大模型标准工作组第一次会议纪要

Meeting Minutes of the First Work Group Meeting

一、大模型标准工作组介绍

Introduction of LLM Standards Working Group

1. 詹剑锋研究员介绍评价科学与工程 Evaluatology；
   Professor Jianfeng Zhan introduces Evaluatology, the science and engineering of evaluation.
2. 高婉铃副研究员介绍工作组架构、工作计划、工作机制、工作方法等；
   Associate Professor Wanling Gao introduces the Working Group's Structure, Plans, Mechanisms, and Methodologies.

二、标准工作组成员自我介绍

LLM Standards Working Group Self-introduction

1. 汤飞 （浪潮数据）我是来自浪潮科技的汤飞。我是一名云计算架构师。我对 RAG（检索增强生成，Retrieval Augmented Generation）技术非常感兴趣。我正在与 uni-RAG 技术合作进行人工智能计算，以帮助我们的逻辑销售部门将我们的原型产品介绍给客户。
   I’m Fei Tang, from Inspur Data. I am a cloud architect. I am interested in RAG technology. I am working on the ai computing together with uni-RAG technology to help our sales of logic combind to introduce our prototypes to our customer.
2. 陆刚 （阿里巴巴）来自阿里云高智能网络基础设施部门。我专注于 NVIDIA 芯片相关的网卡技术领域，致力于提升网络性能和稳定性。此外，我还带领团队进行 AI 仿真工作。
   I am Gang Lu from Alibaba cloud, specializes in network card technology related to NVIDIA chips, focusing on enhancing network performance and stability. Additionally, I lead a team in AI simulation efforts.
3. 严昱瑾（联通数科）团队主要负责联通数科智算的建设与后期运营工作。我之前在计算所进行高性能计算和并行计算优化相关，之前也接触过 top500 超级计算机的评测相关工作。芯片厂商和 GPU 种类繁多，联通数科关注芯片和架构的选择。
   Yujin Yan (China Unicom Digital Technology) leads a team responsible for the construction and subsequent operation of smart computing initiatives at China Unicom Digital Technology. My previous work involved high-performance computing and parallel computing optimization at the Institute of Computing Technology, and I have also been

involved in evaluating top 500 supercomputers. With a variety of chip manufacturers and GPUs, China Unicom Digital Technology focuses on chip and architecture selection.

4. 张（华为）simulation product optimize 数通产品线大模型训练过程的仿真模拟，类似工作，好奇大模型的评估指标，以指导仿真框架设计。

   Zhang (Huawei) is involved in optimizing simulation products for the data communication product line, specifically the simulation of large model training processes. Similar to others here, I am curious about evaluation metrics for large models to guide the design of simulation frameworks.

5. 刘悦（上海大学）计算机学院 最初对机器学习的泛化性和可解释性感兴趣，后来转向研究机器学习与材料科学的交叉应用。近期正在探索大模型与材料学的融合。今天的讨论与此密切相关，因为我正在进行材料数据质量的评估和提升工作。希望把评估的工作上升到学科高度。

   Yue Liu (Shanghai University), from the School of Computer Science and Engineering, initially focused on the generalization and interpretability of machine learning before shifting to the intersection of machine learning and materials science. Currently, she is exploring the integration of large models with materials science. Today's discussion is closely related to her ongoing work on evaluating and enhancing material data quality, which she hopes to elevate to an academic level.

6. 丁鹏（中国电信研究院）我的工作重心在于大型模型的评测领域。我们团队在中国电信内部建立了一个专门用于大型模型评测的平台。该平台旨在对中国电信自主构建的大型模型进行全面的性能评估，同时也针对那些用于赋能不同业务领域的大型模型，如网络大模型，进行性能测试。迄今为止，我们已经完成了数十款通用大型模型的评测工作，并且我们每季度都会发布一份通用大型模型的排行榜，以供行业参考。我们的目标是希望在通讯子领域内，通过这些评测工作，能够推动技术的持续进步和创新。

   Peng Ding (China Telecom Research Institute) focuses on the evaluation of large models. His team has established a dedicated platform within China Telecom for this purpose, aiming to comprehensively assess the performance of large models built in-house and those used to empower various business areas, such as network-centric large models. To date, they have evaluated dozens of general-purpose large models and publish a quarterly ranking for industry reference. Their goal is to drive continuous technological advancement and innovation in the telecommunications subdomain through these evaluations.

7. Fox：我对 AI for Science 领域充满兴趣，特别是致力于构建适用于特定科学领域的大型语言模型（LLM）。我的目标是将这些模型应用于化学和物理学等通用科学问题，以引领科学研究的进展。

   Fox: I am deeply interested in AI for Science, particularly in building large language models (LLMs) tailored for specific scientific domains. My objective is to apply these models to general scientific problems in fields like chemistry and physics, aiming to steer the progress of scientific research.

8. 詹剑锋（中科院计算所）对基本理论进行了评价。他指出，当前许多领域的评价方法往往是经验性的，这种所谓的“测量科学”实际上更接近于量子科学，其中每一个测量结果都有一个确切的真值。然而，在评价领域，很多人对于测量和评价的区别并不清晰。詹教授强调，进行评价时，我们需要定义一个“最小系统”。例如，在评价一个 CPU 或处理器时，我们必须考虑到它所在的计算机系统，包括内存、操作系统、编译器等组件。同样，对于大型语言模型的评价也是如此。评价的最小系统并不需要面面俱到，例如，在评价 CPU 时，我们通常不需要考虑网卡。但是，在构建分布式系统时，网卡则是不可或缺的。这样一个最小系统应当满足三个基本要求：第一，它必须能够独立运作；第二，它需要包含所有关键的变量；第三，它应当涵盖利益相关者的需求。例如，大型语言模型在中国和外国可能会面临不同的政治相关需求，而在教育领域的大预言模型则需要避免色情内容，同时考虑到家长的需求。今天的讨论，我们首先需要明确评价对象，并考虑到不同利益相关者的关切。我们需要针对不同领域的需求，构建大型预言模型的最小系统。我们将讨论相关工作的现状，结合并整合各位的已有成果。我们期望这项工作能够在国际上产生实际影响，并与国际竞争者相媲美。

   Jianfeng Zhan (Institute of Computing Technology, Chinese Academy of Sciences) evaluated the fundamental theories. He pointed out that evaluation methods in many fields are often empirical, and this so-called "measurement science" is actually closer to quantum science, where each measurement has a definitive true value. However, in the field of evaluation, many people lack a clear distinction between measurement and evaluation. Professor Zhan emphasized that when conducting evaluations, we need to define a "minimum system." For example, when evaluating a CPU or processor, we must consider the computer system of which it is a part, including components such as memory, the operating system, and compilers. Similarly, this applies to the evaluation of large language models (LLMs). The minimum system for evaluation does not need to be comprehensive; for instance, when evaluating a CPU, we usually do not consider the network card. However, in building distributed systems, the network card is indispensable. Such a minimum system should meet three basic requirements: first, it must be able to operate independently; second, it needs to include all critical variables; third, it should encompass the needs of stakeholders. For example, large language models may face different politically related requirements in China and abroad, while large predictive models in the education sector need to avoid pornographic content while considering parents' needs. In today's discussion, we first need to clarify the evaluation object and consider the concerns of different stakeholders. We need to construct a minimum system for large predictive models tailored to the needs of different fields. We will discuss the current status of related work, combining and integrating everyone's existing achievements. We hope this work will have practical international impact and rival that of international competitors.

9. 林放（浪潮信息）ai 和 hpc 的产品方案研发优化验证，公司内部需要进行大模型相

关的评价和标准制定。希望脱离朴素的运行 benchmark 形式。

Fang Lin (Inspur) is involved in the R&D, optimization, and validation of AI and HPC product solutions. Within the company, there is a need to conduct evaluations and establish standards related to large models, aiming to go beyond basic benchmark testing.

10. （华为）学习评价学。负责网络建模仿真和 benchmark 评测。学习类似的 SimAi 工作。

    (Huawei) Specializes in Learning Evaluation Science, responsible for network modeling, simulation, and benchmark evaluations. Studying similar work such as SimAi.

11. 张仪（华为）2012 实验室。主要关注 ai 集群可靠性。希望做 ai 网络相关的 benchmark。

    Yi Zhang (Huawei 2012 Lab) focuses mainly on AI cluster reliability and hopes to develop benchmarks related to AI networks.

12. 高婉铃（中科院计算所）关注人工智能的评价，以及在不同行业应用 ai 的评价

    Wanling Gao (Institute of Computing Technology, Chinese Academy of Sciences) is concerned with the evaluation of artificial intelligence and its application in various industries.

13. 罗纯杰（中科院计算所）芯片预研阶段的性能、功耗评价，希望借助 ai 工具

    Chunjie Luo (Institute of Computing Technology, Chinese Academy of Sciences) evaluates performance and power consumption during the pre-research stage of chip development, with a hope to leverage AI tools.

14. 鲁海荣（联通）模型微调之后的评测。

    Hairong Lu (China Unicom) focuses on evaluation after model fine-tuning.

15. 张星洲（中科院计算所）专注于边缘计算与物联网领域，致力于探索设备统一管理与编程的高效途径。我曾对多种编程框架在边缘计算环境下的性能进行过评估。然而，由于框架版本众多，配置参数复杂，评估结果存在不稳定性，且复现过程面临挑战。为此，我希望通过评价学的视角，对大型模型在边缘端的性能进行深入评估，并在此次会议中与大家分享这一案例。阮里（北航）前期更多参与超算相关研究。现在关注 ai 多模型多负载推理，和相关基准构建。

    Xingzhou Zhang (Institute of Computing Technology, Chinese Academy of Sciences) specializes in edge computing and IoT, dedicated to exploring efficient ways for unified device management and programming. He has evaluated the performance of various programming frameworks in edge computing environments. However, due to the numerous framework versions and complex configuration parameters, evaluation results are unstable, and reproducibility is challenging. Therefore, he hopes to conduct in-depth evaluations of large model performance on edge devices from an evaluation science perspective and share this case at this conference. Li Ruan (Beihang University) was previously more involved in supercomputing research but now focuses on multi-model, multi-load inference in AI

and related benchmark construction.

16. 孙启明（58 同城）从事基础大语言模型的相关研究，并在预训练后通过训练蒸馏技术将其应用于业务线。他尝试了多种评价框架和模型实验，但发现评价结果难以完全契合下游行业的特定需求，至少在置信区间方面存在挑战。他愿意为工作组提供相关案例和评价方法，期望能够加速形成一套成熟的评价体系，并尽快在应用层面落地。

    Qiming Sun (58.com) conducts research on foundational large language models and applies them to business lines through training distillation techniques after pre-training. He has tried various evaluation frameworks and model experiments but found that evaluation results do not fully meet the specific needs of downstream industries, particularly in terms of confidence intervals. He is willing to provide relevant cases and evaluation methods to the working group, hoping to accelerate the formation of a mature evaluation system and its application.

17. 石晶（北京大学）在读博士生 专注于分布式系统的性能测试研究。在实验过程中，她遇到了一个关键问题：实验结果难以复现。

    Jing Shi (Peking University), a doctoral candidate, focuses on performance testing research in distributed systems. During experiments, she encountered a key issue: reproducibility of results.

18. 刘（信通院）专注于大语言模型的研究，尤其是针对终端优化的模型。他在手机、PC、车载等终端设备上，研究完成特定任务时的性能计量和测试方法。评价过程中，他关注了性能、内存占用、时延、功耗、温度等多种指标。前期，他已推动相关团体、行业以及国家计量技术规范的进展。他期望这些方法能够具有可追溯性，并且更加准确和一致。

    Liu (China Academy of Information and Communications Technology) specializes in large language models, particularly those optimized for terminals. He has researched performance measurement and testing methods for completing specific tasks on mobile phones, PCs, automotive devices, and other terminals. During evaluations, he focused on various indicators such as performance, memory usage, latency, power consumption, and temperature. Earlier, he has promoted the progress of related groups, industries, and national measurement technical specifications. He hopes these methods will be traceable, more accurate, and consistent.

19. 陆明（联想集团）云计算的架构师。主要研究方向为云计算和智能运维。他曾与清华大学合作，完成了大模型在智能运维领域的评测工作。目前，他正专注于时空大模型相关的探索和研究

    Ming Lu (Lenovo Group) is a cloud computing architect whose main research directions are cloud computing and intelligent operations. He has collaborated with Tsinghua University to complete evaluations of large models in the field of intelligent operations. Currently, he is focusing on exploration and research related to spatio-temporal large

models.

20. 马弗里（国家空间中心）目前正致力于推进空间科学领域的大语言模型和 AI for 空间科学的科学大模型两个方向的研究。依托数据中心，对空间科学领域的数据进行汇总存储，并推动相关研究。目前，他重点关注的是这两个模型的评价，尤其是模型的可使用性和可发布性的评估。

    Ma Fuli (National Space Science Center) is currently dedicated to advancing research in two directions: large language models in space science and AI for scientific large models in space science. Relying on data centers, he aggregates and stores data in the field of space science and promotes related research. Currently, he focuses on the evaluation of these two models, especially their usability and releasability.

21. 杨郑鑫（中科院计算所）研究方向主要集中在 AI 的通用评价方法。这包括对 AI 推理系统以及传统深度学习算法的评估。目前，他专注于 AI 生成 AI 的任务。他指出，传统的 MR 性能评估工具在 AI 系统评估中存在不客观性，因此他的研究主要致力于通过评价学构建合理且客观的研究方法。

    Zhengxin Yang (Institute of Computing Technology, Chinese Academy of Sciences) focuses his research on general evaluation methods for AI, including evaluations of AI reasoning systems and traditional deep learning algorithms. Currently, he specializes in AI-generated AI tasks. He points out that traditional MR performance evaluation tools lack objectivity in evaluating AI systems. Therefore, his research is mainly dedicated to constructing reasonable and objective research methods through evaluation science.

22. 祝弘华（中科院计算所）研究的是大语言模型在芯片设计中对芯片指标的辅助预测能力，探索如何利用这些模型提高芯片设计的效率和准确性。

    Honghua Zhu (Institute of Computing Technology, Chinese Academy of Sciences) studies the auxiliary predictive capabilities of large language models in chip design for chip indicators, exploring how to use these models to improve the efficiency and accuracy of chip design.

23. 徐俊刚（国科大计算机学院）从事大模型相关的验证和评测工作，并与信通院合作制定评测标准，以确保大模型的质量和性能能够得到有效评估。

    Jungang Xu (School of Computer Science and Technology, University of Chinese Academy of Sciences) engages in verification and evaluation work related to large models and collaborates with the China Academy of Information and Communications Technology to develop evaluation standards to ensure that the quality and performance of large models can be effectively assessed.

三、标准工作组深入研讨

LLM Standards Working Group In-depth Discussion

詹剑锋老师指出，参与工作组的各位成员背景多元，涵盖了设备厂商、应用领域专家、科研机构和高校，这样的组合能够从不同角度进行深入交流。

关于工作组的特点，詹老师强调了以下三点：

1. 开放性：工作组对个人和机构开放，不设立加入门槛。
2. 公益性：工作组的标准评价活动不会用于盈利，不会借此收取费用。
3. 知识产权保护：会议将有记录者记录每位成员的发言和贡献，最终形成的标准将按照贡献进行署名。

詹老师接着分享了自己在评价学领域的研究进展，他区分了测量和评价的概念：
1. 测量是直接的，通常涉及一个物理量，通过实验和工具进行量化。
2. 评价则是间接的，与上下文紧密相关，很多上下文因素实际上会影响到评价结果，而这些往往在现有评价中被忽视。

他以 SpecCPU 和网络安全评价为例，说明了编译器优化等级和网络配置等上下文因素对评价结果的影响。詹老师还指出，在大语言模型评价领域，由于涉及的利益相关方众多，背景差异大，数据集构建多样，评价的复现性成为一个挑战。他提醒，如果在报告工作时挑选或构造对自己有利的测试集，很容易导致误导性的结论。

Professor Jianfeng Zhan pointed out that the diverse backgrounds of working group members, encompassing equipment manufacturers, application domain experts, research institutions, and universities, facilitate in-depth exchanges from various perspectives.
Regarding the characteristics of the working group, Professor Zhan emphasized the following three points:
Openness: The working group is open to individuals and institutions without any barriers to entry.
Public Welfare: The standard evaluation activities of the working group are not for profit and do not involve any fees.
Intellectual Property Protection: Meetings are recorded to document each member's contributions, and the final standards will be attributed according to these contributions.
Professor Zhan then shared his research progress in the field of evaluation studies, distinguishing between measurement and evaluation:
Measurement is direct and typically involving a physical quantity, quantified through experiments and tools.
Evaluation is indirect and closely related to context, with many contextual factors actually influencing evaluation results, often overlooked in existing evaluations.
He cited SPECCPU and network security evaluations as examples to illustrate the impact of contextual factors such as compiler optimization levels and network configurations on evaluation results. Mr. Zhan also noted that in the field of large language model evaluation, due to the numerous stakeholders involved, diverse backgrounds, and varied dataset constructions, reproducibility of evaluations poses a challenge. He cautioned that selecting or constructing test sets favorable to oneself when reporting work can easily lead to misleading conclusions.

丁朋宇 中国电信

丁朋宇 中国电信 介绍了中国电信在通用大模型和行业大模型方面的工作。他们主要聚焦于通信领域的大模型，这些模型旨在为核心网络智能化服务，辅助工作人员完成客户工单处理、网络监控、运维以及跨专业网络的故障分析定位等任务。为了达到这些目标，他们主要利用行业知识对模型进行优化。
丁朋宇指出，通信大模型需要具备以下能力：

1. 时序数据处理能力，以应对网络中的时间序列数据。
2. 多轮对话能力，以满足跨省工程师在维护网络时的大量交互需求。
3. 高确定性，因为网络故障可能带来严重的后果。

在评价大模型的复杂推理能力方面，丁朋宇提到目前存在一些不足，评估技术滞后于大型语言模型（LLM）的发展。尽管各公司如 OpenAI 的 O1 模型相比之前有了显著提升，并且提出了许多新技术，如思维链 CoT、RAG 等，但在复杂推理评价方面仍然缺乏相应的评价标准和数据集。在通用领域，虽然存在一些数学和物理数据集，但在通信等行业相关的数据集却是缺失的。此外，推理过程的步骤难以精确划分。

丁朋宇还提到，在实时通话等应用领域中，对大模型的实时性、人类情感辨别等指标的评估也相对缺乏，这些都是未来需要关注和改进的方向。

Pengyu Ding, China Telecom

Pengyu Ding from China Telecom presented China Telecom's efforts in developing general and industry-specific large models. Their primary focus is on large models for the communications sector, aimed at enabling intelligent core network services and assisting staff in tasks such as customer work order processing, network monitoring, operations and maintenance, as well as fault analysis and localization across different network disciplines. To achieve these goals, they primarily optimize the models using industry knowledge.

Ding highlighted that communication large models need to possess the following capabilities:

Time-series data processing to handle temporal data in networks.

Multi-turn dialogue to meet the extensive interaction needs of engineers across provinces when maintaining networks.

High determinism due to the potentially severe consequences of network failures.

Regarding the evaluation of large models' complex reasoning abilities, Ding mentioned current deficiencies, with evaluation techniques lagging behind the development of large language models (LLMs). Although companies like OpenAI's O1 model have shown significant improvements and introduced many new techniques such as Chain of Thought (CoT) and Retrieval Augmented Generation (RAG), there is still a lack of corresponding evaluation standards and datasets for complex reasoning evaluations. In the general domain, while there are some datasets for mathematics and physics, there is a lack of relevant datasets for industries such as communications. Additionally, the steps in the reasoning process are difficult to precisely delineate.

Ding also noted that in application areas such as real-time calls, there is a relative lack of evaluation for metrics such as real-time performance and human emotion recognition of large models, which are directions for future attention and improvement.

刘悦 上海大学

刘悦 上海大学 材料领域重视大模型，在材料领域对大模型的重视程度日益增加，尤其在学术会议中对此表现出浓厚的兴趣。她的课题组由机器学习、计算和材料实验三个部分组成，这些团队需要协作闭环，共同完成材料研究，并利用实验结果来评价大语言模型。

刘悦提到，传统的材料研究方法类似于“炒菜式”的探索，其中材料的组分、结构、工艺等因素都会影响最终的性能。为了设计出能满足多个性能指标的材料，研究者需要关注性能与这

些因素之间的映射关系。过去的研究主要依赖于物理化学原理的计算和大量的实验探索，她提出了“材料基因”的概念，希望通过材料计算学、机器学习、人工智能和大语言模型等工具，找到决定材料性能的关键因素，类似于人类基因组的作用。

在她的课题组中，大语言模型之前更多地被视为一个黑盒，而现在她们希望开发出针对垂直领域的大模型。刘悦指出，高校在大语言模型研究中面临的主要制约因素是算力不足，以及数据质量的重要性。尽管已有一些数据质量评估和提升的框架，但这些框架尚未形成完整的体系，不足以成为全行业的金标准。
她目前关注的重点包括：

1. 材料领域知识的量化与评估难题。
2. 数据标准化的进一步完善。
3. 垂直领域大模型的评价需要考虑科研工作者的应用反馈，以评估是否真正促进了材料研究流程的改进。
4. 大语言模型与传统研究范式之间的差异，以及其结果的不可解释性，可能会引起对其准确性和可靠性的质疑。
5. 材料实验的风险管理，确保大模型的设计失败不会引发伦理和安全问题。

此外，港科大的郭老师提出，大语言模型的不确定性和不可解释性可能更适合应用于文学和艺术领域。

刘悦还提到，与其他领域相比，材料领域的数据库大多开放，可以通过爬虫技术获取数据，这为材料领域的研究提供了便利。

Yue Liu, Shanghai University
Yue Liu from Shanghai University emphasizes the growing importance of large models in the field of materials science, particularly evident in the keen interest shown during academic conferences. Her research group comprises three parts: machine learning, computation, and materials experimentation, which collaborate in a closed loop to conduct materials research and utilize experimental results to evaluate large language models.
Liu Yue mentions that traditional materials research methods resemble an exploratory "cooking" process, where factors such as material composition, structure, and processing all influence the final properties. To design materials that meet multiple performance indicators, researchers must focus on the mapping relationships between performance and these factors. Past research primarily relied on calculations based on physicochemical principles and extensive experimental exploration. She introduces the concept of a "material gene," aiming to identify key factors determining material properties through tools like materials informatics, machine learning, artificial intelligence, and large language models, analogous to the role of the human genome.
In her research group, large language models were previously seen more as black boxes, but now they aim to develop large models tailored for vertical domains. Liu points out that the main constraints faced by universities in large language model research are insufficient computing power and the importance of data quality. Although there are some frameworks for data quality assessment and improvement, they have not yet formed a complete system sufficient to serve as an industry-wide gold standard.

Her current focus includes:
The challenge of quantifying and evaluating knowledge in the field of materials.
Further refinement of data standardization.
Evaluating vertical domain large models by considering feedback from scientific researchers to assess whether they truly facilitate improvements in materials research processes.
The difference between large language models and traditional research paradigms, as well as the uninterpretability of their results, which may raise questions about their accuracy and reliability.
Risk management in materials experimentation to ensure that design failures of large models do not lead to ethical and safety issues.
Additionally, Professor Guo from the Hong Kong University of Science and Technology suggests that the uncertainty and uninterpretability of large language models may make them more suitable for applications in literature and art.
Yue Liu also mentions that, compared to other fields, most databases in the field of materials are open and accessible through web scraping techniques, providing convenience for materials research.

王一帆 计算所
现在做的与大模型相关的内容主要有两部分：1. 对现有分布式算力软件构建测试工具；2. 提供一个算力网的新架构。支撑大模型的训练和推理，今天来学习一下。

Yifan Wang, Institute of Computing Technology, Chinese Academy of Sciences
My current work related to large models mainly consists of two parts: 1. Building testing tools for existing distributed computing power software; 2. Providing a new architecture for the computing power network to support the training and inference of large models. Let's delve into this today.

汤飞 计算所
现在负责智能问答大模型系统。这个项目主要是分析内部文档，帮助提供更详细的售前服务，相当于一个内部的行业大模型。就是关于 Open Stack、K8S 这样的云操作系统的大模型。会问一些简单的问题，比如就是怎么创建虚拟机、怎么调度、这个系统支持什么功能这些？举个例子吧，售前可能他不了解这个产品支持什么功能，但是他投招标的时候。他可能要求你这个产品必须得支持某个功能。这样这个时候你可以直接问这个大模型，是不是支持这个功能，然后大模型告诉你这个产品是不是支持这个功能的？这是个简单的例子。

因为我们每个组的人，可能他并不是对各个功能都了解嘛，对吧？还有可能查一些文档什么的，这个知识文档 API 文档什么的，他也可以查一下嘛，对吧？

我就觉得就说通用的大模型。他可能在通用能力上可能比较强，但是在垂直领域他能力是需要单独的评测 100 个行业的话，可能就得分别针对这 100 个行业做 100 个 Benchmark，100 个评价标准。
我觉得如果说用一套标准的话，其实并不现实的，但是能不能用一套标准的一个数据生成方法，或者说数据的生成方法。能够用户上传这个行业的数据，然后就能够用这套方法生成这个行业的数据集或者是测试问题。
虽然说我们没法就是构建这个标准的方法。能够抽取这个数据集分析这个数据集，然后能分析数据集 meta data 的或者什么的。构构造不同针对性的问答数据集，可能一般来说是问答数据集。因为问答数据集比较简单嘛，还有当然也有可能是其他的数据集，做出功能对话的

一个数据集其实我觉得也是很需要的。当然，构造这样的数据集，其实也挺难的。其实我提的一个想法，就是说能不能构造一种这样的方法。之前其实清华大学，有个工作就是为不同的行业构建不同的数据集，但是他们工作最大的 bug 是它在构建设计的过程中需要人为的干预来造 meta data（数据集的属性等），这个数据构造的过程中，它不能自动化。所以我想我们创建一个标准化的数据集构建方法，而且这个数据集构建方法是自动化的，这样我觉得能解决很大的问题。

Fei Tang, Inspur

I am currently responsible for the intelligent Q&A large model system. This project primarily analyzes internal documents to assist in providing more detailed pre-sales services, essentially serving as an internal industry-specific large model. It focuses on large models for cloud operating systems like OpenStack and Kubernetes.

The system can answer simple questions, such as how to create virtual machines, how to schedule tasks, and what features the system supports. For instance, during bidding, pre-sales personnel may not understand what features a product supports. In such cases, they can directly ask the large model whether the product supports a specific feature, and the model will provide the answer.

Each team member may not be familiar with all features, so they may need to consult documents or APIs. While general large models may excel in general capabilities, evaluating their performance in vertical domains requires separate benchmarks and evaluation standards for each industry. Using a single set of standards is impractical. However, could we use a standardized data generation method that allows users to upload industry-specific data and generate corresponding datasets or test questions?

Although we cannot develop a standard method to construct and analyze datasets, we can consider creating a method to generate targeted Q&A datasets. Constructing such datasets is challenging, but I propose developing an automated and standardized approach for dataset creation. Previously, a team at Tsinghua University attempted to build different datasets for various industries, but their method required manual intervention to create metadata, preventing full automation. Therefore, I believe creating an automated and standardized dataset construction method could address many issues.

詹剑锋研究员：你刚才说要 meta data，那你这些数据的话，到时候我们要做的话有一些数据能力能公开吗？

汤飞（浪潮）：浪潮可以公开数据。

Professor Jianfeng Zhan: You mentioned metadata just now. Can some of these data and capabilities be made public when we proceed with the project?

Fei Tang (Inspur): Inspur can make the data public.

陆明（联想）：清华大学的裴丹老师早年的学生裴长华老师也在思考这个问题，就是怎么样去生成这个数据集？但是在这个地方的话，我会有些疑问，就是这个数据集在生成了之后对于我们所要做的领域它是有偏的，这是一个地方。还有一个作为这个行业的这种数据集，它可能会存在一种情况，就是它抓来的数据大量的也属于通识数据，通识数据一线的运维工程师他其实不会去用，他平常的工作里面他都已经掌握这些知识了。什么样的数据集最后才能是有价值的？这个事情其实我也没有想的很明白。

Ming Lu (Lenovo): Professor Changhua Pei, an early student of Professor Dan Pei at Tsinghua University, is also contemplating how to generate datasets. However, I have some doubts. Firstly, the dataset generated may be biased towards the field we intend to work on. Secondly, as an industry-specific dataset, it may contain a large amount of general knowledge that frontline operation and maintenance engineers already possess and do not need. What kind of dataset will ultimately be valuable? I haven't figured this out yet.

汤飞（浪潮）：所以数据集的质量问题我觉得也很重要，就是怎么评价数据的质量，这个事儿也挺难的。
Fei Tang (Inspur): Therefore, I believe the quality of the dataset is also crucial. Evaluating the quality of data is quite challenging.

徐俊刚
评价有时候是一个方法，但是是不是要做一些系统？有的评价需要人，但是不是有些需要系统，有些系统自动化的去评价。我们想尽快的分到组里面，另外就是希望有系统的评价。
Jungang Xu:
Evaluation can sometimes be a method, but do we need to develop systems for it? Some evaluations require human involvement, while others can be automated through systems. We hope to assign tasks to teams as soon as possible and have systematic evaluations in place.

严（联通）
我们主要是负责联通整个智算的建设和运营。它主要是两个阶段：

一个是建设阶段，因为现在对于咱们国产化的要求比较高，因为智算集群涉及到算、存、网三个部分，然后算力有卡，网络有 IB、RoCE 等不同的网络，然后存也有很多东西，这种不同的组合方式会导致我们整个智算机器上会产生不同的性能。包括我们还有客户，因为我们客户不可能只跑一种模型，可能会有不同的要求，然后跑的是不同的精度、不同体系规模。我们会有 AI for Science 客户、政府客户的需求，会导致我们需要去了解的环节比较多，需要去处理的内容比较纷杂。所以现在不是很确定我们要以什么样的标准去建设一个什么样体系的机器。目前可以做一些测试，然后但是我们的测试还是在探索学习的阶段，测一些已有的各种 AI 的模型，但是没有一个比较全面的评价，比如说怎么通过评价去确定我们这个机器这样的配置是不是好的、是不是好用的，甚至包括不同机器的报价、能耗也是需要考虑的。很多机器有液冷、风冷，他现在因为很多能耗比较高的时候要用到液冷，液冷有改造机房造价特别贵的问题，这些东西都在我们的考虑范围。

二是运维阶段，目前我们使用 MFU 作为评价标准，这个评价标准是否是科学的、可令人接受的也是我们目前在考虑的问题。

包括后期运营优化，我们也有一些评价标准，如果达不到这个标准是不是有优化的必要，优化到什么限度是我们可接受的，这都是我们所在意的一些东西。
Yan (China Unicom):
We are primarily responsible for the construction and operation of China Unicom's entire intelligent computing ecosystem. It mainly involves two stages:

The first stage is construction. Currently, there are high requirements for localization, as intelligent computing clusters involve computing, storage, and networking components. Different combinations of computing cards, networking technologies like IB and RoCE, and storage solutions can lead to varying performance on our intelligent computing machines. Additionally, our customers have diverse needs, running different models with varying precision and system scales. We cater to clients from AI for Science and government sectors, which requires us to consider numerous aspects and handle complex content. Therefore, we are uncertain about the standards for constructing a machine with a specific architecture. Currently, we are conducting tests, but these are still in the exploratory learning phase, focusing on existing AI models without a comprehensive evaluation. For instance, we lack a method to determine if a machine configuration is optimal or user-friendly based on evaluations, and factors such as pricing and energy consumption, including the high cost of liquid-cooled machines for high energy consumption and the need for expensive data center modifications, are also considerations.

The second stage is operation and maintenance. Currently, we use MFU as an evaluation standard, but we are still assessing its scientific validity and acceptability.

Regarding later-stage operational optimization, we also have evaluation standards. We need to determine if optimization is necessary when standards are not met and to what extent optimizations are acceptable. These are all areas of concern for us.

詹剑锋研究员：你们现在用的一些大模型的评价有没有发现有什么问题？
严：前两天清华的老师做了 AIPerf 的测试，我们试着去跑了一下，清华实际上是以 ops 为评价标准，这个评价标准第一是需要一些配置、参数的设置，我们发现反倒不如在国产机器上跑出来的效果高，参数的设置很影响整个效率波动的情况；第二是国产化的卡它不是所有的模型都能跑的，我们想要一个好上手的东西，所有国产化的东西都能跑，而不是让我们把所有现有的大模型都跑一遍，这个工作量是很大的、很困难的。

严：现在国产化卡的问题是它不一定所有的都能跑，即使能跑在跑的过程中甚至结果都是错的，那我们怎么去评价这个国产化的卡我们是否需要去采购它。包括现在提出的异构混训，那有没有这种可能性：不同的卡之间一块去跑一个测试？这种如何去评价？
Professor Jianfeng Zhan: Have you encountered any issues with the evaluation of the large models you are currently using?
Yan: A few days ago, teachers from Tsinghua University conducted the AIPerf test, and we tried it out. Tsinghua actually uses ops as the evaluation standard, which requires some configuration and parameter settings. We found that the performance on domestic machines was actually higher. The setting of parameters greatly affects the efficiency fluctuations. Secondly, not all models can run on domestically produced cards. We want something that is easy to use and compatible with all domestically produced items, rather than having to test all existing large models, which is a huge and difficult workload.
Yan: The current issue with domestically produced cards is that they may not be able to run all models, and even if they can, the results may be incorrect during the process. How do we evaluate whether we should purchase these domestically produced cards? Additionally, with the proposal of

heterogeneous mixed training, is it possible to run a test on different cards simultaneously? How should we evaluate this?

陆刚
我这边的工作是大模型模拟器，这主要是应对 AI 基础设施的缺乏或者它演进比较快的特点，我们要做一些模拟器来做方案的评估、架构的选型。

詹剑锋研究员：你这个是开源的对吧？
陆刚：对这个已经开源。

这个工作里面有一部分是 benchmark，怎么设置和选择 benchmark。因为这个模拟器你选的参数、框架、模型都会很多，我们其实是根据阿里云自己的经验，阿里云作为 AI 集群的提供商，我们知道主要客户用的是什么样的模型或者什么样的参数去训练，所以我们其实有一个典型的负载作为我们阿里云的过去对底层基础设施的评价指标或者说是评价的 workload，从现有工作来说我觉得这块是可以贡献出来的。

詹剑锋研究员：刚才她（指严）的需求你在模拟器上都能搞起来吗？都能满足吗？
陆刚：可以深度交流一下，目前我们适用的场景还挺广的。

詹剑锋研究员：还有个问题就是，你现在发现这些大模型评价里面有些东西哪些你觉得满足不了需求？
陆刚：我觉得大模型的这个目前来说我感觉缺乏一个评价的工作，就是我们现在做大模型基础设施的时候，有一个怎么评价基础设施的稳定性，因为整个 AI 集群的稳定性实际上是一个非常重要的话题，所谓的稳定性就是 AI 集群里面的硬件、软件会经常出现故障，那么问题就是你什么时候能够快速恢复。
詹剑锋研究员：阿里云来说有些数据嘛，比如说多大程度上干扰的业务？这个方便透露吗？
陆刚：这个可能得回去问一下，因为稳定性是阿里云 AI 基础设施的核心竞争力。
汤飞：我可以补充一下，去年有个 OSDI 文章就是做这个的，故障的快速的 checkpoint。
陆刚：但我刚才提的挑战不止 checkpoint，是怎么评价还有稳定性。

Gang Lu
My work focuses on large model simulators, primarily addressing the lack of AI infrastructure or its rapid evolution. We develop simulators for solution evaluation and architecture selection.

Professor Jianfeng Zhan: Is this open-source?
Gang Lu: Yes, it is already open-source.
Part of this work involves setting up and selecting benchmarks. Because there are many parameters, frameworks, and models to choose from in the simulator, we base our choices on Alibaba Cloud's own experience. As an AI cluster provider, Alibaba Cloud knows what models and parameters our main customers use for training. Therefore, we have typical workloads that serve as evaluation indicators for our underlying infrastructure, which we believe can be contributed.

Professor Jianfeng Zhan: Can your simulator meet her (referring to Yan) needs?
Gang Lu: We can have an in-depth discussion. Our applicable scenarios are quite broad at present.

Professor Jianfeng Zhan: Another question is, what do you think is lacking in the evaluation of these large models?
Gang Lu: I think there is a lack of evaluation work for large models. When we build large model infrastructure, one challenge is how to evaluate the stability of the infrastructure. The stability of the entire AI cluster is actually a very important topic. By stability, I mean that hardware and software in the AI cluster will often malfunction. The key issue is how quickly you can recover.

Professor Jianfeng Zhan: Does Alibaba Cloud have some data on this, such as the extent to which it interferes with business operations? Is this convenient to share?
Gang Lu: I may need to go back and ask about that, as stability is a core competitive advantage of Alibaba Cloud's AI infrastructure.
Fei Tang: I can add that there was an OSDI paper last year on rapid checkpointing for fault tolerance.
Lu Gang: But the challenge I mentioned earlier goes beyond just checkpointing; it's also about evaluation and stability.

北大
我主要做的研究方向偏软件测试这一块儿，然后软件测试想要尽可能的把一些测试的环节自动化。我感觉在自动化这块我们是不是可以做一些工作。因为 benchmark 肯定是需要集成很多不同的大模型，然后大模型进来以后比方说比较 detail 的是：这个接口怎么去统一；是不是要制定一个统一的接口的标准。对于测试而言我们规定了对应的一些输入，那么大模型吐出的输出是不一样的，不同的输出它对应不同的语义，那么语义和语义之间应该怎样去对齐？

已有的一些研究工作是通过比方说 Agent 这种方式去做对齐，那这种对齐它的准确性又如何？然后，我们这块有没有一些可以探索的空间能够尽量让这些输出对齐，包括像比如：不同的大模型集成进来的这些配置，从软件、硬件不同的方面应该怎样去对齐，这一块也是需要考虑的。

詹剑锋研究员：你提到的是说在整个评价过程中怎么通过自动化的测试去自动地测试框架和方法，对吧？
北大：是的

Peking University
My main research direction focuses on software testing, with an emphasis on automating as many testing processes as possible. I feel that we can contribute to this automation aspect. Since benchmarks definitely need to integrate many different large models, upon their integration, some detailed considerations arise, such as how to unify their interfaces and whether it is necessary to establish a unified interface standard. In testing, we specify corresponding inputs, but the outputs produced by large models vary, with different outputs corresponding to different semantics. How should we align these semantics?

Some existing research efforts align semantics through methods like agents, but how accurate is this alignment? Additionally, is there room for exploration in our field to align these outputs as much as possible, including aligning configurations introduced by different large models, from both software

and hardware perspectives? This is also an area that needs consideration.

Professor Jianfeng Zhan: You're talking about automating the testing framework and methodology through automated testing throughout the evaluation process, right?
Peking University: Yes.

刘思
我介绍一下中国信通院在终端大模型评价的两块工作。第一部分是对终端本身这个设备进行评价，第二部分是对模型进行评价。

首先介绍一下终端评价。思路是比较简单的，就是我们拿一些第三方或者开源的能在本地运行的大模型，比如像 LIama2，或者一些生图的 SD 这类模型，并且构建好一个测试数据集，然后把它导入到一个终端设备上，像手机、PC 或者车载，然后去运行大模型。

做终端评价主要有两个目标：第一是如何去评价它的一个真实性能，这里说的真实性就是用户在实际操作场景下能达到的，比如响应时间、内存占用、功耗。原本传统的人工智能模型规模体积比较小，可以直接导入到手机上本地运行，不需要做任何调整。但是大模型领域就需要针对具体的终端，包括它的硬件平台去做一个具体的适配。比如像在高通平台上，肯定要适用于高通，解决方案需要对大模型的框架进行量化、进行压缩、进行硬件加速等等。第二是如何构建一个可重复、可比较、可朔源的方法。这块就是要把大模型在硬件上进行运行，找到对它性能的影响因素。这块可以总结成四个部分：一是大模型的选取，目前我们主要选用一些比较经典的。然后在数据集上我们一方面选取一些典型的任务，再找一些常用的第三方开源数据集，对这些开源数据集进行筛选，尽可能保证输入到设备上的数据集保持相同的程度。第三部分是测试工具，目前正在开发，主要影响模型的一些超参数的选择，希望构建一个业内比较公认且贴合实际场景的测试评价方法，去满足输出性能的可重复性。最后是它的硬件和框架，因为我们评价的主要是终端。

Si Liu
I would like to introduce two areas of work by the China Academy of Information and Communications Technology (CAICT) in evaluating terminal large models. The first part focuses on evaluating the terminal device itself, and the second part on evaluating the model.

Let's start with terminal evaluation. The approach is straightforward: we take some third-party or open-source large models that can run locally, such as LIama2 or image-generating SD models, build a test dataset, and import it onto a terminal device like a smartphone, PC, or in-car system to run the large model.

Terminal evaluation has two main objectives: Firstly, how to evaluate its real-world performance, which refers to what users can achieve in actual operating scenarios, such as response time, memory usage, and power consumption. Traditional AI models were smaller in size and could be directly imported onto mobile phones for local execution without any adjustments. However, in the realm of large models, specific adaptations are required for different terminals, including their hardware platforms. For example, on Qualcomm platforms, solutions need to be tailored to Qualcomm, involving quantization, compression, hardware acceleration, etc., of the large model framework.

Secondly, how to establish a repeatable, comparable, and traceable method. This involves running the large model on hardware to identify factors influencing its performance. This can be summarized into four parts: Firstly, selecting large models, where we mainly choose classic ones. Secondly, selecting datasets, where we choose typical tasks and commonly used third-party open-source datasets, screening them to ensure that the datasets input into the device are as consistent as possible. Thirdly, developing testing tools, which mainly affect the selection of hyperparameters for the model. We aim to build a widely recognized testing and evaluation method that aligns with real-world scenarios to ensure repeatability of output performance. Lastly, hardware and frameworks, as our main focus is on terminals.

詹剑锋研究员：主要还是手机对吧？
刘思：手机、PC、车载，目前主要是这三块，我觉得边缘服务器也适用于这类方法，其实就是一些算力不那么强的设备。
詹剑锋研究员：现在这些大模型能在上面跑起来吗？
刘思：目前旗舰款的一些设备能够加载，只要模型量化到英特四这种水平是能够跑起来的。如果优化的比较好，大概的 Tokens 能到三四十每秒。
詹剑锋研究员：现在在终端里面有一些实际需求的场景。
刘思：手机上目前一些国内品牌，比如 OV、荣耀，它其实都是手机上有一个端测模型，这个模型能在断网条件下运行，功能上是针对具体的一个场景，比如通话的摘要提取、翻译、语音助手等，形成这样一个手机智能体。
詹剑锋研究员：如果网络联通的话，是不是都放到云端去了？
刘思：对，主流的其实还是通过云端，用服务器的方式进行。

然后就是模型评价。模型评价其实就是针对具体终端的具体场景去构建一个评价任务和数据集。我简单说一下，像手机最重要的就是通话功能，比如通话的摘要提取以及语音助手。PC 上主要关注生产力工具。汽车这块更加聚焦，因为它能影响物理空间，我们在评价的时候会评价它对汽车的操纵性能、问题回答功能等等。

Professor Jianfeng Zhan: It's mainly about mobile phones, right?
Si Liu: Mobile phones, PCs, and in-car systems are the main three areas currently. I think edge servers also fit into this method, basically devices with less powerful computing capabilities.

Professor Jianfeng Zhan: Can these large models run on them now?
Si Liu: Currently, some flagship devices can load them, as long as the models are quantized to a certain level like INT4, they can run. If optimized well, the processing speed can reach around 30 to 40 tokens per second.

Professor Jianfeng Zhan: Are there any actual demand scenarios for these in terminals?
Si Liu: Currently, some domestic brands like OPPO, Vivo, and Honor have a terminal model on their phones that can run offline. This model functions for specific scenarios, such as call summary extraction, translation, voice assistants, etc., forming an intelligent entity on the phone.

Professor Jianfeng Zhan: If the network is connected, are these functions all moved to the cloud?
Si Liu: Yes, the mainstream approach is still through the cloud, using servers.

Then, let's move on to model evaluation. Model evaluation basically involves constructing an evaluation task and dataset for specific scenarios on specific terminals. Let me briefly explain. For mobile phones, the most important feature is the call function, such as call summary extraction and voice assistants. For PCs, the focus is mainly on productivity tools. For cars, it's more focused because it can affect the physical space. When evaluating, we will assess its control performance for the car, question-answering capabilities, and so on.

陆明
我来介绍一下我们这边在智能运维大模型下的一些思考。我们在做的智能运维大模型，我们是做了一些评测的工作，主要围绕云计算的基础设施：包括 IAAS、PAAS、以及 devops 相关的工作。

做完这个工作后我们会从几个方面去看：一个方面是评测的效果我们看完后的感受。我们出的题目是工程师结合工作和客户的理解，一点点写出来的。这些题目确实帮我们测了大模型，但是有个问题是，这里面用户懂得问题比例会不会比较高。当把难得问题排除掉、又有通识问题的时候，虽然会保证一定的分看起来比较高，但是实际对用户的价值能有多少？这个是我们构造完问题后考虑的问题，什么样的题目是好的题目。

再一个评价的这个数据成本很高。在出这个题目时很多同事参与评测，但是有的题目会从里面排除掉。问题出完能够用到最后用作测试的，保守来说也就 30%上下。

我的一个思考是大模型的评测可能要跟它的提示工程结合在一起，这样对于一些特定的业务场景来说更有意义。

我们在测完之后 ZeroShot 比 FewShot 的效果要好，而其他领域的一些测试有的是 FewShot 比 ZeroShot 效果好。那么就带来一个问题：当我们进行 FewShot 预测时，给它提供什么案例会帮助大模型的评测效果变好。

我们也在做通过大模型进行人机交互的意图识别，识别后会由它进行一定的数学建模，然后帮助我们解决云平台的调度问题、预测问题。在做的时候遇到的困难是：当完成这样一个建模，我们怎么去评价这个建模的质量是否是好的、测试数据如何构成、怎么去保证测试覆盖率。

詹剑锋研究员：第二个是在什么具体场景？
陆明：举个例子，当我们进行私有云交付时，私有云的很多规则不是通用的，是根据不同企业的业务场景的。如果我们给一套资源调度方案，过段时间业务场景发生变化，方案也要变。如果我们的人到用户现场进行重新建模和交付，代价太大了，如果通过大模型让用户自助的完成大部分工作，复杂和难的问题我们再过去，这样对于用户也简单、对于我们也简单。但是我们遇到的挑战就是测试覆盖率怎么保证，我们也思考过用仿真、模拟的方法，但是仍然不确定如何保证覆盖率的问题。因为如果我们把这样一个运筹优化的东西放到私有云平台上，假如真的有问题，可能会导致平台性的故障。

Ming Lu
Let me introduce some of our thoughts regarding the large model for intelligent operations and

maintenance. We have conducted evaluations on our large model for intelligent operations and maintenance, focusing on cloud computing infrastructure, including IAAS, PAAS, and DevOps-related tasks.

After completing this work, we examined it from several perspectives. One is the impression we gained after reviewing the evaluation results. The questions we created were written by engineers based on their understanding of their work and clients. These questions did help us evaluate the large model, but one issue is whether a high proportion of the questions were too familiar to users. When difficult questions are excluded and common-sense questions are included, although it ensures a certain score looks high, how much actual value does it provide to users? This is a question we considered after constructing the questions: what makes a good question?

Another consideration is the high cost of evaluation data. Many colleagues participated in evaluating the questions, but some were excluded. Conservatively speaking, only about 30% of the questions created were ultimately used for testing.

One of my thoughts is that the evaluation of large models may need to be combined with prompt engineering, which makes more sense for specific business scenarios.

After testing, we found that ZeroShot outperformed FewShot in our scenario, while in other domains, FewShot sometimes outperformed ZeroShot. This raises a question: when performing FewShot predictions, what kind of cases should we provide to improve the evaluation results of the large model?

We are also working on intent recognition through large models for human-computer interaction. After recognition, it performs mathematical modeling to help us solve scheduling and prediction problems in cloud platforms. The difficulty we encountered is: how do we evaluate the quality of this modeling once it's completed? How do we construct test data and ensure test coverage?

Professor Jianfeng Zhan: The second question is about a specific scenario.
Ming Lu: For example, when delivering private clouds, many rules of private clouds are not universal but are based on different business scenarios of enterprises. If we provide a resource scheduling solution, and the business scenario changes over time, the solution will also need to change. If our team members go to the user's site to remodel and redeliver, the cost would be too high. If we use a large model to allow users to complete most of the work themselves, and we only handle complex and difficult issues, it would be simpler for both users and us. However, the challenge we face is ensuring test coverage. We have considered using simulation methods, but we are still uncertain about how to guarantee coverage. Because if we place such operational research optimization on a private cloud platform, any real issues could lead to platform-wide failures.

58 同城

模型评测我们总体分成了两个环节。
一个是训推环节，它衡量的是偏硬件相关的，比如训练的时候资源占用率。另一个是模型效果，大体分成三类去做评测：第一类是达到要求。比如：国家主权这类问题，它是红线类问

题，答错一道题大模型就备案失败。这种达到要求类再业务场景也是存在的，宁可你说我不知道，但是不能胡说。这种问题是可以提前准备的，它更接近于大模型底层的通用能力建设，这种我们认为是最容易评价的，这个其实我们现在是做了。

第二类是证明模型能力的评价，证明我的模型比其他的模型性能更有优势。其实我们做了大量的工作，但是目前看来和真正的业务场景相关性没有那么高。我说一下做的核心工作：首先会看影响大模型最终效果的因素有哪些；另外就是这个模型的参数，模型有很多超参数，设置不同的参数最终对模型的效果是怎么样的；另外就是模型本身有很多属性，刚才大家也说了，比如它是量化过的，实际参数可能不是原来的精度了，这种情况下还是同一个模型吗，评价是分开评价还是一起评价。另外这个模型部署在不同的芯片上、使用不同的框架、使用不同的卡、使用不同的推理引擎，都有可能导致模型的精度损失，它本质上是一个模型但是推理出来的结果是不一样的，这种情况我们怎么去做评价。

大模型本身是一个多步决策类的任务，那比如我前面某半部分效果好、后面某半部分效果不好，最终评估的是一个非常靠后的结果，在工业界一般不会直接去评价这么靠后的结果。所以我们一直在想能不能把评测前置，不去直接评价可理解的指标，目前我们在做这样子的尝试。然后我们做了大量的自动化测试，基于不同的 prompt、不同的 cot 去评测模型，得到差异非常大的结果。

然后大家可以想一想，在这个领域做评价和其他评价有很大区别。因为大模型是一个飞速发展的领域，那我们怎么可能指望一个相对固定的评价体系去做很好的评测呢。所以我认为这个评价体系如果要做的话，它应该是一个动态更新且更新频繁的。

然后就是大模型我要评价微调之后的效果怎么样。我们做过的测试是在第二阶段选取了几十个模型，微调后比较模型的效果，实际和第一阶段的相似度不太高。

58.com
Our model evaluation process is generally divided into two parts.
One is the training and inference stage, which measures hardware-related aspects, such as resource occupancy during training. The other is the model performance evaluation, which is roughly divided into three categories:

The first category is meeting requirements. For example, questions related to national sovereignty are red-line issues. If one such question is answered incorrectly, the large model will fail to be registered. Such requirements also exist in business scenarios. It's better to say "I don't know" than to give a wrong answer. These questions can be prepared in advance and are closer to the development of the large model's underlying general capabilities. We believe this category is the easiest to evaluate, and we have already done so.

The second category is evaluating the model's capabilities to demonstrate its advantages over other models. We have done a lot of work in this area, but it currently seems less relevant to real business scenarios. Let me mention some core work we have done: Firstly, we analyze the factors that affect the final performance of the large model. Secondly, we consider the model's parameters, including many hyperparameters, and how different settings affect the model's performance. Additionally, the

model itself has many attributes. For example, if it has been quantized, its actual parameters may not have the original precision. In this case, is it still the same model? Should we evaluate it separately or together? Furthermore, deploying the model on different chips, using different frameworks, cards, or inference engines may lead to a loss of model accuracy. Essentially, it is the same model, but the inferred results may differ. How do we evaluate in such cases?

The large model itself is a multi-step decision-making task. For example, if the first half of the model performs well but the second half does not, the final evaluation may focus on a relatively later result. In the industry, we generally do not directly evaluate such late results. Therefore, we have been thinking about whether we can advance the evaluation and avoid directly evaluating understandable indicators. Currently, we are making such attempts. We have also conducted a large number of automated tests, evaluating the model based on different prompts and chains of thought (cot), and obtained vastly different results.

You can think about it: evaluation in this field is very different from other evaluations. Since the large model field is rapidly developing, how can we expect a relatively fixed evaluation system to perform well? Therefore, I believe that if such an evaluation system is to be implemented, it should be dynamically updated and frequently updated.

Lastly, we need to evaluate the effectiveness of the fine-tuned large model. In one test, we selected dozens of models in the second stage and compared their performance after fine-tuning. The results were not very similar to those in the first stage.

空天

我汇报一下我们这个领域大模型的进展。我们现在深度合作的是阿里，在千万基础上提供了大概 200 多万条语料给大模型去训练空间科学领域。其实刚才各位同事提到的我们什么时机适合去发布模型，这个是我们一直没有把握准的问题。另外一个我们现在在语料库这块，我们其实也在做 AI ready 数据集，这块的问题是，我们具体把 AI ready 做到什么程度才能够让大模型在训练的时候更加高效。然后另外一个在整个评价方面现在主要靠我们信息的反馈跟模型的交互，这块我们还没有利用一个比较标准的评价工具。我们也非常希望能采用一些比较标准的通用化的工具，这样能更好的评价我们领域的模型。

Aerospace

I'd like to report on the progress of large models in our field. We are currently in deep collaboration with Alibaba, providing over 2 million additional corpus entries on top of the tens of millions to train large models in the field of space science. Actually, as colleagues have mentioned earlier, we have been uncertain about the right timing to release our model. Another issue we are facing is with our corpus. We are also working on creating an AI-ready dataset, but the question is, to what extent should we refine it to make large model training more efficient. Additionally, in terms of evaluation, we currently rely mainly on feedback from information and interaction with the model. We have not yet utilized a more standardized evaluation tool, and we really hope to adopt some standardized and generalized tools to better evaluate our domain-specific models.

张星洲

刚才各位老师讲了大模型的各种评价，在我看来其实是分为两大类：一类是对模型本身功能

性的评价，比如准确率、算法层面；第二类是对它真实运行的性能评价，更多偏向于计算机系统内存占用量、延迟。所以我想的是我们能不能先基于大类去分，然后从大类具体细分。比如先从功能去分，然后再分成不同场景、不同场景下的功能指标；如果从性能上，那可能是重延迟、功耗、吞吐等等，然后再往下细分。

Xingzhou Zhang

Just now, various evaluations of large models were discussed. In my opinion, they can be broadly divided into two categories: one is the functional evaluation of the model itself, such as accuracy and algorithmic aspects; the other is the performance evaluation of its actual operation, which is more focused on computer system metrics like memory usage and latency. So, I was thinking, can we first categorize based on these two broad categories and then subdivide them? For example, starting with functional categorization and then further dividing into different scenarios and functional indicators within those scenarios; for performance, it could be heavy on latency, power consumption, throughput, etc., and then subdividing further.

鲁海荣（联通）

我们这边也是在做一个模型评测的工作，也是偏向于应用这一块。在做的过程中，有一些主观题，它是一些开放性的问题，目前是通过裁判模型来做。通过裁判模型来做就会有一个问题，它会倾向于答案长的，然后还有可能跟先后顺序有关系，会有一些不确定性的答案。还有些情况下，裁判模型不一定能够就是标准答案，对于预测的和标准的它没法做一个准确的评判。

还有就是在做专业领域的数据时，因为专业领域对做这个数据的专业知识要求非常高，特别像医疗方面，做这种测试集就比较费时费力。训练出一个模型，怎么让这个模型的覆盖范围更全面，怎么去评判训练的模型。

推理的时候，因为目前模型都比较大，它对用户请求的时候并发量要求比较高，如果显卡用起来多的话，那硬件程度就比较高了。

Hairong Lu (China Unicom)

We are also working on model evaluation, with a focus on application. In the process, we encounter subjective questions, which are open-ended. Currently, we use a referee model for these. However, using a referee model poses issues, such as a bias towards longer answers and potential dependency on the order of answers, leading to uncertain results. In some cases, the referee model may not provide the standard answer, making it difficult to accurately evaluate predictions against standards.

Moreover, when creating data for specialized fields, the expertise required is very high, especially in areas like healthcare. Creating such test sets is time-consuming and labor-intensive. After training a model, how do we ensure its comprehensive coverage and evaluate its training effectiveness?

During inference, since current models are relatively large, they have high concurrency requirements for user requests. If multiple GPUs are used, the hardware requirements become significant.

浪潮

我们公司大模型的评价也是在起步阶段，它的准确性评价、价值评价能否产生我们想要的价

值。大模型的榜单到底能不能产生对我们有价值的输出。其次在性能方面，因为模型本身也会用到算力，所以性能是我们非常关注的问题，目前大模型最主要的性能指标就是它的吞吐量，现在模型都特别大，跑它的时候会用到特别多的算力，有没有可能把模型所占用的算力和它最后产生的价值联系起来，类似于你用了多少算力，我就期待着能够产生多少价值。

Inspur

Our company is also in the initial stages of evaluating large models, focusing on whether their accuracy and value evaluations can deliver the outcomes we desire. Can rankings of large models actually produce valuable outputs for us? Secondly, in terms of performance, since models themselves consume computing power, performance is a major concern for us. Currently, the primary performance indicator for large models is their throughput. Given the sheer size of these models, running them requires substantial computing power. Is it possible to establish a connection between the computing power a model consumes and the value it ultimately generates? Something akin to, "the more computing power you use, the more value you can expect in return."

詹剑锋研究员：新的指标？

刘思：对，能不能有个方法能将性能和它产生的价值联系起来，用最少的算力产生更多有价值的输出。或者说哪怕我多堆点算力，能不能给我多点价值的输出。

Professor Jianfeng Zhan: New indicators?

Si Liu: Yes, can we develop a method to link performance with the value generated, ensuring more valuable outputs with minimal computing power? Or alternatively, if I invest more computing power, can I expect proportionately more valuable outputs?

郭晶晶（不清楚名字对不对）：工作组最后大的目标是要形成一套系统吗？

詹剑锋研究员：包括理论方法、一些工具、数据集、标准都是。在智能方面我们要形成一个智能评价学，大语言模型是智能的一种嘛，刚才 58 同城老师讲的，可能我们也有一些前瞻性的评价。一方面有学术的影响力，另一方面给工业界一些东西。有些工具学术界可以拿来做研究，通过大家共享，比如您这边有一些数据集可以拿出来、一个平台拿出来。它一定是个具体的东西，刚才讲的有理论、有方法。我们现在是一个线下活动，后面有些线上的、定期的活动来推动。

Jingjing Guo (unclear if the name is correct): Is the ultimate goal of the working group to develop a comprehensive system?

Professor Jianfeng Zhan: It encompasses theoretical methods, tools, datasets, and standards. In the realm of intelligence, we aim to establish an evaluative science of intelligence, and large language models are a form of intelligence, as mentioned by the representative from 58.com earlier. We may also incorporate forward-looking evaluations. This will have both academic impact and provide something tangible for the industry. Some tools can be used by the academic community for research and shared among everyone. For instance, you might contribute datasets or a platform. It must be a concrete entity, encompassing both theory and methodology, as discussed earlier. We are currently conducting offline activities, but there will be follow-up online and periodic activities to drive this forward.

汤飞：我感觉生成一个 benchmark，类似于榜单让大家用起来。现在有 Ollama 一个工具，它能很方便的构建一个本地的 AI 模型。能不能出这样一个榜单，Ollama 上所有的模型用不同

的卡、不同的精度都测一遍，发布一个榜单。

詹剑锋研究员：肯定要有榜单，你像我们搞得那个 OpenMeter，我们把开源的两亿多个项目，不同领域都有一个榜单，可以促进工作宣传。

Fei Tang: I feel like creating a benchmark, similar to a ranking, for everyone to use would be beneficial. There's a tool called Ollama that conveniently builds local AI models. Could we have a ranking where all the models on Ollama are tested using different GPUs and precisions, and then publish the results?

Professor Jianfeng Zhan: Definitely, there will be rankings. Like our OpenMeter, which ranks over 200 million open-source projects across different domains, promoting work and awareness.

四、总结

Conclusion

詹剑锋研究员发表总结讲话

1、 智能的角度的评价学。结合工业界的需求，从智能的角度，本身在评价的方面，这一块是可以产出的。

2、 数据质量。从数据集的构造方面，其实整个领域现在没什么特别的方法。这是教与学的问题：本质上训练就是教它们，我怎么教它、教它什么样的知识；另一个就是学。刚才讲了很多每个行业怎么来出题，其实我觉得本质上就是教育学的问题，这些方法的东西我们肯定是要研究它的。包括数据质量，就是一些基础的方法。

3、 其实方法处理我们也讲过，一个真实世界的评价系统到一个理想化的，早期是仿真，它能模拟不同的情况，但是它的精度可能有差异，它没办法完全考虑所有的细节。到实际的一个评价系统，刚才也有讲评价完了以后到下游可能不是这样，整套的方法我们也可以做一些工作，从方法论的角度。具体我觉得有几个东西是可以做的，比如：AI 模拟器、数据集的测试工具。

4、 另外有一些优先的领域，我觉得 AI for Science 是一个重要的领域，材料和天文两部分作为重点。

5、 行业应用和通用应用之间的关系（汤飞介绍的工作）。

6、 关键的因素对结果的影响。这个涉及到最后的评价的细则问题，也是很重要的问题。

7、 国产芯片和英伟达的差距，生态方面的一些建议。

Professor Jianfeng Zhan Delivers a Concluding Speech

1、 **Evaluatology from an Intelligence Perspective**: Combining the needs of the industry, evaluations from an intelligence perspective can indeed be produced in this area.

2、 **Data Quality**: In terms of dataset construction, there are currently no particularly distinctive methods in the entire field. This is an issue of teaching and learning: essentially, training involves teaching AI models, how to teach them, and what knowledge to impart. The other aspect is learning. Much has been discussed about how each industry can create test questions, but I believe this essentially boils down to issues in pedagogy. We must definitely study these methodological aspects, including fundamental approaches to data quality.

3、**Methodological Processing**: We have also discussed the transition from a real-world evaluation system to an idealized one. Early stages involve simulation, which can mimic different scenarios but may vary in accuracy and cannot fully consider all details. When transitioning to an actual evaluation system, as mentioned earlier, the results after evaluation may differ in downstream applications. We can make contributions from a methodological perspective. Specifically, there are several areas where we can take action, such as developing AI simulators and dataset testing tools.

4、**Priority Areas**: I believe AI for Science is an important area, with a focus on materials and astronomy.

5、**Relationship Between Industry-Specific and General Applications** (work introduced by Tang Fei).

6、**Impact of Key Factors on Results**: This relates to the final evaluation guidelines, which are also crucial.

7、**Gap Between Domestic Chips and NVIDIA, and Suggestions for Ecosystem Development**.