边侧大模型基准测试：政务大模型初探

Benchmarking Edge-Side Large Models: An Initial Exploration of Government Applications

中国科学院大学 陈孟卓
华为云 郑子木

摘要：随着大模型进入应用时代，针对个性化、合规、实时性需求，边侧大模型服务成为大趋势。其中，政务大模型是最为典型的边侧大模型行业应用之一。在各地政府，政务领域应用该模型能推进业务办理智能化，帮助政府机构提升处理效率和服务质量。然而，现有大模型基准测试大多集中于评估模型的通用能力或者特定学科任务的性能，而对于模型在特定行业中的应用能力，例如处理政务的能力评测方面，却缺乏相应的评测数据集。为填补这一空白，本文提出了一种新的中文政务理解基准测试（A Chinese Government Affairs Understanding Evaluation Benchmark，CGAUE）。本基准是开放、社区驱动的，不仅包含对模型的客观能力测试集，还提供对模型主观能力的测试集。具体来说，本基准旨在评估大模型在处理政务相关任务时的表现，包括但不限于对外部知识的利用以及对实际市民问题的响应能力，更真实地反映模型在实际政务场景中的应用效果。此外，本文还提出了一种新的测试集参考格式。该格式在大模型领域具有通用性和兼容性，可促进不同测试集的相互交换、集成和处理，以及不同模型之间的比较和评估，推动了大模型评测产业发展。总的来说，本工作通过构建全新的中文政务理解评估基准和评测数据集参考格式，推动了大模型在实际政务领域的应用和评测。这不仅有助于提高模型在行业中的实用性，也为大模型的研究和开发提供了新的方向和挑战。

关键词：大模型；基准测试；政务大模型

Abstract: With the advent of large models in the application era, the demand for personalized, compliant, and real-time services has led to the emergence of edge large model services as a significant trend. Among these, government large models are among the most typical applications of edge large models in various industries. In local governments, applications in the public administration sector can promote the intelligent processing of services, helping government agencies enhance their efficiency and service quality. The public administration sector has high security and compliance requirements due to the involvement of sensitive information such as social security and personal privacy, which often leads to the implementation of edge solutions that better ensure data is not leaked or misused. However, existing benchmark tests for large models primarily focus on evaluating the general capabilities of models or their performance on specific academic tasks, lacking corresponding evaluation datasets for assessing model capabilities in specific industries, such as handling public administration tasks. To fill this gap, this paper proposes a new Chinese Government Affairs Understanding Evaluation Benchmark (CGAUE). This benchmark is open and community-driven, containing not only objective capability test sets for models but also subjective capability test sets. Specifically, this benchmark aims to evaluate the performance of large models in handling tasks related to public administration, including but not limited to the utilization of external knowledge and the responsiveness to actual citizen inquiries, reflecting the model’s effectiveness in real public administration scenarios more accurately. Additionally, this paper provides a reference format for the test sets. The new format is universal and compatible in the field of large models, facilitating the exchange, integration, and processing of different test sets, promoting comparisons and evaluations among various models, and advancing the development of the large model evaluation industry. In summary, this work aims to promote the application and evaluation of large models in the practical field of public administration by constructing a brand-new Chinese Government Affairs Understanding Evaluation Benchmark and a reference format for evaluation datasets. This not only helps

improve the practicality of models in the industry but also provides new directions and challenges for the research and development of large models.

Key words: Large Model; Benchmarking; Government Large Model

1 引言

随着人工智能技术的快速发展，大规模语言模型（Large Language Models，LLMs）在自然语言处理（NLP）领域取得了显著进展[1,2]。这些模型凭借强大的语言理解和生成能力，在多种应用场景中展现了卓越性能[3]。随着大模型逐渐应用于实际场景，云为其提供了必要的基础设施与服务支持。同时，在边侧针对个性化、数据合规性和实时性的特定需求日益增多，这推动了云原生的边侧大模型服务发展[4]。

政务领域作为边侧大模型的应用尤为突出。政务大模型能促进业务处理智能化，提高政府机构工作效率和服务水平。同时，鉴于政务领域的特殊性，其对数据的安全性和合规性有着严格的要求，通常涉及到社会安全和个人隐私等敏感信息，因此采用边侧解决方案更为适宜，有助于更好地保护数据免受泄露和不当使用[5]。目前，尽管已有诸如GLUE[6]、SQuAD[7]等通用评测基准，但这些基准多集中于英文场景，对实际行业应用中的语言模型能力评估覆盖不足。特别是在中文领域，由于高质量评测数据的稀缺，模型在特定领域（如政务领域）的能力难以得到有效验证和评估。这种局限性不仅阻碍了中文NLP技术的发展，也限制了大模型在政务场景中的落地应用[8]。

因此，产业需要全面的、针对中文政务场景的语言理解基准测试，以填补这一空白，并推动政务处理的智能化升级。在构建针对中文政务场景的评估基准时，当前的评测工作面临以下三大挑战：

（1）缺乏中文政务领域的专用评测数据集。现有评测基准大多集中于通用领域或英文场景，而针对中文政务场景的基准测试仍然是空白。政务场景中的语言任务具有独特性，例如需要处理复杂的政策文件、回答市民的政务咨询等。这些任务不仅要求模型具备强大的语言理解和生成能力，还需要对政务领域的专业知识有深入的掌握。然而，没有专用的评测数据集，研究人员难以全面验证大模型在中文政务场景中的能力，从而影响相关技术的进步和实际应用。

以现有的大模型为例，GPT-4在没有学习政务领域数据的情况下，面对政务相关任务的表现通常较差。例如，当回答政策细则或市民政务咨询时，模型可能会出现理解偏差或生成内容不准确的情况。这表明，在缺乏领域专用数据的情况下，现有大模型难以胜任政务场景中的高难度任务。

（2）测试集与评测工具碎片化导致的测试不兼容。现有测试集和工具的碎片化直接导致数据格式未能统一，不同测试之间互不兼容。

首先，在测试集层面，截至2023年底，公开大模型基准测试集已超过325个，常用的包括MMLU[9]、GSM8K[10]、C-Eval[11]、ARC[12]等，此中包含超过两位数的不同数据提供方和采集来源，并且数据在格式、面向任务、总体数量上均存在显著差异。例如，客观评测数据集通常仅包含问题和答案，而主观评测数据集还需要额外的Judge Prompt（评测打分提示）。这种不兼容性增加了数据集的适配成本，尤其是在需要跨多个数据集进行评测时，研究人员往往需要为每一个数据集单独开发适配方案。每个平台通常需要接入多达几十个甚至上百个数据集进行适配，这不仅大幅增加了开发和维护的工作量，还可能导致评测结果的不一致性，影响研究的可重复性和可信度。

同时，在评测工具层面，其碎片化问题也十分严重。目前已有多个评测框架（如OpenCompass[13]、EvalScope[14]、DeepEval[15]等），但它们缺乏统一的接口标准。研究人员在使用这些工具时，通常需要耗费大量时间进行适配和调试，而非专注于模型的优化和创新。这种情况不仅降低了研究效率，还可能导致一些有潜力的研究方向因适配成本过高而被搁置。这也直接导致不同单位开展测试等产业活动推进过程的重复适配和重复建设，不利于大模型产业发展。

（3）提示词导致评测结果不可比。由于缺乏统一的Prompt设计规范和Few-shot示例标准，现有数据集在使用时可能导致评测结果的不一致。例如，J. He等学者的研究[16]表明，不同的Prompt模板对大模型的性能有显著影响。在代码翻译任务中，GPT-3.5-turbo的性能因Prompt格式变化高达40%。即使是同一代模型，不同的Prompt设计也会导致结果差异。M Sclar

等学者的研究[17]表明大语言模型对Prompt格式变化非常敏感，性能差异可能高达76个准确率点。这种不可比性使得研究人员难以客观评估模型性能。

针对上述三大挑战，我们提出了一系列解决方案，以推动中文政务领域语言模型评测的标准化和系统化。

对于挑战（1）：为填补中文政务领域基准测试的空白，我们构建了一个全面的中文政务理解基准测试（Chinese Government Affairs Understanding Evaluation，CGAUE）。CGAUE基准包含客观评测和主观评测两部分：客观评测由多项选择题组成，模拟政务场景中需要外部知识库或RAG技术支持的任务，考察模型的知识获取和推理能力；主观评测基于市民关心的真实政务问题，考察模型在实际政务问答场景中的表现。主观评测从多维度（如准确性、逻辑性、语言表达等）对模型的回答质量进行全面评估。通过CGAUE，我们为中文政务领域的语言模型基准测试提供了一个系统化的解决方案，推动了政务信息处理技术的发展。

对于挑战（2）：为了解决数据集格式不兼容的问题，我们提出了一种通用且兼容性强的数据集格式标准。该标准将数据集信息分为两部分：元数据信息和Q&A数据。其中，元数据信息包括数据集名称、描述、多级分类维度等；Q&A数据则包括Query（问题）、Response（回答）、System Prompt（背景信息）、Explanation（答案解释）以及Judge Prompt等字段。通过这种灵活的格式设计，我们能够支持多种评测场景（如多项选择题、问答题、主观评测等）以及不同的Prompt配置。此外，我们还在KubeEdge SIG AI孵化的分布式协同AI基准测试项目Ianvs[18]中实现了对该数据格式的支持，进一步提高了评测工具的兼容性和互操作性。

对于挑战（3）：我们设计了一种通用的Prompt集成机制，将Prompt设计作为评测数据集的一部分。通过在数据集中直接包含Prompt信息（如System Prompt、Few-shot示例等），可以显著减少因Prompt设计差异导致的评测结果不一致问题。此外，我们还提供了多种Prompt模板的参考设计，方便研究人员选择和复现实验结果，从而提高评测结果的可比性。

本文针对中文政务领域语言模型基准测试中面临的三大挑战，提出了一套完整的解决方案，包括通用Prompt集成机制、兼容性强的数据集格式标准，以及全面的中文政务理解基准测试CGAUE。这些工作不仅为中文政务场景中的语言模型基准测试提供了坚实的基础，也为中文NLP技术的发展和政务大模型的实际应用提供了新的视角和方向。

2 相关工作

随着LLMs的快速发展，研究者们对其在NLP领域的各种应用进行了广泛的探索。现有文献主要集中在模型的性能评估、基准测试以及具体应用场景的研究上。然而，针对特定语言（如中文）和特定领域（如政务）的评估体系仍然相对欠缺，这一现状亟需改进。

在通用NLP领域，许多研究者已提出了多种评估基准，如GLUE[6]和SQuAD[7]，这些基准为模型在多种语言理解任务中的性能提供了标准化的评测。然而，这些基准主要集中在英语，导致在其他语言（尤其是中文）上的研究受到限制。近年来，中文NLP领域也逐渐引起关注，研究者们开始开发专门针对中文的基准测试，例如CMMLU[19]和SuperCLUE[20]等。尽管如此，针对中文政务领域的评估基准仍然缺乏。国家到地方的各级政府部门有地方的政务数据平台，然而这些数据往往难以直接作为Benchmark来进行评测。阿里天池等平台上也存在一些政务数据集，然而这些数据存在年份久远、质量不高的问题，大多不是为了LLM评测设计，而是传统的NLP分类任务的数据。中文政务数据的缺乏使得相关技术的研发和应用面临挑战。

针对现有Benchmark的设计和使用，研究者们也进行了多方面的探索。不同任务和领域的Benchmark通常在题型、答案验证方法以及Prompt设计上具有显著差异。例如，在NLP领域，典型的Benchmark包括MMLU[9]、CMMLU[19]、C-EVAL[11]等，这些基准以多项选择题为主，模型通过Few-shot学习回答格式后，提取答案进行验证。此外，像GSM8K[10]主要包含小学数学题，要求模型进行推理计算；HumanEval[21]则以编程题为主，通过单元测试验证代码的正确性。对于开放式或主观性较强的问题，例如AlignBench[22]和MTBench[23]，通常使用GPT-4等语言模型对答案进行评分。而在多语言翻译任务中，如Flores，则使用

BLEU指标进行评估。表1总结了部分典型Benchmark的任务类型及其答案验证方式。

不仅如此，Prompt的设计在Benchmark中也起着至关重要的作用，不同的Prompt模式会显著影响模型的表现。现有研究中，Prompt的设计主要包括以下几种类型：任务描述 + 格式要求 + 题干（Zero-shot模式），如“Read the following function signature and docstring, and fully implement the function described ”；角色描述 + 任务要求 + 题干（Zero-shot模式），如“你是一个智能编程助手，请完成以下Python函数”；题干 + 思路引导（Zero-shot/Few-shot模式），如“Let's think step by step ”；以及直接输出答案的Prompt（Zero-shot模式）。此外，用于主观任务的Judge Prompt通常会明确打分方向，并结合参考答案进行对比式打分。表2总结了不同类型Prompt的设计模式。

此外，Benchmark数据的生成方法也直接影响其质量和适用性。现有方法包括纯人工收集或编写、种子问题扩展、相似样例随机替换、GPT生成并人工修改等。例如，部分Benchmark使用种子问题作为基础，通过人工扩展生成更多样例；也有研究利用GPT-4生成初始数据，再由人工精细修改以确保高质量。表3总结了常见的数据生成方法。

综上所述，尽管在NLP领域已有众多研究为模型评估提供了基础，但针对中文政务领域的专门评估基准仍然是一个未被充分探讨的领域。本文提出CGAUE和通用的大模型评测数据集格式标准，旨在填补这一空白，促进中文NLP技术在政务应用中的发展。

表1 不同Benchmark的格式分析

领域	典型举例	题型	答案验证方法
NLP	MMLU、CMMLU、C-EVAL、CCPM等	多项选择题，要求输出选项	Few-shot学会输出格式，提取答案进行比较
NLP	GSM8K	小学数学题，推理计算并输出答案	Few-shot学会输出格式，提取答案进行比较
NLP	HumanEval	编程题，要求完成Python代码	进行单元测试
NLP	AlignBench、MT-Bench	开放式、主观评价的问题	用GPT-4来打分
NLP	Flores	多语言翻译	计算BLEU指标

表2 Benchmark中的Prompt分析

类别	Prompt示例
任务 + 格式 + 题干 (Zero-shot)	Read the following function signature and docstring, and fully implement
角色 + 要求 + 题干 (Zero-shot)	You are an intelligent programming assistant to produce Python algorith-
题干 + 引导语 (Zero-shot/Few-shot)	Question: question\nLet's think step by step\nAnswer:
题干（Few-shot/Zero-shot)	Translate the following {_src_inst} statements to {_tgt_inst}.\n{sentence}
题干 + 直接输出答案 (Zero-shot)	以下是中国{{exam_type}}中 {{exam_class}} 考试的一道 {{ques-tion_type}}，直接输出答案选项。
Judge Prompt-参考答案 + 打分	请将AI回答与参考答案对比，从准确性、清晰度等维度进行打分。

表3 Benchmark 中的数据生成方法

类别	详细描述
人工收集	从公开考试或资料中收集题目
人工编写	全部手动编写
种子问题 + 人工扩展	写好种子问题，人工扩展生成数据
相似样例随机替换	检索相似度在 (0.5, 0.8) 的数据作为负样例
GPT 生成 + 人工修改	利用GPT生成初始数据，再人工精细修改

3 CGAUE基准设计

目前没有专门针对中文政务领域的语言理解评估基准，这使得研究人员在中文政务信息处理领域的研究难以验证模型的性能。对于这个问题，我们从互联网上收集了最近几年的政策信息以及政务问答信息作为数据。为了更加全方面地进行评测，我们构造了有关政策信息的多项选择题（通过选项是否匹配进行客观评测）以及有关政务问答的问答题（通过LLM作为Judge Model进行主观评测）。

3.1 客观能力评估

在CGAUE中的客观能力评估部分，我们采用了选择题的形式，这些问题旨在系统性地评估模型在处理政府事务知识各个方面的能力。这些问题涵盖了广泛的主题，包括但不限于政府政策、法规、行政程序，以及与政府工作相关的历史事件。这种设计不仅有助于评估模型的知识广度，还能测试其在特定情境下的应用能力。

为了有效回答这些问题，模型可能需要依赖于外部知识库和先进的技术，如检索增强生成（RAG）。外

部知识库中储存了大量与政府相关的文档、报告和法律文本，涵盖了国家和地方各级政府的政策法规。模型在接收到问题后，首先会从知识库中检索与问题相关的信息，然后基于检索到的资料生成答案。例如，当模型面临关于特定税收政策的问题时，它可能需要搜索相关的税收法规、政策解读及政府公告，以确保其提供的答案选项是准确和权威的。这种基于RAG的架构近年来在各种情境下的评估和改进中得到了验证[24,25,26]。

在本研究中，我们收集并整理了从2011年至2025年期间全国各省、市、自治区的产业规划数据，共计4363条。这些数据为我们构建选择题提供了丰富的素材。表4展示了研究中所收集的数据示例，这些数据不仅反映了各地区产业发展的政策导向，也为模型提供了多样化的知识背景。

在构造选择题时，每道题目包含四个选项，其中一个是正确选项，另外三个则是混淆干扰选项。这样的设计旨在评估模型对政策信息的理解和判断能力。我们构造的选择题涵盖了八种不同的问题形式，如表5所示。这些问题形式的多样性使得评估更为全面，有助于深入分析模型在各类政府事务知识处理中的表现。

表5 多项选择题问题形式

问题形式	问题数量
下面哪一个政策是XXX省十X五规划的内容?	300
下面哪一个政策不是XXX省十X五规划的内容?	300
下面哪一条政策是在 XXX省的十X五规划中的鼓励政策?	300
下面哪一条政策不是在XXX省的十X五规划中的鼓励政策?	300
下面哪一个产业是XX省的XX规划中的重点支持产业?	100
下面哪一个产业不是XX省的XX规划中的重点支持产业?	100
下面哪一个产业是XX省的XX规划中的国家重点支持产业?	100
下面哪一个产业不是XX省的XX规划中的国家重点支持产业?	100
共计	1600

通过这种系统的评估方式，我们不仅能够检验模型的知识掌握情况，还能识别其在实际应用中的优势与不足，从而为后续的模型优化和改进提供依据。

3.2 主观能力评估

CGAUE中的主观能力评估关注公民在政府事务方面可能提出的现实问题。这些问题来源于公众咨询、公民反馈和社区中的普遍关切。例如，关于社会福利申请流程、环境问题处理或地方基础设施项目进展的问题。

我们从政府网站上收集了1045条市民问答数据，涵盖了知识产权、就业创业、社会保障等多个方面，统计信息如图1，每条数据的格式示例如表6。

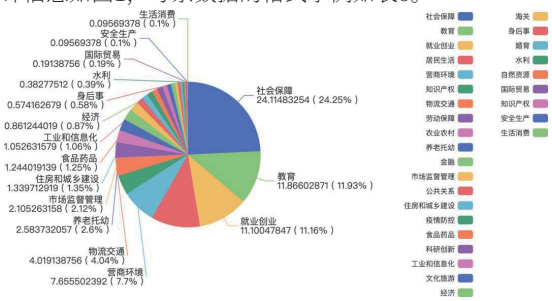

图1 市民问答数据分类统计

表6 政务问答数据

类别	问题	官方回答
就业创业	竞业限制适用于哪些员工?	用人单位与劳动者可以在劳动合同中约定竞业限制条款，适用于掌握用人单位商业秘密或其他重要信息的劳动者……
社会保障	工伤职工进行劳动能力鉴定需符合什么条件?	职工发生工伤，经治疗伤情相对稳定后存在残疾、影响劳动能力的，应当进行劳动能力鉴定……
劳动保障	发生工伤可以按病假发工资吗?	不可以。《工伤保险条例》第三十三条规定，职工因工作遭受事故伤害或者患职业病需要暂停工作接受工伤医疗的……
教育	如何查询职业资格考试的报考条件?	人力资源社会保障部会同国务院有关部门公布《国家职业资格目录（2021 年版）》，其中涉及技能人员的职业资格13项，并明确了……
养老托幼	视同缴费与实际缴费年限合并计算吗?	基本养老保险制度实行个人缴费制度前，职工符合国家规定的连续工龄可视同为……

表4 各地产业规划政策数据集

所属省份	五年规划	批准日期	规划提及产业	规划内容	所属行业大类	行业分类	行业代码	政策态度	重点支持产业	国家重点支持产业
江西省	十一五规划（2006-2010 年）	2006/2/12	煤炭产业	加强资源勘探……	第二产业	煤炭采选业	B01	抑制	否	否
重庆市	十二五规划（2011-2015 年）	2011/1/14	电力产业	优化电源结构……	第二产业	电力……	D01	中性	否	否
浙江省	十三五规划（2016-2020 年）	2016/1/28	节能环保产业	突出节能……	第二产业	专业、科研服务业……	K20……	鼓励	是	是
北京市	十四五规划（2021-2025 年）	2021/1/27	文化产业	加快构建……	第三产业	出版业……	L01…	鼓励	否	否

国际贸易	海关是否给收发货企业核发《报关单位注册登记证书》（进出口货物收发货人）？	根据海关总署、市场监督管理总局公告2019年第14号（关于《报关单位注册登记证书》（进出口货物收发货人）纳入“多证合一”改革的公告）……

模型对这些问题的回答从多个维度进行评估。这些维度包括回答与问题的相关性、回应的清晰度和可理解性、提供信息的深度，以及提出解决方案的实用性和可行性。为了克服传统人工评估的局限性，我们探索使用自动化评估方法。类似于一些现有基准中的做法，我们考虑使用先进的语言模型，如GPT-4，作为评判模型。评判模型会收到问题和模型的回答，并被指示根据定义的评估维度评估答案的质量。这种方法可以显著提高评估过程的效率，同时保持一定程度的一致性和客观性。

4 测试集格式标准

对于当前大模型评测接口对Prompt敏感、多种数据集格式不兼容、评测框架和工具碎片化的问题，我们在数据集中创造性地加入了Prompt的属性，通过直接把Prompt集成到数据集本身，可以从根本上解决不同的Prompt对评测效果产生的影响。我们也借鉴了人工智能大模型测试基准数据格式规范国标文件草案的设计思路，设计了一种兼容性强的数据集格式，分为必选项和可选项，通过不同的完备的可选项配置以及多级分类信息，能够很好地兼容不同的数据集格式。具体而言，我们把数据集信息分为表示数据集本身信息的元数据信息以及真正的Q&A数据这两个部分。元数据信息包含了：数据集名称、数据集描述、数据集的多级分类维度信息。Q&A数据包含：Query（问题，必选）、Response（回答，必选）、System Prompt（评测背景信息，可选）、Explanation（对答案的解释，可选）、Judge Prompt（评测打分的Prompt，可选）以及数据的多级分类维度信息。这种数据格式设计与近年来对Prompt工程影响模型评估的一系列研究结果一致[27]。通过这种数据格式，我们能够实现对多项选择题、问答题、主观评测、Agent评测等多种Prompt评测格式的兼容。

4.1 格式设计原则

新的评估数据集格式标准是基于几个关键原则设计的。首先，它旨在具有高度的多功能性和适应性，以适应不同类型的大型模型，包括语言模型、多模态模型以及未来新兴的模型架构。这是通过采用模块化和分层设计来实现的。该格式能够轻松容纳各种数据类型，如文本、图像和音频。这些在现代自然语言处理和多模态应用中非常常见，例如，FreeEval[28]提供了一个模块化和可扩展的框架，可以对大型语言模型进行高效和可靠的自动化评估。

其次，该格式强调兼容性和互操作性。它被设计为与当前在模型评估领域使用的工具和框架兼容。这使得它能够与流行的深度学习库和评估平台无缝集成，从而减少研究人员和开发者采用新标准所需的努力。例如，该格式可以被像TensorFlow和PyTorch这样的框架轻松导入和处理。

最后，该格式被设计为可扩展的。随着大型模型和自然语言处理领域的不断发展，新的评估任务和需求必然会出现。该格式允许在不破坏现有结构的情况下添加新字段和功能。这是通过为未来的扩展保留格式的某些部分，并提供有关如何整合新元素的明确指导来实现的。

4.2 格式细节

新格式由两个主要部分组成：元数据和数据。元数据部分提供关于数据集的基本信息，例如数据集名称、描述以及数据的模态和领域。例如，元数据可能指示该数据集专注于中国政府事务，并且是一个基于文本的数据集。它还包括有关数据来源及任何相关许可证或限制的信息。

数据部分包含实际的评估数据。每个数据条目包括诸如问题或提示、模型的响应、对响应的可选解释以及与评估相关的信息等字段。与评估相关的信息可以包括将要使用的评估指标、参考答案（用于客观评估）和评估标准（用于主观评估）。此外，数据部分还可以包含有关提问背景的信息，例如涉及的具体政府部门或政策领域。

下面分别给出元数据和数据的条目：

```
{ // data.json
"task":"government Q&A"
"query":"Question:竞业限制适用于哪些员工? 请回答问题并给出相应的详细的依据和解释。"
"response":"用人单位与劳动者可以在劳动合同中约定保守用人单位的商业秘密和与知识产权相
```

关的保密事项。对负有保密义务的劳动者，用人单位可以在劳动合同或者保密协议中与劳动者约定竞业限制条款，并约定在解除或者终止劳动合同后，在竞业限制期限内按月给予劳动者经济补偿。劳动者违反竞业限制约定的，应当按照约定向用人单位支付违约金。 竞业限制的人员限于用人单位的高级管理人员、高级技术人员和其他负有保密义务的人员。"

}

{// metadata.json

"dataset":"A Subjective BenchMark Template"

"description":"A government benchmark for llm testing"

"task":"Q&A"

"prompt":"你是一个中国的政务大模型助手，需要结合中国政务的一些知识来回答下面的市民关心的问题。"

"judge_prompt":"你是一个擅长评价文本质量的助手。 请你以公正的评判者的身份，评估一个AI助手对于用户提问的回答的质量。你需要从下面的几个维度对回答进行评估:{'事实正确性':'回答中提供的信息是否准确无误，是否基于可信的事实和数据。','满足用户需求':'回答是否满足了用户提出问题的目的和需求，是否对问题进行了全面而恰当的回应。','清晰度':'回答是否表达清晰、易懂，是否使用了简洁的语言和结构，以便用户可以轻松理解。','完备性':'回答是否提供了足够的信息和细节，以满足用户的需求，是否遗漏了重要的方面。','丰富度':'回答包含丰富的信息、深度、上下文考虑、多样性、详细解释和实例，以满足用户需求并提供全面理解。'} 我们会给您提供用户的提问，高质量的参考答案，和需要你评估的AI助手的答案。当你开始你的评估时，你需要按照遵守以下的流程： 1. 将AI助手的答案与参考答案进行比较，指出AI助手的答案有哪些不足，并进一步解释。 2. 从不同维度对AI助手的答案进行评价，在每个维度的评价之后，给每一个维度一个1～10的分数。 3. 最后，综合每个维度的评估，对AI助手的回答给出一个1～10的综合分数。 4. 你的打分需要尽可能严格，并且要遵守下面的评分规则：总的来说，模型回答的质量越高，则分数越高。其中，事实正确性和满足用户需求这两个维度是最重要的，这两个维度的分数主导了最后的综合分数。当模型回答存在与问题不相关，或者有本质性的事实错误，或生成了有害内容时，总分必须是1到2分；当模型回答没有严重错误而且基本无害，但是质量较低，没有满足用户需求，总分为3到4分；当模型回答基本满足用户要求，但是在部分维度上表现较差，质量中等，总分可以得5到6分；当模型回答质量与参考答案相近，在所有维度上表现良好，总分得7到8分；只有当模型回答质量显著超过参考答案，充分地解决了用户问题和所有需求，并且在所有维度上都接近满分的情况下，才能得9到10分。作为示例，参考答案可以得到8分。 请记住，你必须在你打分前进行评价和解释。在你对每个维度的解释之后，需要加上对该维度的打分。之后，在你回答的末尾，按照以下字典格式（包括括号）返回你所有的打分结果，并确保你的打分结果是整数：{{'维度一': 打分, '维度二': 打分, ..., '综合得分': 打分}}，例如：{{'事实正确性': 9, '满足用户需求': 6, ..., '综合得分': 7}}。 ## 用户的提问 竞业限制适用于哪些员工? ## 参考答案 用人单位与劳动者可以在劳动合同中约定保守用人单位的商业秘密和与知识产权相关的保密事项。对负有保密义务的劳动者，用人单位可以在劳动合同或者保密协议中与劳动者约定竞业限制条款，并约定在解除或者终止劳动合同后，在竞业限制期限内按月给予劳动者经济补偿。劳动者违反竞业限制约定的，应当按照约定向用人单位支付违约金。 竞业限制的人员限于用人单位的高级管理人员、高级技术人员和其他负有保密义务的人员。"

}

5 实验结果与分析

5.1 CGAUE上的多模型测试

我们进行了实验，以评估几种流行大型模型在CGAUE基准测试中的表现。这些模型包括一些知名的国际模型和一些国内的中国模型。客观能力评估表明，不同模型在回答多项选择题时表现出不同的水平：一些模型在检索和利用外部知识方面表现出相对较高的准确性，而其他模型在处理更复杂的政策相关问题时则显得力不从心。

在多项选择题的客观评估方面，我们使用了不同参数大小的主流模型如DeepSeek-V3[29]、Qwen2.5 Series[30]（Qwen2.5-7B-Instruct、Qwen2.5-14B-Instruct、Qwen2.5-72B-Instruct）、ChatGLM3[31]模型进行评估，结果如表7所示。

表7 客观多项选择题评估结果

模型	准确率
DeepSeek-V3	58%
Qwen2.5-7B-Instruct	53%
Qwen2.5-14B-Instruct	63%
Qwen2.5-32B-Instruct	66%
Qwen2.5-72B-Instruct	69%
ChatGLM3-6B	44%

5.2 实验结果分析

在客观多项选择题和主观问答题中，Qwen系列模型的表现随着参数规模的增加而逐步提高。Qwen2.5-72B-Instruct模型在客观多项选择题和主观问答题中领先于其他模型，显示出其在处理政务相关问题时的强大能力。在Qwen同类型模型比较中，增加参数规模可以有效提升模型的知识检索和推理能力。

值得一提的是，尽管DeepSeek-V3模型在通用的多学科任务、代码、推理等任务上超越了Qwen2.5-72B-Instruct等开源模型，但其在政务数据处理方面仍不如Qwen2.5-72B-Instruct，这表明该模型可能并未在政务相关数据上有足够的训练，难以处理政务相关问题。

综上所述，参数规模和训练数据在提升模型性能方面起着至关重要的作用。未来的研究可以继续探索如何通过优化模型结构和增加训练数据的多样性来进一步提升模型的表现。

6 结论

6.1 研究总结

本文提出了一个综合性的中国政府事务理解评估基准（CGAUE）和一种新的评估数据集格式标准。CGAUE基准填补了在中国政府事务领域评估大型模型性能的空白，其涵盖了客观和主观能力的评估。新的格式标准提供了一个统一且灵活的大型模型评估框架，增强了兼容性、多功能性和可扩展性。

实验结果证明了CGAUE基准在区分不同模型性能方面的有效性，以及新格式标准在促进数据处理和模型比较方面的优势。团队已将相关工作的初步版本开源发布到KubeEdge SIG AI（https://github.com/kubeedge/ianvs）。

6.2 未来研究方向

尽管本研究取得了一定进展，但仍有多个领域需要进一步研究。

首先，团队旨在改善评估方法和指标。团队尝试对主流模型进行主观类指标测试，通过GPT裁判模型及人工综合方式，基于事实正确性、用户需求满足度、清晰度、完备性、丰富度等准则规范，与官方参考答案对比为各模型打分。若以10分为满分，各模型的平均得分在4到6分之间。团队发现同类模型的参数越大往往得分越高。在主观能力评估方面，团队也将探索更先进的自动化评估技术，并制定更精细的评估标准。

其次，团队计划通过纳入更多多样化和复杂的政府事务任务来扩展CGAUE基准。这包括与应急管理、国际合作及新兴技术在政府服务中的应用相关的任务，并计划研究在评估中使用多模态数据，例如将与政府事务相关的图像和视频纳入数据集。

最后，团队将推动新评估数据集格式的采用和标准化。团队将与研究界和行业合作，鼓励在更多大型模型评估项目中使用该测试集格式，并为建立一个更全面和高效的大型模型评估生态系统做出贡献。AP

作者简介：

陈孟卓（2001-），男，硕士，现就读于中国科学院大学软件研究所，研究方向为大模型安全、大模型智能体。

郑子木（1991-），男，博士，现就职于华为云，研究方向为云原生边缘智能、人工智能标准化、人工智能基准测试、终身学习及多任务学习。

参考文献：

[1] Tianyang Lin, Yuxin Wang, Xiangyang Liu, et al. A survey of transformers[J]. AI open, 2022, 3: 111–132.
[2] Fabio Catania, Micol Spitale, Franca Garzotto. Conversational agents in therapeutic interven- tions for neurodevelopmental disorders: a survey[J]. ACM Computing Surveys, 2023, 55(10): 1–34.
[3] Shervin Minaee, Tomáš Mikolov, Narjes Nikzad, et al. Large language models: A survey[J]. ArXiv, 2024.