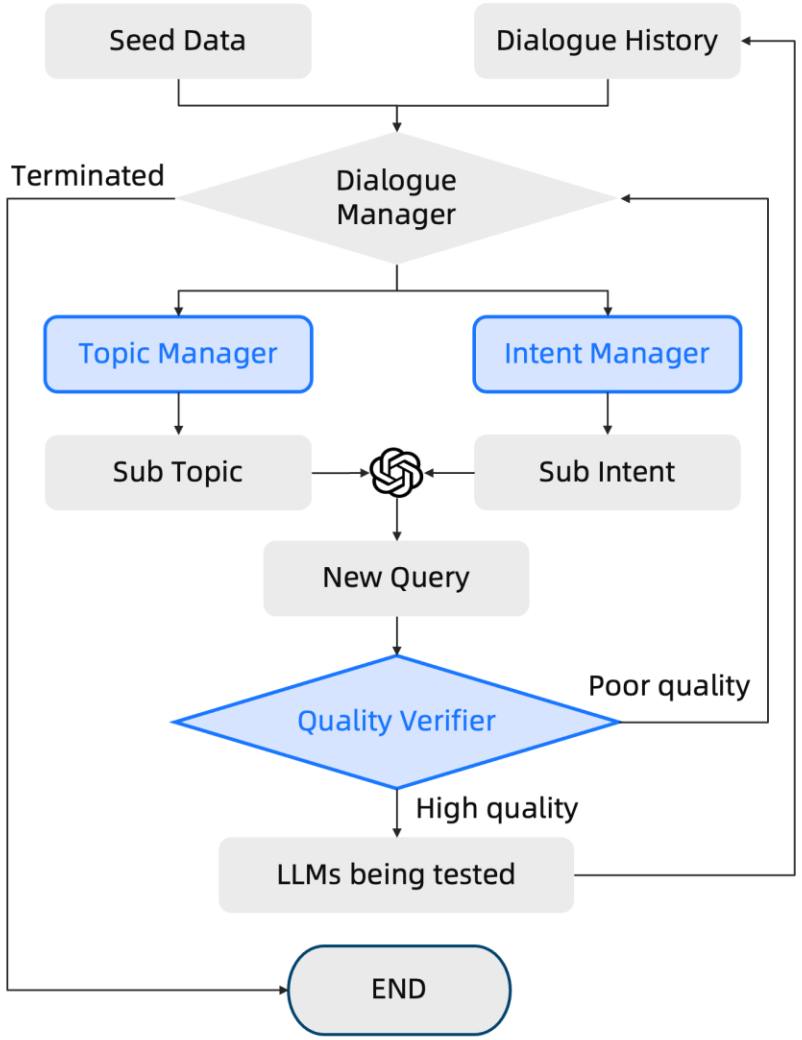


Evaluation Intents	Descriptions
Identity Recognition Eval	Evaluates the model's capability to recognize the identity information of the role being portrayed.
Role-specified Knowledge QA Eval	Evaluates the model's capability of role-specified knowledge acquistion.
Personality Trait Eval	Assesses the model's proficiency in emulating the linguistic style and personality trait of the role being portrayed.
Knowledge Boundary Eval	Evaluates the model's ability to recognize the knowledge boundaries of the role being portrayed. For instance, an individual from antiquity should not possess knowledge of how to write Python code.
Casual Conversation Steering Eval	Evaluates the model's capability for conversation steering during casual chat.
Professional Skill Eval	Evaluates the model's mastery level of professional skills pertinent to the role being portrayed. For example, a chef character should demonstrate knowledge of culinary techniques and meal preparation.
Game Interaction Eval	Assesses the model's ability of interaction in orchestrating and propelling the progression of gameplay scenarios.

Table 1: The evaluation intents and corresponding descrip-tions.

动态评测在多轮对话上的应用案例

采用一个用户模拟器，生成动态多轮对话

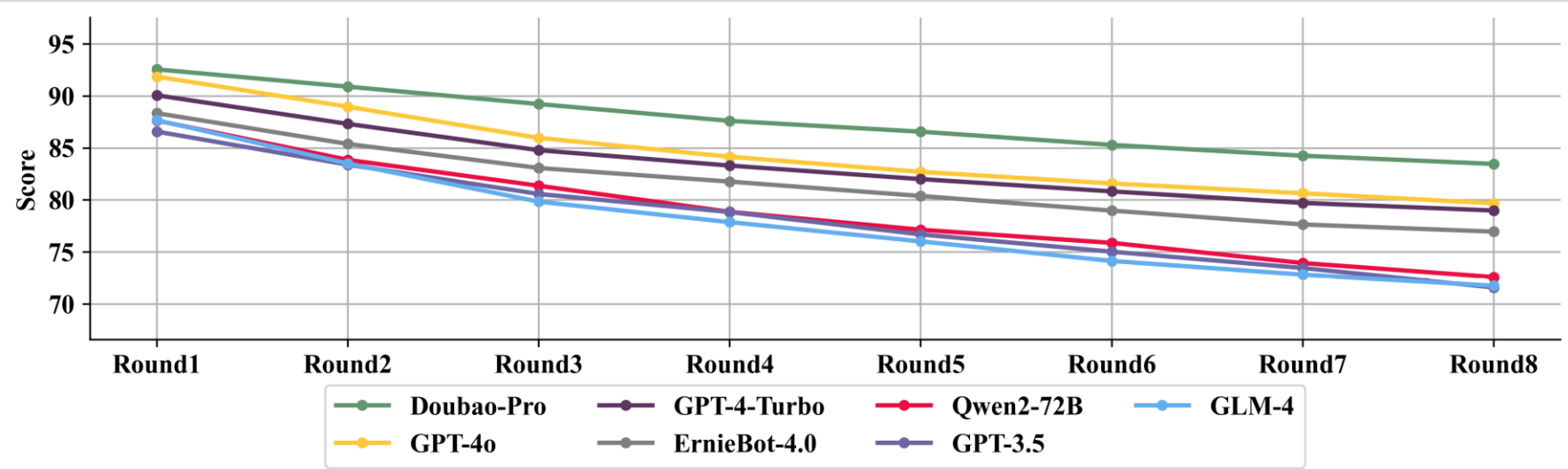


评测指标，基于 Qwen 开源模型微调，与人工标注可达 95% 一致性

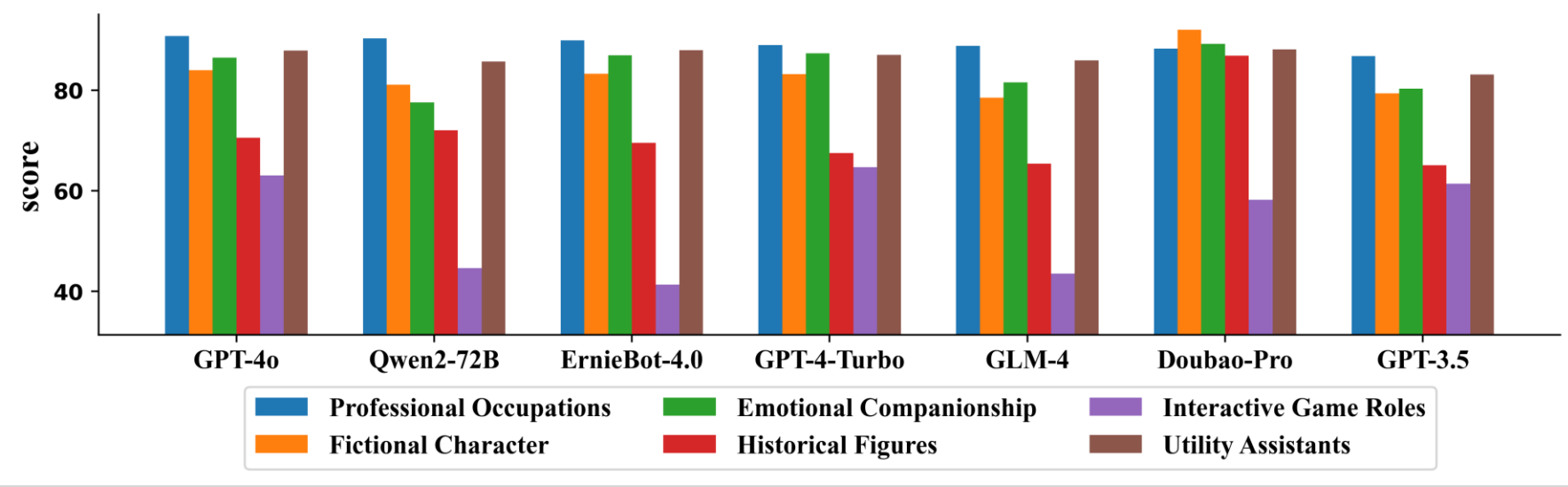
Role Types	Metrics
Fictional Characters	RE, Flu., Coh., Cons., Div., HL, KA, KH, KE, PT
Historical Figures	IF, Flu., Coh., Cons., Div., HL, KA, PT
Professional Occupations	IF, Flu., Coh., Cons., Div., HL, KA, KH
Emotional Companion-ships	IF, Flu., Coh., Cons., Div.,HL, Emp., PT, Inte.
Utility Assistants	IF, Flu., Coh., Cons., Div., HL, KA, KH
Interactive Game NPCs	GCD

Table 2: Metric allocation strategy for different role types.

动态评测在多轮对话上的应用案例



1. 得分随对话轮次增加递减
2. 不同模型递减的速度不同



1. 不同角色类型上分差明显

为什么需要 LLM Judge

传统自动化指标代表： BLEU, ROUGE

☒ 缺乏语义理解，不适用于开放式、创造性任务

☒ 无法评估“好”的程度

☒ Reasoning 模型的输出多样化，数学推理等任务的复杂化，导致 rule-based 不可靠

结论：传统指标对于评估现代 LLM 的综合能力，有很多局限

人类评估：这是评估模型质量的“金标准”

☒ 极其昂贵且耗时

☒ 难以规模化

☒ 主观性与不一致性

结论：人类评估虽然是金标准，但其高昂的成本和缓慢的速度使其成为模型开发的巨大瓶颈

LLM Judge 的缺陷 – 偏见问题

这是 LLM Judge 最核心、最普遍的缺陷。

1. **位置偏见**：这是最被广泛证实的一种偏见。在进行 A/B 对比评测时，LLM Judge 倾向于更喜欢排在第一个位置（Answer A）的答案。即使将两个答案的顺序调换，它也可能仍然选择第一个，导致评估结果不一致和不准确。
2. **长度偏见**：LLM Judge 通常会偏爱更长、更详细、看起来更全面的回答，即使这些回答可能包含冗余信息、不相关内容甚至是“幻觉”。这会鼓励被评估的模型生成冗长而非精炼的答案。
3. **迎合偏见**：LLM Judge 在评估其他模型的回答时，可能会偏爱那些风格、格式、观点与其自身相似的回答。它倾向于奖励那些“看起来像自己”的答案，而不是真正更优的答案。这会导致评估的同质化，阻碍模型多样性的发展。
4. **格式偏见**：对特定格式（如使用Markdown列表、加粗等）的偏好，即使内容质量相当，格式更规整的回答也更容易获得高分。

LLM Judge 的缺陷 – 评估能力局限

1. **事实性校验能力弱**：LLM Judge 本身也存在“幻觉”问题。它很难可靠地验证答案中的事实性错误，尤其是在专业或冷门领域。
2. **无法理解深层逻辑和创造力**：对于需要复杂推理、数学计算、代码执行或深度创造性的任务，LLM Judge 往往只能进行表面评估。
    - 代码： 它可以判断代码风格是否良好，但无法实际运行来验证其正确性或效率。
    - 数学： 它可以判断解题步骤是否“看起来”合理，但很难发现计算过程中的细微错误。
    - 创意： 它对幽默、讽刺、高级比喻、诗歌等主观和文化强相关的创造性内容，其评估标准可能非常刻板和肤浅。

LLM Judge 的缺陷 – 方法论风险

1. **对提示词敏感**：评估结果很大程度依赖于你如何设计提示词，提示词中一个关键词的改变就可能影响评估结果。这使得设计一个公平、稳定、普适的评估提示词本身就是一个巨大的挑战。
2. **评估标准无法完全对齐人类偏好**：LLM Judge 的“价值观”和判断标准来自于其训练数据，这不一定完全符合真实、多样化的人类偏好。过度依赖 LLM Judge 进行模型迭代，可能会导致模型“过拟合”到 Judge 的偏好上，而不是真正的人类用户偏好。
3. **谁来评估裁判的循环问题**：如何确定一个 LLM Judge 本身是高质量的？我们通常需要用高质量的人类标注数据来验证它。但这又回到了最初试图用 LLM Judge 来解决的问题——对人类标注的依赖。这形成了一个方法论上的循环困境。

尽管存在这些缺陷，LLM Judge 仍然是一个非常有价值的工具，尤其适用于大规模、初步的、非关键性任务的评估，可以快速筛选和迭代模型。

未来的研究方向在于如何**缓解这些偏见**、**将 LLM Judge 与人类评估相结合**，以及开发更强大、更专业的“裁判模型”，使其评估能力更接近甚至在某些方面超越非专家人类。

如何衡量 LLM Judge 的准确性

核心思想是将其判断结果与更可靠的“金标准” 进行对标，目前，最可靠的黄金标准仍然是高质量的人工标注

1. **建立高质量的“金标准”测试集**
   - 多位专家标注： 我们不会依赖单一的人类标注员。我们会邀请多位（通常是 3 位或更多）在该领域有经验的专家，对同一批模型输出进行独立打分或排序。
   - 清晰的标注准则 (Rubrics)： 专家们会遵循一套非常详细、明确的评分标准。
   - 一致性检验与仲裁： 我们会计算标注员之间的一致性。对于不一致的样本，会由更资深的专家进行最终仲裁，或者通过讨论达成共识。
2. 有测试集后，用量化的指标来衡量模型的表现：一致性、相关性系数等。
3. **Meta Evaluation：**让模型在给出分数的同时，也生成一段理由 ，评估理由的合理性和错误的归因。然后请人类专家来评估：
   - 理由的合理性： 即使分数与人类一致，裁判模型的理由是否抓住了关键点?
   - 错误的归因： 当分数不一致时，裁判模型的理由是否暴露了它的认知偏差或知识盲区?

如何提升 LLM Judge 准确性

- ✓ **优化提示工程**
  1. CoT： 要求模型在给出最终判断前，先逐步分析、推理
  2. 精细化的评分标准 (Detailed Rubrics)： 分解多个正交的、可量化的维度
  3. 少样本提示
- ✓ **模型与数据层面的优化**
  1. 使用更强的基础模型
  2. 在“困难样本”上进行微调
  3. 多模型投票： 不依赖单一的裁判模型
- ✓ **人机协同（半自动评估）：不追求用 LLM 完全取代人类，而是构建一个人机协同的评测系统**
  1. 分层审核： 使用 LLM Judge 进行大规模的、初步的筛选和打分，将模型打分处于“模糊地带” 的交由人类专家进行精细复核
  2. 持续校准： 定期用最新的“金标准”测试集来校准裁判模型

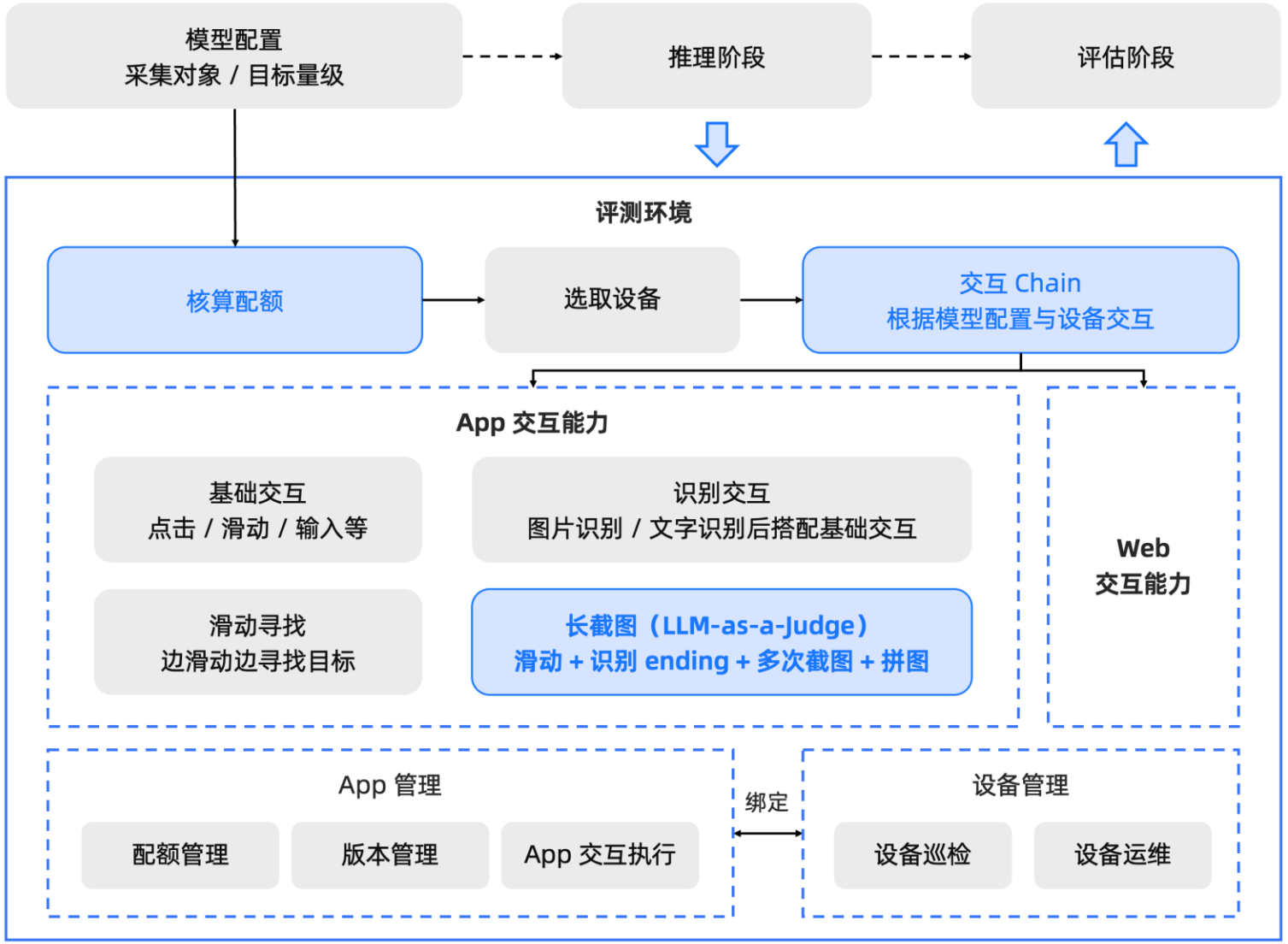
开放动态的评测环境 - 构建交互模拟式评测

问题：

使用 API 做评测，**与真实用户体感不能完全对齐**

解法：

- ✓ 大模型应用**真机交互式评测，100% 模拟真实环境的交互**
- ✓ 也支持 PC Web 页面交互



PART 03

总结与展望

总结与展望

评测与模型能力共同演进

- 下个阶段：Diffusion LLM？与真实世界交互？
- 阶段四：交互模拟
- 阶段三：推理模型评测
- 阶段二：全模态评测
- 阶段一：评测集全面、准确、细粒度

模型评测是一件重要、复杂、且有趣的工作，欢迎共同探讨

AiDD 8th 2025

第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

查看会议详情

下一站预告 11/14-15 | 深圳站 12/19-20 | 上海站

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

AiDD

「深行·浅智」

Walk Deep, Think Light.

2025.11.16

AiDD首届麦理浩径徒步

KEYLINK ing

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

AiDD峰会详情

上海站

AI+研发数字峰会

AiDD 峰会

时间：

2025.05.23-24

北京站

AI+研发数字峰会

AiDD 峰会

时间：

2025.08.08-09

深圳站

AI+研发数字峰会

AiDD 峰会

时间：

2025.11.14-15

香港站

「深行·浅智」

AiDD 首届麦理浩径徒步

时间：

2025.11.16

上海站

AI+产品创新峰会

AiDD × 联合峰会

时间：

2025.12.19-20

第7届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料