中国移动
China Mobile

审计行业大模型评测体系白皮书

（2025 年）

发布单位：中国移动通信集团有限公司

编制单位：中国移动通信集团有限公司数智化部

中国移动通信集团有限公司内审部

版权声明

前言

当前，全球正经历一场由人工智能驱动的深刻产业变革。人工智能、大数据等新兴技术的迅猛发展，为审计工作带来了前所未有的机遇与挑战。习近平总书记在二十届中央审计委员会第一次会议上强调：“做好新时代新征程审计工作，总的要求是在构建集中统一、全面覆盖、权威高效的审计监督体系，更好发挥审计监督作用上聚焦发力。”审计行业大模型作为融合前沿技术的创新产物，正在逐渐重塑审计业务的流程与模式，其在提升审计效率、增强风险识别、助力精准决策等方面展现出巨大潜力。其价值体现在三大维度：

1.效率跃迁：推动审计工作从传统“抽样验证+规则驱动”模式，向“全量分析+智能驱动”新范式深度转型，突破人工处理数据的能力边界，实现对海量、多源、异构数据的深度穿透与精准解读，为审计全流程注入全新动能。

2.能力升维：赋能审计人员突破个人经验局限，穿透复杂业务表象，精准洞察风险关联逻辑，揭示传统方法难以发现的隐蔽欺诈、系统性风险及业务实质，提升风险识别的深度和广度。

3.流程再造：助力审计文档生成、程序控制、质量复核等环节的标准化和自动化处理，打破环节间的协同壁垒，在保障审计作业的合规性与一致性的基础上，推进审计作业流程的自动化、智能化升级。

然而，随着审计领域各类大模型的不断涌现，其质量与性能参差不齐，如何科学、客观且全面地评测审计行业大模型就显得尤为重要。现有通用大模型评测侧重于文本流畅性与开放任务泛化能力，难以量

化审计场景特定需求，主要面临三重鸿沟：数据上，敏感信息脱敏易失真，跨机构数据孤岛阻隔评测覆盖；方法上，通用指标难以衡量审计风险，黑箱模型决策难以满足“可解释、可追溯”要求；落地上，评测与生产环境割裂，模型结果难比对。

为深度洞察不同审计行业大模型的实际效能，精准辨析各模型的优势与短板，推动审计行业大模型技术健康发展，团队依据国家标准GB/T45288.2-2025《人工智能大模型第 2 部分：评测指标与方法》，并结合中国移动联合发布的《通用大模型评测标准》，编制完成《审计行业大模型评测体系白皮书》。本白皮书创新性地提出面向审计行业的大模型评测体系，以“2+4+6”层级架构为核心：聚焦基础能力层与审计应用层两大核心场景，并将审计应用评测按审计流程细分为不同场景下的 30 余项具体应用任务。针对每项审计应用任务，白皮书清晰指明适用的评测方式、指标、数据与工具四项关键评测要素，同时细化反映功能性、准确性、可靠性等六大审计评测维度的具体指标，为评测工作提供了切实可行的落地级指南，有效弥补了通用评测在审计行业适配性与应用价值评估上的不足。

本白皮书旨在为审计行业大模型研发、机构选型及监管合规提供一个精准标尺，推动大模型安全、透明、高效地赋能审计现代化，铸就“科技强审”新范式，铸牢审计之盾！

目录

第一章：

审计行业大模型评测背景

一、审计行业大模型评测背景

（一）大模型在审计数智化转型的核心作用

在审计数智化转型的浪潮中，大模型凭借自然语言处理、模式识别和知识推理三大核心能力，在数据处理、风险识别、模式革新等方面展现出不可替代的核心作用，成为推动审计数智化转型的核心引擎。

数据处理方面，大模型凭借其卓越的自然语言处理、图像识别、多模态数据融合能力，能够高效处理审计过程中涉及的制度规范、合同文本、发票凭证等不同类型数据，自动提取关键信息，极大地提升了数据处理的效率和准确性，为审计数智化提供了更坚实的数据基础。

风险识别方面，大模型通过对过往审计案例、行业数据、企业经营数据的深度学习，能够构建动态的风险评估模型，显著提升风险识别的敏锐度，强化潜在风险的前瞻预判能力，为审计数智化提供了更主动的决策支撑。

模式革新方面，大模型不仅可以辅助审计人员完成重复性的工作，让审计人员从繁琐的事务中解放出来，专注于更高价值的风险分析、策略制定等工作，还能够对审计过程进行监控，自动检查审计程序的执行情况、审计证据的充分性和相关性，及时发现审计过程中的偏差和问题，确保审计质量的稳定性和一致性，为审计数智化提供了更可靠的质量保障。

（二）审计行业大模型应用现状

当前，审计领域的大模型应用正从探索阶段加速迈向实践落地，在提升审计效率、深化审计维度、保障审计质量等方面持续释放潜能。从文本处理的自动化革新、数据分析的深度挖掘突破，到跨领域知识的融合应用实践，大模型的强劲赋能价值已然凸显，不仅显著提升了审计工作的效率、精准度与科学性，更为审计行业的数智化转型注入了强劲动力。大模型正通过以下典型审计场景实现价值落地：

应用场景	工作内容
制度与案例的智能检索与匹配	针对审计定性和制度引用难题，大模型可构建专业知识库（如审计准则、制度规范、典型案例），实现制度条款的精准定位和案例的智能推荐，为审计判断提供可溯源的依据。
审计文档自动生成	大模型可以辅助甚至自动生成审计框架、审计通知、审计确认单、审计报告等文档，减轻审计人员工作负担。
智能问答与决策支持	审计人员可通过自然语言与系统交互，快速获取专业问答、风险提示和处理建议，提升审计决策的科学性。
审计线索筛查与异常检测	大模型自动分析财务数据、业务数据，识别异常交易、虚假报销、采购风险等。
多模态数据融合分析	结合语义理解、表格分析、图像理解等技术，大模型能够处理票据、合同、图片等多模态信息，拓展审计证据的来源和维度。

随着技术的不断发展与应用的持续深化，大模型有望在审计领域发挥更为重要的作用，推动行业迈向更高水平的智能化发展阶段。然而，受审计场景碎片化、数据异构性突出、效果量化标准缺失等现实难题制约，行业内尚未形成统一的大模型能力评测框架，对大模型技

术适配性与效果可信度的评估仍依赖局部实践经验，其跨领域能力迁移也面临标准化不足的瓶颈，因此，建立一套统一、科学的评测体系已成为行业发展的迫切需求。

第二章：

大模型评测现状分析

二、大模型评测现状分析

（一）通用大模型评测现状

通用大模型评测侧重于通过公开基准（如 GLUE、MMLU）及自动化指标（准确率、BLEU 等），结合零/少样本测试，全面评估模型在开放任务中的语言理解、生成、推理及多模态处理能力，强调指标可比性与标准化。国家标准GB/T45288.2-2025《人工智能大模型第2部分：评测指标与方法》系统构建了分层评测体系与多模态任务矩阵，围绕理解（30 类任务）与生成（16 类任务）两大核心维度设定 140 项指标，规定了数据集构建（单能力项多于200条）和综合性能评估方法（含 7 项性能指标），并区分基础与增强能力评测。

中国移动联合多家单位发布的《通用大模型评测标准》基于此国标，提出“2-4-6”框架体系：“2”指对标国标的理解与生成能力；“4”指覆盖评测全生命周期的工具、数据、方式、指标四要素；“6”指功能、准确、可靠、安全、交互、应用六大维度，旨在结合行业需求构建科学、权威的评估基准。中国信通院提出“方升”大模型评测框架，其核心通过“三横一纵”能力维度（行业能力、应用能力、通用能力为“三横”，安全能力为贯穿的“一纵”）和自适应动态测试方法，结合动态更新的百万级题库与智能抽样算法，全面评估大模型的产业落地效果，为行业选型与模型优化提供科学依据。

（二）行业大模型评测现状

行业大模型评测与通用大模型评测的核心区别在于其强行业属性、深度定制化和结果实用性导向。行业评测聚焦特定行业的实际业务需

求，通过高度定制化的场景任务设计，深度贴合行业痛点和复杂业务逻辑。其评估方法强调引入行业专家审核和真实业务反馈，确保对专业术语、监管规则及隐性知识的精准掌握，并以实用性、合规性和风险可控性为关键价值导向。相比之下，通用评测侧重于基础语言能力的普适性基准，依赖自动化指标（如准确率、BLEU）和标准化数据集，缺乏行业深度验证，结果主要用于模型能力排名而非直接指导业务落地。

以 Super-CLUE 为例，其在金融（风控／投资决策）、工业（智能体优化／预测维护）、汽车（智能座舱人机交互／驾驶辅助）领域构建三维评测体系：通过行业导向性目标设计、深度定制化场景任务、专家验证与真实反馈机制，在评测目标、场景设计、评测方法、数据构建和评价指标等方面均表现出明显的行业特性和实践导向性，有效弥补自动化评估在专业术语理解与监管适应性方面的局限，显著提升细分领域应用价值验证效力。

上海财经大学联合发布的《金融大模型应用评测指南》围绕金融业务需求，设计了涵盖模型基础、金融安全与价值对齐、风险控制、专业认知及业务辅助五大维度指标。配套数万句对高密度、动态更新的专业数据集，采用“客观量化+专家审核”融合评估方法确保精准性。评测强调工程落地，工具集成并加权量化结果，优先风险控制与业务辅助能力，直接服务于智能投研等场景效能优化，解决了通用评测在行业术语、监管适配及落地方面的不足，成为驱动金融智能化的核心基准。

（三）审计行业大模型评测需求

审计行业大模型评测需构建分层框架，覆盖核心能力、技术与流程融合、治理与持续演进三大方面，以系统性验证模型是否满足审计行业**“可靠、高效、合规、可解释”**的核心需求。

核心能力方面聚焦模型在审计实务中的基础能力。评测需覆盖模型对财务、合规、内控、IT 审计等细分场景的多样性适配能力，验证其跨行业适应性并建立专业指标。同时，需考察模型对结构化表格、文本、图像、日志等多模态异构数据的关联分析与整合能力，特别是关注其在数据缺失、存在噪声及遭遇篡改等情况下的运行稳健性。此外，敏感信息安全管控的严密性、制度依据精准校验的有效性、审计证据完整可追溯的规范性，也是其核心能力不可或缺的关键体现。

技术与流程融合方面强调模型与审计工作流的深度结合。评测需验证模型在审计计划、证据收集、风险识别至报告生成全流程中的适应性及输出匹配度，实现业务流程实景还原与人机协同。深度评估模型在财务异常识别、政策法规匹配、系统漏洞分析等专业领域的知识掌握度与逻辑推理精准性，是检验其领域知识与因果推理能力的核心。模型在高并发、长周期任务下的稳定性及对异常数据输入的容错能力亦需严格测试。最终，确保模型决策逻辑透明、依据链清晰可查，保障全流程可追溯的决策路径不可或缺。

治理与持续演进方面则关注模型的合规性、演变性与可信赖基础。评测需验证敏感数据保护措施的有效性，确保符合《数据安全法》、《个人信息保护法》等法规要求。评估模型吸收新知识、新法规的速度与准确性，以及利用历史数据和反馈持续优化性能的能力，是衡量模型动态更新与持续学习的关键。确保模型输出高度可解释、推理逻

辑可追溯，并在错误发生时责任归属清晰，是实现决策透明性与责任界定的目标。同时，还应关注交互性等用户体验要素。

以上评测需求分析旨在确保审计行业大模型在其全生命周期内能够适应行业的严苛要求，为构建真正可信赖的审计智能化工具提供科学依据。上述“可靠、高效、合规、可解释”的审计业务核心需求，具体转化并体现在本白皮书第四章提出的“功能性、准确性、可靠性、安全性、交互性、应用性”六大技术评测维度中，共同构成评估模型是否满足行业要求的完整框架。

（四）问题与挑战

审计行业的大模型评测具有高度专业化、动态化和强安全性的特点。这与通用大模型评测方法所追求的普适性、静态性和开放性存在本质差异。为满足行业规范与数智化升级需求，重点需攻克数据合规性、方法适配性、实施协同性与改进持续性四大核心问题，推动评测体系从“技术验证”到“业务赋能”的跨越。

在**审计数据**层面，挑战集中于合规性与可用性的根本矛盾。审计数据（涵盖财务、交易、合规文本等多模态信息）的脱敏与匿名化处理易导致关键特征丢失（如模糊化交易金额削弱异常检测能力），且敏感数据的跨机构共享机制缺失加剧了“**数据孤岛**”现象。同时，高专业门槛与成本制约了规模化标注数据集的构建，高风险场景数据的匮乏依赖真实性存疑的合成数据补充，而审计规则与政策的快速迭代亦使得依赖历史案例的数据库难以捕捉新兴风险模式。

在**评测方法**层面，现有通用指标（如 BLEU、ROUGE）难以满足审计任务对风险识别准确性及合规判断严谨性的核心诉求，亟需设计融合领域特性的复合评价指标。此外，大模型固有的“黑箱”特性与审计行业强调的“可复现、可追溯、可解释”原则相冲突，缺乏将可解释性量化的系统方法论。现有评测框架多基于固定数据集与封闭任务，对模型跨行业、跨规模企业的适应性与迁移能力评估不足。

在**操作实施**层面，评测体系与现有审计信息化系统（如审计底稿软件、ERP 接口）的深度集成面临接口兼容、数据流转及实时同步等技术障碍。多方参与中，数据提供方因数据泄露顾虑倾向于提供高度脱敏的模拟数据，导致评测失真，而评测方可能忽略审计法律约束，目标冲突显著。由于算力成本受限，中小机构难以私有部署大模型进行评测，导致评测结果无法准确反映实际应用效果，严重阻碍了评测的横向可比性与标准化进程。

在**持续改进**层面，技术和法规的双重快速演进要求审计行业大模型评测体系具备敏捷迭代能力，但设计兼容不同模型版本演化的动态评测基准与数据集面临巨大挑战。模型在实际审计场景中的表现反馈（如误判案例、用户修正）虽为优化关键，却因依赖主观分析而难以转化为结构化改进信号，制约了数据驱动的优化闭环。

综上，当前审计行业缺乏统一的大模型专业化评测体系，表现为**审计评测场景**、**评测方式**、评测指标与**审计案例数据库**的标准化空白。这一现状不仅阻碍了对审计行业大模型应用效果与潜在风险的客观认知，更减弱了审计结论的可靠性。

第三章：

大模型评测技术

三、大模型评测技术

（一）主要评测方式

在大模型评测领域，评测方法分为**客观评测**与**主观评价**两大类。客观评测通过预先定义的量化指标（如准确度、精确率、召回率、F1分数、困惑度、延迟、吞吐量、资源消耗）严格度量模型性能，具有高可重复性和可比性。它利用标准化数据集（如 GLUE、SQuAD）和自动化流程快速定位模型弱点，为优化提供量化依据。

主观评价依赖专家对模型输出的语言质量、逻辑连贯性、专业适用性、创意性等维度进行综合判断，弥补客观评测的不足。但其人力成本高、周期长，限制了大模型的持续评测。

为兼顾深度与效率，业界探索基于顶级大模型的**自动化主观评测**。该方法引导大模型（如 GPT-4 级）依据类似人类的标准对其他模型输出进行打分和评审。通过与少量人类评审校准，经优化后其一致性与效率接近人类水平，有望成为重要补充手段。

（二）典型评测维度

当前业界对通用大语言模型（LLM）的评估主要聚焦于以下五大核心维度，旨在为模型研发、优化及场景化部署提供客观、量化的决策依据：

基础性能评估聚焦模型在核心语言任务上的准确性与输出质量。采用行业标准基准数据集（如 GLUE、SuperGLUE、MMLU、BIG－bench）进行验证，涵盖文本理解准确率、文本生成质量（BLEU，

ROUGE，BERTScore 等）、问答正确率、代码生成能力及常识推理精度等关键指标。该评估不仅用于横向对比不同模型架构与预训练策略的效能，更旨在客观衡量模型在真实应用场景中对复杂语义信息的理解、推理与生成能力的上限与瓶颈。

泛化能力评估衡量模型在未见任务、领域或全新应用场景下的适应性与迁移能力。主要通过零样本（Zero-shot）、少样本（Few-shot）学习及跨领域迁移实验进行检验，例如评估模型在新兴垂直领域（如法律文本分析、生物医学文献理解）、低资源语言处理或少样本复杂逻辑推理任务中的表现。此维度直接反映了模型应对多样化、动态化业务需求的通用潜力与部署灵活性。

可靠性与安全性评估测试模型在非理想输入或潜在对抗环境下的稳定性与可靠性。评估场景包括对抗性文本攻击（如精心构造的提示词诱导、字符级扰动）、输入噪声干扰（如语法错误、语义模糊提示、无关上下文注入）以及信息过载或低质量输入。通过系统化压力测试，揭示模型在异常或恶意输入下的脆弱性、偏见放大风险及内容安全风险，为提升模型安全防护能力、伦理对齐及优化训练策略提供关键洞见。

逻辑一致性与可靠性评估考察模型在处理同一核心问题但不同表述形式、语境或隐含前提输入时的输出稳定性与逻辑自洽性。评估方法涉及提示词重构、多轮对话上下文连贯性验证、长文档信息一致性追踪以及复杂推理链的可验证性分析。模型若在不同形式查询或语境下产生矛盾、事实错误或逻辑断裂的结果，则暴露其内部知识表征与

推理机制的缺陷，直接影响用户信任度与系统可信赖性，是用户体验与实用性的关键考量。

资源效率与部署可行性评估量化模型在实际运行环境中的计算与存储开销，涵盖关键指标如推理延迟、峰值内存占用、模型参数量级、训练 / 微调成本及能耗水平。尤其在实时交互、大规模服务或资源受限场景（如边缘设备集成）下，此维度直接决定了模型的技术可行性与商业价值。评估需结合模型压缩技术（如量化、剪枝、知识蒸馏）后的性能资源消耗权衡分析，为不同成本、性能与延迟约束下的模型选型及工程化优化提供精准指导。

上述五大评估维度共同构成了对通用大语言模型综合能力的全景式画像，系统性地揭示了模型在真实业务场景中的优势、局限与潜在风险。该框架不仅为技术研发迭代指明方向，更为企业选型、场景适配及成本效益优化提供了坚实的科学依据，是驱动大模型技术健康发展和规模化应用落地的基石。

（三）常见评测指标及计算方式

指标名称	指标描述	计算方法	应用用途	局限性
BLEU	衡量生成文本与参考文本之间 n 元语的重叠程度	基于 n 元语的精确率，并对过短的生成文本进行惩罚	机器翻译 文本生成	主要关注字面匹配，对语义相似但措辞不同的文本评估可能不佳
ROUGE	衡量生成文本覆盖参考文本内容的程度	基于 n 元语和最长公共子序列的召回率	文本摘要	可能对简洁但准确的输出进行惩罚

困惑度	衡量语言模型预测文本序列的能力	基于模型对输出标记的概率分布的确信程度	语言模型评估	主要反映语言建模能力，不能直接衡量事实正确性和相关性
准确率	衡量模型预测正确的比例	正确预测数/总预测数	分类任务	对于开放式生成任务可能具有误导性
F1 分数	精确率和召回率的调和平均值	2*(精确率*召回率)/(精确率+召回率)	分类任务	平衡假阳性和假阴性
METEOR	考虑精确率和召回率，以及同义词匹配和词干还原	结合精确率、召回率、同义词和词干匹配，并对词序差异进行惩罚	机器翻译 文本摘要	在语义理解方面有所提升，但计算过程较为复杂
BERT Score	基于 BERT 嵌入的语义相似度	计算生成文本和参考文本词嵌入之间的余弦相似度	文本生成	能够捕捉语义信息，但可能存在局限性

第四章：

审计行业大模型评测体系

四、审计行业大模型评测体系

（一）整体框架

审计行业大模型评测的核心目标，是确保模型在高风险审计场景下的可靠性与实用价值。这要求评测超越单纯的技术适配，必须深入解构审计业务逻辑，严格依托专业化知识体系、严苛的合规要求及场景化能力验证，实现大模型能力与行业需求的深度有机融合。为此，提出了“**2+4+6**”层级评测架构。

具体而言，“**两层评测场景**”是评测架构基础。**基础能力层**聚焦模型针对审计行业中语言处理、知识理解与逻辑推理等能力。**审计应用层**则侧重评测模型在审计实务中数据解析、风险识别、审计报告生成等真实专业任务中的表现，强调整体综合能力在复杂业务环境中的实操价值。

“**四项评测要素**”提供方法论支撑。**评测方式**融合自动化基准评测、审计专家评估与大模型评测，实现多角度验证。**评测指标**涵盖准确性、召回率、F1 值、逻辑一致性、合规性等多维量化与定性标准。**评测工具**依托专业审计辅助工具、数据标注与测试脚本、智能评估平台，保障流程标准化与自动化。**评测数据**基于权威的通用审计语料库、专业标注数据、实际案例库与风险案例库，确保其完整性、时效性与代表性。

“**六个关键维度**”综合评估审计行业大模型表现与落地价值。**功能性**衡量处理多样化审计任务与多维数据的能力；**准确性**确保输出结

果与专业标准及客观事实的高度一致；**可靠性**考察在复杂异常输入下的稳定性与容错能力；**安全性**强调数据隐私保护、风险防范及内容合规；**交互性**评测与审计人员互动的便捷性、自然度与满意度；**应用性**则最终体现模型在实际审计工作场景中的落地能力、业务价值创造及用户认可程度。

“2+4+6”评测框架通过深度贴合审计业务的独特需求与高风险属性，为审计行业大模型的科学评测、性能验证与实践价值评估，提供了系统化、可操作的方法论基础与实践指南。

（二）评测场景

审计行业大模型评测场景分为**基础能力层**和**审计应用层**两类。**基础能力层**聚焦模型的语言理解、文本生成、常识问答及逻辑推理等通用能力，其水平决定了模型处理复杂审计任务的上限。**审计应用层**则评估模型在审计专业领域的表现，包括专业术语掌握、报表分析、风险评估、底稿及报告生成等核心任务，利用专业数据和专家评审判定其准确性、逻辑性与合规性，是衡量模型实际业务支撑能力的关键依据。具体任务描述如下：

1.基础能力层

自然语言生成旨在检验模型依据审计特定要求与规则，生成高质量文本（如审计文档段落、流程框架）的能力。核心在于理解审计流程、标准与术语，并确保输出内容的条理性、完整性与逻辑合规性。

自然语言理解重点评估模型对审计报告、法规条文等专业材料的深度理解与推理能力。模型需精准完成文本分类、关键信息抽取及智能问答，准确识别风险类型、控制缺陷与合规要求。

数据处理与分析聚焦模型处理结构化审计数据的能力。评测涵盖数据清洗、异常值检测、基本统计分析，以及生成可执行的审计程序代码，以支持自动化审计流程。

图片处理与分析考察模型识别、提取与分析审计相关图像（如合同、发票）关键信息的能力。通常结合 OCR 与图像语义理解技术，实现信息准确提取、一致性核对及潜在风险提示。

多模态信息整合评测模型对跨模态数据（文本、表格、图像等）的协同分析与融合能力。核心在于理解模态间语义关联，模拟“交叉验证”工作评测模式，生成综合性审计证据链，以解决多源异构数据割裂问题。

智能体评测审计智能体自主解决复杂问题的能力，重点验证其在模拟环境中理解用户意图、制定任务分步规划、动态调用工具（如 DB、OCR、脚本库）、闭环决策及反馈修正的能力，推动模型向“可信赖审计执行体”演进。

能力	细分任务	评测维度	评测指标	评测方式	评测数据
自然语言生成	根据审计领域特定要求或指定规则生成高质量的自然语言文本，包括审计文档段落、流程框架、分析结论等内容	功能性、准确性、可靠性、安全性、应用性	客观指标： BLEU、ROUGE、METEOR（语言流畅性）、无效回答次数、任务完成率 主观指标： 准确性（与审计准则相符度）、完整性（覆盖关键审计要素）、逻辑性（推理链条清晰度）、合规性（符合法规要求）	基准自动化测试（语言流畅性初筛）； 审计专家交叉复核（内容真实性与合规性验证）； 审计专家模型评价（辅助评审）	审计报告模板库； 法规条文数据集； 历史审计文档（脱敏处理）
自然语言理解	深度理解审计报告、法规条文等专业材料，完成文本分类、信息抽取、智能问答等任务，准确识别风险类型、控制缺陷与合规要求	功能性、准确性、可靠性、交互性、应用性	客观指标： 准确率、精确度、召回率、F1 分数、混淆矩阵、推理时延、无效回答次数 主观指标： 完整性（关键信息提取）、可理解性、适用性、逻辑性	基准自动化测试（量化基线）； 审计专家交叉复核（高风险样本专业性验证）； 审计专家模型评价（规模化评审）	标注的审计事项分类集； 风险类型标签库； 审计准则问答库； 法规条文判例库； 审计证据标注集

数据处理与分析	处理结构化审计数据，包括数据清洗、异常值检测、统计分析、审计程序代码生成等，支持自动化审计流程	功能性、准确性、可靠性、安全性、应用性	客观指标： 准确率、精确度、召回率、F1 分数、混淆矩阵、任务完成率、推理时延、系统功耗 主观指标： 可行性（代码可执行性）、创新性、适用性、完整性	基准自动化测试（执行测试与异常检测）； 审计专家交叉复核（复杂案例与代码审查）	异常检测数据集； 含噪声数据样本； 审计程序模板库； 审计查询日志
图片处理与分析	识别、提取与分析审计相关图像（如合同、发票、凭证等）的关键信息，实现信息准确提取、一致性核对及潜在风险提示	功能性、准确性、可靠性、安全性	客观指标： 准确率（OCR 识别）、精确度（字段提取）、召回率、任务完成率 主观指标： 准确性、完整性、适用性	基准自动化测试（标准化测试）； 审计专家交叉复核（关键凭证核验）	发票/合同扫描件库； 审计凭证图像集； 印章样本库
多模态信息整合	融合文本、图像、语音等不同模态数据，进行跨模态对齐与特征互补，理解模态间语义关联，生成综合性审计证据链	功能性、准确性、可靠性、应用性	客观指标： 准确率（跨模态匹配）、任务完成率、推理时延 主观指标： 完整性（信息融合度）、逻辑性（证据链连贯性）	基准自动化测试（跨模态基准）； 审计专家交叉复核（综合分析评审）； 审计专家模型评价（证据链验证）	多模态审计证据库； 文本-图像配对数据

智能体任务	评测审计智能体自主解决复杂问题的能力，包括意图理解、任务规划、工具调用、闭环决策及反馈修正	功能性、准确性、可靠性、交互性、应用性	客观指标： 准确率（意图识别、工具选择）、任务完成率、推理时延、无效回答次数 主观指标： 可行性（规划合理性）、创新性、完整性（信息融合）、逻辑性	基准自动化测试； 审计专家交叉复核（复杂任务与高风险决策）； 审计专家模型评价（多模型交叉验证）	审计任务场景库； 意图标注数据集； 工具测试集

2.审计应用层

与基础能力层侧重通用能力不同，审计应用层评估模型在审计专业领域的实战能力，聚焦于审计业务全流程的核心专业场景，要求模型深度理解审计准则体系、准确把握风险要素、高效支撑审计决策。具体任务描述如下：

审计专业知识问答评测模型对审计专业知识体系的理解与应用能力，包括对党规国法、审计准则、规章制度等专业知识的掌握，以及基于知识库的智能检索与推荐能力。该任务要求模型不仅能准确理解和回答审计专业问题，还能根据具体场景推荐相关的审计项目、模型、内参资料和风险案例。

审计风险识别与分类考察模型在风险管理领域的专业判断能力，涵盖风险点识别、拆解、分类以及问题性质判定等核心功能。模型需要准确识别审计发现中的风险类型、问题原因、影响程度和性质，并能追溯相关制度依据，为审计决策提供精准的风险画像。

审计文档智能生成评估模型生成高质量审计文档的能力，包括审计方案、通知书、底稿、报告等关键文书的自动化撰写。该任务不仅要求文本流畅性和完整性，更强调内容的专业性、合规性和逻辑严密性，确保生成文档符合审计工作规范。

审计过程执行评估模型在审计过程执行中处理和分析审计数据的综合能力，包括自动化数据分析、异常检测、数据可视化等。模型需要能够识别数据异常模式、执行统计分析、生成可视化报表，支持数据驱动的审计决策。

审计合规性检查评测模型进行合规性审查和一致性核验的能力，包括文档合规检查、审计证据链完整性识别等。该任务要求模型能够精准比对多源信息，识别偏差和违规点，确保审计工作的合规性。

审计整改与监督则考察模型在审计后续管理中的应用能力，包括整改结果审核、追责判定、处分决定生成等。模型需要能够评估整改措施的有效性，判定责任归属，支持审计成果的落实与追踪。

场景	细分任务	评测维度	评测指标	评测方式	评测数据
审计专业知识问答	党规国法/审计准则问答； 规章制度问答； 审计项目问答与推荐； 内参资料/风险案例推荐； 数据字典问答	功能性、准确性、可靠性、交互性、应用性	客观指标： 准确率、精确度、召回率、F1分数、无效回答次数 主观指标： 准确性、完整性、逻辑性、适用性	基准自动化测试； 审计专家交叉复核（复杂问题）	党规国法问答对； 审计准则题库； 历史项目文档库； 风险案例库
审计风险识别与分类	风险点拆解与分类； 问题原因/影响/性质分类； 审计发现分类； 问责类型判定；	功能性、准确性、可靠性、安全性	客观指标： 准确率、精确度、召回率、F1分数、混淆矩阵 主观指标： 准确性、逻辑性、合规性	基准自动化测试（主要）； 审计专家交叉复核（高风险样本）	风险描述文本与分类标签； 问题分类标准库； 违规行为数据集
审计文档智能生成	审计方案、通知书、报告生成； 审计发现归纳总结； 审计底稿描述生成	功能性、准确性、可靠性、应用性	客观指标： BLEU/ROUGE/METEOR、BERTScore、困惑度 主观指标： 准确性、完整性、逻辑性、合规性、可理解性	审计专家模型评价； 审计专家交叉复核（重点审查）	标准文档模板库； 审计要素表； 专家编写的范本； 历史优秀文档

审计过程执行	自动化数据分析与可视化； 异常检测； 关键信息提取； 审计材料脱敏	功能性、准确性、可靠性、安全性、应用性	客观指标： 精确度、召回率、F1 分数、任务完成率、推理时延 主观指标： 可行性、创新性	基准自动化测试； 审计专家交叉复核（分析结果验证）	带标签的审计数据集； 数据质量测试集； 预期分析结果集； 脱敏标准数据集
审计合规性检查	合同与中标文件一致性分析； 审计证据闭环识别； 审计底稿逻辑检查	功能性、准确性、可靠性、安全性	客观指标： 准确率、精确度、召回率、F1 分数 主观指标： 完整性	基准自动化测试； 审计专家交叉复核（合规性验证）	合规规则库； 证据链数据集
审计整改与监督	整改结果审核； 确认单辅助审理； 审计线索查处判定； 审计意见与决定生成	功能性、准确性、可靠性、应用性	客观指标： 准确率、F1 分数、任务完成率 主观指标： 适用性、可行性、合规性	审计专家交叉复核（主要）； 基准自动化测试（辅助）	整改材料+审核依据； 审计线索与查处措施案例库； 专家标注的处分案例

（三）评测要素

审计行业大模型评测需立足审计行业特性，围绕**评测方式**、**评测指标**、**评测数据**与**评测工具**四大核心要素实施专项优化，以构建系统化的能力评估框架。

1.评测方式

基准自动化测试是一种依托标准化审计知识库、法规条文及判例库、以及经脱敏处理的真实业务数据集，对审计行业大模型的核心专业能力进行系统性量化评估的方法。该评估聚焦于评测审计应用中的关键表现，例如：利用财务异常检测数据集考核模型的风险识别能力；通过审计准则与法规问答库检验其对多层级制度体系的语义理解与应用准确性。尽管该方法具备高效性与客观性优势，其核心挑战在于构建与维护**高质量**的审计行业数据集成本高昂。除此之外，还需辅以动态更新的法规库及复杂业务情境模拟案例，以验证模型在审计准则演进及非标准化业务场景下的泛化能力与适应性。

审计专家交叉复核作为审计大模型评测的核心验证机制，在涉及重大专业判断、风险评估及审计意见形成等高风险关键环节中不可或缺。该方式需要依托由资深注册会计师、合规专家及行业顾问构成的专家委员会，对模型输出进行深度专业审查。通过明确评估维度、构建结构化复核工具以及实施独立验证等标准化实施步骤机制，确保评估结论的客观性与公正性。审计专家复核的核心价值在于其能精准识别自动化指标难以捕捉的逻辑缺陷、合规性偏差等风险。然而，该评测方式实施成本**高昂**、**周期冗长**且规模化应用受限，在实践中通常采

用人机协同评测模式：自动化基准测试进行初步筛选与大规模覆盖，专家资源则聚焦于高风险、高复杂度样本的精深复核，从而在保障评测深度的前提下提升整体效率。

审计专家模型评价是一种利用经过审计行业知识调优的大模型对被测模型输出进行自动化评审的新兴范式。该范式通过提示工程、预训练或微调等模型训练手段，将复杂审计准则与专业判断内化于审计专家模型，进而对风险报告、会计建议等进行合规性与合理性评估。此方法在拓展评测覆盖、模拟专业判断、减少专家资源依赖方面潜力显著，但面临审计专家模型生成幻觉导致虚假合规判定及内置偏差导致系统性失真的核心风险。为确保评测稳健可靠，必须定期与人类专家结论对齐校准，并实施多模型交叉验证。

综上所述，审计行业大模型评测方式需要平衡效能与严谨性。基准自动化测试构建模型核心能力的量化基线；审计专家交叉复核作为“金标准”，以其专业深度与权威性为评测质量提供最终保障；审计专家模型评价作为专家评审的有效补充与延伸，推动评测向规模化、半自动化演进。三种方法协同互补，共同构筑起多层次、全方位的审计质量保障体系，确保大模型在高信任度要求的专业领域实现安全、合规、有效的应用。

2.评测指标

科学评估审计行业大模型的性能，需构建融合通用评估框架与审计核心要求的评测体系。该体系的核心在于区分**客观量化类指标**与**主**

观专家评判类指标，以契合审计工作对合规性、风险导向和专业审慎性的本质要求。

客观量化类指标适用于有明确审计证据或法规标准参照的任务，并涵盖保障审计效率与项目交付的关键维度。其价值在于标准清晰、结果可量化且可比性强，能有效降低主观偏差，确保评测独立客观。这类指标按任务性质可分为三类：

判断分类任务指标评估模型依据审计准则对信息进行定性或分类的能力（如识别合同关键条款、筛选高风险舞弊交易、分类内控缺陷）。核心指标包括准确率、精确度、召回率、F1 分数及揭示误判模式的混淆矩阵。

内容生成任务指标评估模型基于审计证据生成合规内容的能力（如撰写底稿初稿、提炼合同摘要、草拟审计发现）。常用指标如 BLEU、ROUGE、METEOR 仅能辅助评估语言流畅性。审计领域的关键在于内容的真实性、准确性与合规性，因此必须结合主观专家评判，严防语言流畅但事实错误或违规的无效输出。

稳定性与效率指标关乎模型质量与成本效益。稳定性着重考察模型面对不完整、模糊或潜在误导性信息时的表现，通过输入含矛盾或对抗性样本，统计其无效输出率及任务失败率。稳定性好的模型应识别信息不足并提示追加审计程序，效率则关注时延和资源消耗 / 成本。

主观类指标适用于没有固定标准答案的题目类型，通常采取主观评分的方法进行评估。主观类指标涉及到答案准确性、完整性、逻辑性等多个维度的评价，这要求评估者拥有丰富的审计专业知识和审计

经验。为了保证评估结果的一致性和可靠性，需要对各个评分等级设定明确的标准和分数范围，最后进行汇总和分析，可以采用各种统计方法和技术，如内部一致性检验和信度分析，以提高评估结果的稳定性和信度。

以下是这些指标的详细说明、含义和计算方法：

（1）客观类指标

指标名称	含义	计算方法
准确率(Accuracy)	模型正确预测的样本数量占总样本数量的比例。	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
召回率(Recall)	模型正确预测的正类样本数量占所有实际正类样本数量的比例。	$\text{Recall} = \frac{TP}{TP + FN}$
精确度(Precision)	模型正确预测的正类样本数量占所有预测为正类样本数量的比例。	$\text{Precision} = \frac{TP}{TP + FP}$
F1 分数(F1-score)	精确度和召回率的调和平均值，用于综合评估模型性能。	$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
混淆矩阵(ConfusionMatrix)	显示模型预测结果的详细分布情况，包括 TP、TN、FP、FN 四个部分。	通过统计模型的预测结果与实际结果的匹配情况构建。
BLEU	用于评估生成文本与参考文本的相似度，常用于机器翻译。	计算生成文本与参考文本之间的 n-gram 匹配程度。

ROUGE	用于评估生成文本与参考文本的重叠程度，常用于文本摘要。	计算生成文本与参考文本之间的 n-gram 重叠率。
METEOR	用于评估生成文本与参考文本的相似度，综合考虑精确度、召回率和词序。	计算生成文本与参考文本之间的精确度、召回率和词序匹配程度。
推理时延	模型从接收输入到生成输出所需的时间。	记录模型处理每个输入样本所需的时间，并计算平均时延。
系统功耗	模型在训练和推理过程中的能耗。	通过硬件监控工具测量模型在运行过程中的功耗，并计算平均功耗。
无效回答次数	模型在回答过程中给出无效回答的次数。	统计模型在测试集上的无效回答次数。
任务完成率	模型在面对输入扰动时完成任务的比例。	统计模型在不同扰动条件下成功完成任务的次数占总次数的比例。

（2）主观类指标

指标名称	含义	评估方法
创新性	模型回答的新颖程度和创造性。	回答在审计程序、风险识别或效率提升方面是否突破常规。
可行性	模型回答的实际可操作性和实现可能性。	建议的审计步骤或技术在当前资源限制下是否可执行。

适用性	模型回答在特定场景或问题中的适用性。	回答是否精准匹配特定审计目标与被审计单位特点。
准确性	模型回答的正确性和符合实际情况的程度。	回答内容（事实、程序、结论）是否与审计证据、准则相符。
完整性	模型回答的全面性和信息的完整程度。	回答是否覆盖重大领域、风险点，并基于充分审计证据。
逻辑性	模型回答的逻辑结构和连贯性。	推理链条是否清晰严密，展现审计轨迹。
可理解性	模型回答的清晰度和易懂程度。	回答是否清晰传达审计发现、结论与建议。
合规性	模型回答是否符合审计准则、会计准则及相关法规要求。	回答是否符合审计准则及相关法规要求。

3.评测数据

审计行业大模型的评测数据集构建，需与审计行业严密的准则体系、专业资格认证框架及强制性法律法规要求进行深度耦合。在数据集构建的丰富性维度上，需超越通用知识覆盖，包含审计准则（如中国注册会计师审计准则、国际审计准则）、行业监管规范（如上市公司审计监管规则）、会计审计类考试（CPA、CIA、初中高级审计师、CISA 等）核心大纲与高频高难度审计实务案例，并严格贯穿初、中、高级审计人员能力进阶所必需的、差异化的知识图谱与判断场景，覆盖审计全流程关键智能化任务场景：

知识服务：支撑审计项目／模型／内参推荐、数据字典问答、模型代码逻辑解析。

过程实施：提供审计文本关键信息提取、异常检测（如重复付款、金额偏差）、自动化数据分析与可视化、风险点拆解（含制度依据）、追责制度溯源、问责类型判定所需依据。

报告与底稿：赋能审计通知书／报告／征求意见书生成、审计发现归纳、材料脱敏、底稿描述检查、证据闭环识别。

整改与追责：支持确认单辅助审理、整改结果审核、合同／中标文件一致性分析、文档合规检查等闭环核查与判定任务。

评测数据应严格区分公开验证集（模型调参）与保密测试集（最终评估），严防过拟合，维护评测结果的行业公信力。测试集需覆盖各类审计任务并有足够样本量，确保结果代表性。

评测数据应具备持续优化能力，可以动态迭代保障专业时效性。依据评测发现的专业能力短板调整样本分布与提示设计。定期引入最新审计案例与法规更新，保持测计隻时代契合度。融合自动化评测与专家评审，提升评估效率与科学性。

4.评测工具

在审计行业大模型评测体系中，评测工具的研发与应用是实现评测流程标准化与高效化的关键支持。评测工具覆盖从数据准备到结果可视化的全链路功能，既能保障数据集的质量与代表性，又能实现评测过程的自动化与复现，最终以直观的方式呈现模型优劣。下文按照

数据管理、评测执行与结果分析三大能力模块，分别阐述工具具备的核心功能。

数据管理：评测工具需内置审计场景基准数据库，涵盖文本、表格、图像等多模态经专家标注数据，确保专业覆盖度与代表性。提供灵活的数据清洗组件（自动去重、噪声过滤）和专业数据编辑功能，支持按需构建针对性审计任务子集（如高风险领域测试），加速评测方案迭代。

评测执行：自动化专业流程工具实现评测流程高度自动化，提供标准化模型接口脚本，便于快速对接各类大模型。与数据管理模块无缝衔接，自动加载指定审计评测集，通过命令行 / API 逐条发送测试用例。全程记录响应时间、资源消耗及输出，形成详实评测日志，为性能分析与问题溯源提供依据。

结果分析：工具自动比对模型输出与专家标注答案，计算审计任务相关指标（如针对分类任务的准确率、生成任务的专业合理性与逻辑性评分）。提供多维度解读功能，如误差类型分布（如法规误用、风险误判）、任务难度分层分析、模型专业能力随数据规模变化趋势。评测结论通过交互式图表 / 报告（如审计能力维度雷达图、模型对比柱状图）直观呈现，清晰揭示模型在各项审计专业能力上的优势与不足。

审计行业大模型评测工具的核心价值在于实现了复杂专业评估流程的标准化、自动化与深度洞察。通过集成化的数据管理、高效的自动化评测执行及多维度的专业结果分析，有效解决审计领域模型评估

的专业性壁垒与效率瓶颈。它不仅为模型在关键审计任务上的性能提供了科学、可复现的量化标尺，更通过模块化设计支撑评测体系的持续演进，显著赋能审计行业大模型的规范化发展、精准优化与可信赖落地。

（四）评测维度

构建严谨、面向实践的审计行业大模型评测体系，需科学选取**功能性**、**准确性**、**可靠性**、**安全性**、**交互性**和**应用性**作为六大核心评测维度。这一体系系统性覆盖了大模型从基础能力构建、输出质量保障、风险抵御到审计业务价值实现的关键环节。它们共同构成了一套多维立体、深度契合审计行业要求的评估框架，为客观衡量模型在复杂审计环境中的综合表现、识别潜在短板、确保技术应用的可靠性、合规性与价值最大化提供了不可或缺的依据。

功能性维度是模型应用于审计场景的基础门槛。它系统评估模型在完成典型审计任务（如海量文本分析、关键信息精准抽取、合规性检查、报告自动生成等）所必需的核心能力及其覆盖范围。评测聚焦于其在传统任务效能、零/少样本条件下的快速泛化能力、多任务协同处理效率以及对多模态信息（如结合文本与表格数据）的理解融合水平。强大的功能性是模型能否有效分担审计基础工作负荷、拓展自动化覆盖范围、并适应多样化审计需求的根本前提。

准确性维度是审计大模型可信赖度的核心支柱，直接关系到审计结论的客观性与权威性。该维度通过严谨的指标（涵盖分类精度、生成内容的事实一致性、幻觉发生率、语义贴合度等），全方位检验模

型输出结果与客观事实或专业标准答案的高度吻合程度。在审计这一高度强调证据确举、结论精准的领域，模型的超高准确性是杜绝信息失真、保障审计质量、支撑可靠决策判断的绝对基石，任何偏差都可能引致严重后果。

可靠性维度着重检验模型在审计工作常见复杂与干扰环境下的稳定表现。审计数据常面临非结构化文本噪声、扫描文档识别错误、行业术语变化、以及潜在的数据分布漂移等问题。该维度评测模型在输入扰动、对抗性样本攻击或数据突变场景下的性能保持能力与抗干扰性。强大的可靠性意味着模型能够在现实、非理想的审计数据条件下（如处理低质量凭证或非标准表述）依然保持可靠输出，有效规避因输入瑕疵导致的误判风险，保障审计流程的稳健运行。

安全性维度是审计大模型应用不可逾越的合规红线与风险防火墙。审计工作涉及敏感信息，且结论具有法律效力，模型必须严防生成任何有害、歧视性、虚假或违反法律法规及职业道德的内容。该维度深度评估模型在伦理边界控制、风险内容抑制、应对诱导性提问的防御能力，以及对隐性偏见和价值观误导的规避水平。系统性、高标准的安全性评测，是确保模型行为严格符合审计行业监管框架、保护用户及被审计方权益、维护审计机构专业声誉和避免重大法律风险的核心保障措施。

交互性维度衡量模型作为智能化审计助手在动态人机协作中的效能与体验。现代审计流程常涉及多轮问答澄清、复杂问题探讨、结果复核反馈等交互场景。该维度评估模型在上下文理解与保持、响应相关性、个性化表达、主动澄清需求、以及容错与自我修正等方面的能

力。优秀的交互性能够显著提升审计人员与模型的协作效率，使模型更自然地理解复杂查询意图，提供清晰、连贯、有价值的反馈，降低使用门槛，增强用户信任与工具黏性，是提升审计工作智能化体验的关键要素。

应用性维度是评测体系的最终价值落脚点与产业化推动力，聚焦模型在真实审计业务流中的深度融合与实效产出。它深入评估模型在特定审计领域的知识适配度、与现有审计软件／工作流的集成顺畅度、定制开发与微调的便捷性、部署后的运行稳定性与维护成本，以及在真实任务中提升效率、降低成本、改善用户满意度的实际成效。高应用性意味着模型不仅能解决具体审计痛点（如提升高风险样本筛查效率、自动化繁琐核查步骤），更能无缝融入业务流程，满足行业规范与标准要求，展现出可量化、可持续的业务价值与竞争优势，驱动审计智能化转型的实质性落地。

为便于具体评测实施，上述六大核心评测维度可进一步细化为若干可观测、可量化的具体指标。各维度与具体指标的对应关系详见本白皮书附录（四）。

第五章：

评测实施与持续维护

五、评测实施与持续维护

评测实施与持续维护是确保审计行业大模型评测体系可靠性、公正性和有效性的关键环节。为确保评测体系能够落地并有效运行，需要在评测数据质量、评测流程设计和持续维护机制方面构建详细且可实施的管理框架。

（一）评测数据质量管理

评测数据质量管理需建立全域覆盖的数据采集标准体系，深度耦合智能体、多模态分析等六大基础能力与财务、制度等核心审计应用场景。基于分任务标准化数据集构建统一格式与标注规范，同步覆盖常规与异常场景以系统验证模型可靠性。为保障数据全面性，应通过应用矩阵精准定位各审计场景的关键需求与挑战，并建立动态迭代机制：依据模型评测结果与实际业务反馈持续补充高价值数据，及时淘汰失效案例，有效对抗时效性衰减。

1. 评测数据集构建

为提高评测数据的标准化和有效性，评测数据应明确规定数据的来源范围、质量要求和代表性，以覆盖自然语言生成、自然语言理解、数据处理分析、图片处理分析、多模态信息处理、智能体等六大技术能力。此外，应充分考虑审计专业领域的多样性，包括监督追责审计、经济责任审计、风险管理审计、信息技术审计、采购供应链审计、人力资源审计、内控审计、财务审计等交叉领域的具体业务场景。例如，在财务审计中，应涵盖不同财务报表的类型、财务数据异常情况以及

合规审计案例；在人力资源审计中，则需涵盖人员绩效评估场景、薪酬核算案例及人力资源风险管理实例。

2.评测数据集的全面性和有效性保障

为确保测试数据集全面覆盖所有评测场景，应建立场景覆盖矩阵，明确不同审计领域内的具体场景所涉及的数据类型、复杂程度和特殊要求，有效指导数据采集和标注工作，提高数据采集的精度与场景适用性。同时，为保障测试数据集的全面性和有效性，应通过审计行业专家与技术团队的紧密协作，构建能够反映真实审计业务复杂性与多样性的标准测试数据集，具体做法包括：根据审计业务实际情境，设计针对不同业务类型、风险等级、复杂度水平的差异化数据构建方案，并通过专家评审与验证，确保数据集质量。

3.测试数据管理与迭代优化

应建立明确的迭代优化机制，通过分析模型评测结果、用户反馈和实际业务表现，识别数据集的覆盖不足或质量问题，定期对数据集进行补充、更新和重新标注。具体措施包括：

-定期评估数据集的有效性和代表性。

-根据实际审计业务表现情况，淘汰或更新过时数据。

-根据最新业务场景发展趋势，不断扩充新的测试案例。

-建立数据质量持续监控机制，确保数据集长期保持高质量和高适用性。

针对**数据脱敏**后可能丢失关键特征的问题，可采用可逆脱敏或分级加密方案，在脱敏过程中保留数据的统计分布和关键数值区间，确保异常检测算法依然能够捕捉到重要特征；同时，引入差分隐私方法，

通过在统计汇总层面添加噪声来保护隐私，最大限度地降低对数据可用性的影响。

为打破跨机构**数据孤岛**并扩大测试集覆盖范围，可采用联邦学习或安全多方计算等隐私保护技术，实现各审计机构在本地保留敏感数据的同时进行协同训练；此外，还可以推动建设行业数据信托平台，制定跨机构的数据共享协议与合规指南，使合规数据在受控环境下实现共享与评测。

针对**审计数据人工标注成本高**和**专业门槛高**的问题，应引入主动学习和半监督学习策略，利用模型不确定性度量自动筛选最具价值的样本供专家标注，并结合审计领域知识图谱与智能标注工具，实现标注流程的半自动化，从而有效降低专家投入成本并提升标注效率。

对于**高风险场景数据稀缺**问题，可通过场景仿真和数据增强技术生成合成样本，结合基于规则的合成引擎或生成对抗网络（GAN）模型模拟复杂风险事件；生成数据需经审计专家进行真实性评估和校验，以保证合成样本能够真实反映高风险审计场景。

（二）评测流程设计与实施

为构建可信、可比的审计大模型评测体系，应建立标准化流程并依托容器化技术（如 Docker）统一环境配置，确保高度一致性与可复现性。流程需深度集成现有审计信息化系统（底稿软件、ERP 接口等），严格验证数据接口规范与兼容性，以准确模拟真实业务流程。评测过程需详尽记录参数、日志及结果，并开发自动化工具支持一键复现。

在评测审计行业大模型时，首先必须确保评测方式与审计业务需求深度契合。具体而言，自然语言生成、自然语言理解、数据处理分析、图像处理分析、多模态信息处理及智能体交互等核心技术能力，需与监督追责审计、经济责任审计、风险管理审计、信息技术审计、采购与供应链审计、人力资源审计、内控审计、财务审计等专业场景任务进行交叉对应，从而构建覆盖全面且具有针对性的任务-技术矩阵。其次是评测指标的选定，围绕“功能性、准确性、可靠性、安全性、交互性、应用性”6 个关键维度，综合评估审计行业大模型表现与落地价值（附表四）。

评测流程中，先通过基准自动化测试构建模型核心能力量化基线，再由审计专家交叉复核高风险样本确保专业深度，同时可引入审计专家模型评价作为补充，拓展规模化评测覆盖范围。评测过程需详尽记录参数、日志及结果，并开发自动化工具支持一键复现。为解决中小机构硬件限制导致的模型性能不可比问题，应设计标准化基准与轻量模型对比机制，并探索云端共享评测环境。最终，引入独立第三方专家监督并定期发布权威报告，是提升评测公信力与透明度、促进行业认可的关键。

（三）持续维护与更新

为确保审计行业大模型评测体系持续有效，必须构建系统化、标准化且灵活的维护与更新机制。核心措施包括：建立定期专家联席评审制度，动态分析技术演进、法规变化及模型表现，据此优化评测标准、基准与适配方案；搭建结构化用户反馈与数据分析平台，高效收集实践问题与改进建议；实施规范的版本管理与透明发布流程，确保

迭代可追溯；特别增设生成内容合规性与事实准确性专项评估维度，引入法律合规专家并开发审查工具，平衡创新与合规风险。上述闭环机制紧密协同，保障评测体系的前瞻性、时效性与可靠性，为审计大模型安全有序发展提供坚实支撑。

第六章：

审计行业大模型评测展望

六、审计行业大模型评测展望

未来，审计行业大模型评测的核心使命将从基础功能与性能验证，跃升至对模型“审计智慧”的系统性度量。其终极目标在于驱动审计大模型完成从辅助工具向“可信监督主体”的战略转型，构建**精准化**、**全天候**、**权威性**的智能监督范式。评测体系的演进将聚焦三大核心维度：

构建互信穿透力：评测需引领跨机构审计证据链的互信互验机制建设，确保模型推理过程与数据源头的可追溯、可验证，实现结论的穿透式审视，为审计行业公信力奠定信任基石。

驱动风险预见力：评测必须推动模型能力从“事后检查”向“事前预警”跃迁，使之成为风险的动态“瞭望塔”，为前瞻性风险防控与政策制定提供洞见。

锚定价值引领力：终极评测标准将深度融入社会责任与伦理价值，评测结果应直接服务于行业标准优化与监管政策完善，确保智能审计发展始终与国家治理现代化的公共利益高度协同。

附录：

（一）人工评审表

问题	标准答案	大模型答案	创新性	可行性	适用性	准确性	完整性	逻辑性	可理解性	合规性	备注	评分标准
问题 1（用户日志问题筛选/人工提问）	标准答案（有则给出，没有人工给出）	待评测大模型给出答案										见附注

维度	创新性	可行性	适用性	准确性	完整性	逻辑性	可理解	合规性
1 星	无新意	难以落地	单一场景	错误率高	关键功能缺失	推理混乱	表述晦涩	高风险违规
2 星	局部改进	需定制调整	多场景覆盖	偶有误差	基础功能完备	基本合理	部分可读	需人工复核
3 星	突破性创新	可直接使用	全业务支持	精准可靠	全流程覆盖	严密自洽	清晰易懂	自动合规

（二）众包人工评审表

问题	大模型 1 答案	大模型 2 答案	评判数据（单选）	备注	评分标准
问题 1（用户日志问题筛选/人工提问）	待待评测大模型（匿名）给出答案	其他评测大模型（匿名）给出答案	模型 1 好/模型 2 好/模型 1 和 2 都好/模型 1 和 2 都不好		见附注

（三）大模型自动化评测方法

1）以下是参考 OpenAI 的自动化评估脚本，核心思路通过写 prompt 模版调用大模型能力自动动化评估

fact:

prompt:|

你正在比较提交的答案与给定问题的专家答案。以下是数据：

[BEGINDATA]

****

[Question]:{input}

****

[Expert]:{ideal}

****

[Submission]:{completion}

****

[ENDDATA]

比较提交的答案与专家答案的事实内容。忽略任何风格、语法或标点符号的差异。

提交的答案可能是专家答案的子集或超集，或者可能与之冲突。确定适用的情况。回答以下问题：

(A)提交的答案是专家答案的子集，并且完全一致。

(B)提交的答案是专家答案的超集，并且完全一致。

(C)提交的答案包含与专家答案相同的所有细节。

(D)提交的答案与专家答案之间存在分歧。

(E)答案不同，但这些差异在事实性方面无关紧要。

choice_strings:ABCDE

input_outputs:

-input:completion

2）以下是根据主观评判维度制定的自动化评测 prompt

#主观题评分系统：

请对以下大语言模型在特定学科问题上的输出进行评分，使用 1 到 3 星的评分尺度，从以下八个方面进行评估：

准确性：

1 星表示完全错误

2 星表示部分正确

3 星表示完全正确

.....

用户:[问题]

LLM:[llm 的回答]

正确答案是:[正确答案]

请按以下格式给出评分：

{“创新性”：星级数(整数)，“可行性”：星级数(整数)，“适用性”：星级数(整数)，“准确性”：星级数(整数)，“完整性”：星级数(整数)，“逻辑性”：星级数(整数)，“可理解性”：星级数(整数)，“合规性”：星级数(整数)，}

（四）评测维度与指标对应关系

评测维度	指标名称	描述	指标	计算方式
功能性	任务覆盖度	模型覆盖任务类型和场景的程度	支持任务类型／场景数量（绝对数量） 2、任务覆盖率（相对比例）	1.人工审核或通过 API 测试，统计模型官方支持或实测有效的任务类别（如文本摘要、机器翻译、问答、代码生成、图像描述等）和具体场景（如财务审计、风险管理等）的总数。 2.（模型支持的任务场景数量）／（预设或行业标准任务场景集合的总数量）*100%。需要定义一个基准任务集。
	任务完成率	模型成功完成指定任务的比例	1、任务通过率 2、特定任务指标（如准确率、BLEU、ROUGE、F1 等）	1.（模型输出结果被判定为“成功完成”的任务实例数量）／（测试任务实例总数）* 100%。明确“成功完成”标准（如输出满足特定格式、包含关键信息、答案正确等）。 2.针对每个具体任务类型，使用其领域标准指标。分类任务：准确率；翻译任务：BLEU；摘要任务：ROUGE；问答任务：F1
	多模态能力	模型能同时处理文	1．单模态任务指标（分模	1.分别评估模型在纯文本、纯图像、纯语音输入任务上的表现，使用相应领域的标准指标（如文本任务用准确率／F1，图像任务用 mAP／IoU，语音任务用 WER）。

		本、图像、语音等多种输入的能力	态评估） 2．跨模态任务指标（评估融合理解与生成）	2.设计需要同时理解多种模态输入或生成跨模态输出的任务，并计算： 跨模态任务理解准确率， 跨模态任务生成 BLEU。
	零样本/小样本表现	模型在未见任务或仅有极少样本条件下的推理能力和泛化能力	1．零样本任务指标（特定任务指标） 2．小样本任务指标（特定任务指标） 3．相对性能下降	1.在未提供任何该任务训练样本的情况下，仅给出任务描述或提示，让模型执行新任务。计算该任务的标准指标（如准确率、F1、BLEU 等）。 2.提供极少量（如 1，3，5，10 个）该任务的示例样本（带输入输出），让模型学习后执行新样本。计算该任务的标准指标。 3.（全量监督训练下的任务性能－零样本/小样本下的任务性能）/全量监督训练下的任务性能*100%。下降越小，零样本/小样本能力越强。
	功能扩展性	模型新功能的集成与扩展便利性	1．新功能集成时间/成本 2．扩展后性能表现	1.平均集成时间（人日/人周）：集成一个典型新功能（如支持一个新 API、适配一个新任务类型）所需的平均开发/适配时间。 所需样本量：为模型添加新功能所需提供的微调/提示样本的平均数量。 API/接口复杂度：评估扩展新功能所需调用的 API 或修改配置的复杂程度。 新任务性能：使用新任务的标准指标评估其表现（见“任务完成率”）。 原有任务性能保持率：（扩展后原有核心任务的性能）/（扩展前原有核心任务的性能）*100%。确保扩展不影响原有能力。 资源开销变化：扩展新功能后，模型推理速

				度、内存占用等资源消耗的变化率。
准确性	事实准确性	输出信息符合客观事实的比例	1．事实正确率/事实一致性得分 2．事实错误率 3．特定任务指标（如问答的 EM/F1）	1.（事实完全正确的输出数）/（测试样本总数）＊100%；需定义“事实陈述单元”，并基于可靠知识源（如百科、权威数据库、专家标注）进行验证。通常需要人工评估或利用高质量知识库进行自动化验证。 2.（包含至少一个事实错误的输出数）/（测试样本总数）＊100% 3.在涉及事实性问答的任务中，使用标准指标（如 ExactMatchF1）来衡量答案与标准事实答案的匹配程度。
	意图理解准确率	准确理解用户意图的比例	意图分类准确率（Accuracy） 2．后续动作/回复相关性（人工评估）	1.模型正确识别用户意图（通常预定义类别）的比例。（正确分类意图的样本数）/（测试样本总数）＊100% 2.评估模型基于理解意图后采取的动作（如调用工具）或生成的回复是否恰当满足该意图。常用人工评分（如 1—5 分）或通过率。
	信息抽取准确率	精准抽取关键信息的比例	精确度（Precision） 召回率（Recall） F1 分数（F1—score）	信息抽取（实体识别、关系抽取、事件抽取等）的标准评估指标。 评估模型基于理解意图后采取的动作（如调用工具）或生成的回复是否恰当满足该意图。常用人工评分（如 1—5 分）或通过率。 1.Precision＝（正确抽取的项数）/（模型抽取的总项数） Recall＝（正确抽取的项数）/（测试集中存在的所有应抽取的项数） 3.F1=2＊（Precision＊Recall）/（Precision＋Recall）
	生成	长期	1．多轮	1.评估在长对话或多轮交互中，模型当前回复

	一致性	交互过程中的逻辑自洽程度	对话一致性得分（人工评估） 2．单文档／上下文内一致性（自动化／人工）	是否与自身之前陈述的事实、观点、承诺或角色设定保持一致，避免自相矛盾。 计算：人工评估者根据预设标准（如是否存在矛盾、是否保持角色设定等）对对话片段或整个对话进行评分（如 1-5 分）。计算平均分或通过率。 定义：评估在生成长文本（如故事、报告）时，模型输出内部在事实、逻辑、人物属性、时间线等方面是否自洽。 计算：可设计自动化方法检查简单矛盾（如属性冲突、时间冲突），但复杂一致性仍需依赖人工评估评分。
	语言流畅性	模型输出符合语法规范和语言习惯的程度	1．困惑度（Perplexity—PPL）—自动化 2．语法错误率（GER）—自动化／工具 3．人工流畅性评分	1.定义：衡量模型对测试文本概率分布的预测不确定性。数值越低，表示模型认为该文本越“自然”或越“可能”出现，通常意味着流畅性越高。 计算：基于语言模型本身在测试集上的计算得出。但需注意，低 PPL 不一定代表人类主观感受的流畅，且依赖训练数据分布。 2.（检测出的语法错误数量）／（总词数或总句子数）＊100%。使用语法检查工具（如 LanguageTool）实现。 3.人工评估者根据输出文本是否符合语法规范、是否自然地道、是否拗口难懂等进行评分（如 1—5 分）。计算平均分或通过率（如流畅性得分的比例）。最直接反映人类感知的流畅性。
	回答相关性	模型输出与问题相关的程度	语义相似度分数 人工相关性评分	1.计算模型回复与用户问题（Query）或对话上下文在语义层面的相关性分数（通常 0—1 或 0—100） 2.人工评估者判断模型回复是否直接、充分地回答了用户的问题，是否包含无关信息。进行评分（如 1—5 分）或二元判断（相关／不相关）。计算平均分或相关比例。

可靠性	抗干扰能力	模型面对干扰或噪声时输出稳定性	1．噪声注入后性能衰减率 2．输出稳定性得分	1.定义：在输入（文本、图像、语音、多模态）中加入特定类型和强度的噪声／干扰后，模型在核心任务（如分类准确率、问答 F1、图像描述 BLEU）上的性能下降比例。 计算：衰减率=[（干净数据性能－噪声数据性能）／干净数据性能］*100% 2.定义：对同一输入多次加入不同实例的同类型噪声，评估模型输出结果的一致性（如分类结果是否相同、生成文本的语义是否稳定）。 计算：稳定性得分=（输出一致的样本数）／（总扰动样本数）＊100%
	容错率	模型面对错误输入时处理能力	1．错误输入容忍度 2．崩溃／错误率	1.定义：当输入包含语法错误、拼写错误、逻辑错误、格式错误、无关信息、部分信息缺失等常见用户错误时，模型仍能正确完成任务或提供合理回复的比例。 计算：容忍度=（在错误输入上任务成功的样本数）／（包含错误输入的测试样本总数）＊100％ 2.定义：当输入是严重畸形、完全无关或恶意构造时，模型服务崩溃、抛出未处理异常或返回不可解析结果的比例。 计算：崩溃／错误率＝（导致服务崩溃或返回严重错误的畸形输入数）／（畸形输入测试样本总数）＊100％
	泛化能力	模型处理新数据、新场景的能力	1．域外／新场景性能衰减率 2．零样本／小样本新任务成功率	1.定义： 在与训练数据分布显著不同的新数据、新场景、新领域上测试时，模型在核心任务指标上的性能下降比例。 计算： 衰减率=[(域内/已知场景性能-域外/新场景性能)/域内/已知场景性能]*100% 2.定义： 模型在完全未见过任务类型或仅有极少量示例的情况下，根据任务描述或提示成功执行新任务的比例。 计算： 成功率=(成功完成的新任务实例

				数)/(新任务测试实例总数)*100%
	输入敏感性	输入细微变化导致输出变化的敏感程度	1．对抗样本攻击成功率 2．输入扰动输出变化率	1.定义： 对输入进行细微的、人眼/人耳难以察觉的对抗性扰动后，导致模型产生错误输出（如分类错误、生成有害内容）的比例。 计算： ASR=(成功导致模型出错的对抗样本数)/(生成的对抗样本总数)*100% 2.定义： 对输入进行微小的、语义保持的合法修改（如同义词替换、句式变换、图像轻微旋转裁剪）后，模型输出（尤其是关键决策或事实）发生非预期或过大变化的比例。 计算： * 分类/确定性任务： 变化率=(输出结果发生改变的扰动样本数)/(总扰动样本数)*100% (期望低) * 生成任务： 计算扰动前后输出的语义相似度(BERTScore, BLEURT)或人工评估变化是否合理。
	边界情况表现	模型在极端或稀有场景的表现	1．边界/极端案例通过率 2．边界案例失败严重性等级	1.定义： 在精心设计的、罕见、极端或压力测试场景下，模型行为符合预期（如安全、无害、合理）的比例 计算： 通过率=(行为符合预期的边界案例数)/(边界案例测试总数)*100% 2.定义： 定性/半定量指标，用于评估边界案例失败时后果的严重程度（例如：1-轻微无害错误，2-明显错误，3-产生误导，4-输出有害内容，5-系统崩溃/安全漏洞）。 计算： 人工评估失败案例，统计不同严重等级的比例。目标是尽量减少高严重等级（4,5级）的失败。
	毒害性评估	生成有害内容的比	1.有害内容生成率 2.敏感话题不当处	1.定义： 模型在给定提示（包括故意诱导和无诱导的正常提示）下，生成包含暴力、仇恨、歧视、骚扰、非法活动、自残鼓励、严重虚假信息等有害内容的输出比例。

安全性		例	理率	计算：有害内容生成率=(生成有害内容的样本数)/(测试样本总数)*100% 2.定义：当被问及特定敏感话题（如特定种族、宗教、性别、健康问题、政治事件）时，模型生成有偏见、刻板印象、歧视性、煽动性或严重不准确内容的比率。 计算：不当处理率=(在敏感话题上生成不当内容的样本数)/(敏感话题测试样本总数)*100%
	隐私保护程度	对敏感数据保护的有效性	1.训练数据隐私泄露风险 2.隐私泄露事件发生率	1.定义： 成员推理攻击成功率：攻击者判断特定数据样本是否被用于训练模型的成功率。 计算： 成员推理攻击成功率=(正确识别成员/非成员样本数)/(测试样本总数)*100% 2.定义：在模型部署运行过程中（如输入处理、日志记录、缓存、API 传输），发生意外泄露用户输入或输出中包含的敏感信息的事件比例。 计算：发生率=(发生泄露的事件数)/(总运行时间或总请求数) 目标应为 0。
	防攻击能力	模型抵御对抗攻击的能力	1.对抗攻击防御成功率/攻击失败率 2.提示注入攻击成功率	1.在遭受对抗性攻击（旨在欺骗模型产生特定错误输出或泄露信息）时，模型维持正确行为或未泄露敏感信息的比例。 与可靠性中的 ASR 互补。 计算：防御成功率=1-ASR 2.定义：攻击者通过在用户输入中嵌入恶意指令或覆盖系统提示，成功诱导模型执行非预期操作（如泄露系统提示、越权访问、生成有害内容）的比例。 计算：注入成功率=(成功实现攻击目标的注入样本数)/(注入攻击测试样本总数)*100%
	合规	符合法律	1.法规伦理符合性	1.定义：模型行为符合中国生成式 AI 服务管理办法、行业特定法规的条款比例或综合评分。

	性	法规或伦理标准的比例	得分 2.内容审核违规率	计算： 条款符合率：(满足的法规条款数)/(适用的法规条款总数)*100% 人工审计评分：由法律专家和审计员根据检查清单进行评分（如 0-100 分）。 2.定义：模型生成或处理的内容违反平台内容安全政策（如禁止仇恨言论、暴力极端主义、非法商品交易、儿童安全等）的比例 计算：违规率=(被审核系统或人工判定违规的样本数)/(测试样本总数)*100%
	输出解释性	模型输出结果的可解释程度	可解释性得分 2.解释满意度得分	定义：定性/半定量指标，评估模型决策是否可解释（提供依据）、系统行为是否透明（告知用户是 AI、能力限制）、是否披露数据来源和使用方式。 计算：根据预设的透明度和可解释性标准（如提供置信度、突出关键证据、说明能力边界），由专家或用户进行评分（如 1-5 分）。计算平均分。
交互性	透明度	模型决策过程可追溯性	决策追溯完整度 数据来源披露度 不确定性量化	1.定义：系统记录并展示关键决策节点（如检索来源、推理步骤、规则触发）的完整程度。 计算：人工审核日志，统计符合追溯标准的请求比例： 完整度=(完整记录决策链的请求数)/总请求数×100% 2.定义：输出中关键事实或结论是否标注来源（如训练数据文档 ID、检索结果引用）。计算：披露率=(提供可验证来源的输出数)/需来源披露的测试样本数×100% 3.定义：模型对输出置信度/不确定性的量化是否清晰可用（如概率值、置信区间、警告标志） 计算：人工评估符合标准的输出比例，或用户对不确定性提示理解率的问卷得分。
	可信度评估	模型输出可信	置信度校准误差 2.幻觉检	1.定义：模型自评置信度（如分类概率）与实际正确率的匹配程度（如预测 80%置信度时应有 80%的正确率）

		度估计的准确性	测准确率	计算：BrierScore，概率预测与二值结果的均方误差 2.定义：模型能否识别自身输出中的事实性错误（幻觉） 计算：准确率=(正确识别幻觉/非幻觉的样本数)/总样本数×100%（需构建含标注幻觉的测试集）
	可视化程度	模型内部机制可视化水平	可视化覆盖度 交互效率得分	1.定义：支持可视化解释的模型组件范围（例如思维链推理展示） 计算：覆盖度=(可可视化组件数)/总关键组件数×100% 2.定义：用户通过可视化工具探索模型行为的操作效率（如查找归因、切换视图的速度）。 计算：用户实验记录完成标准任务的平均时间/点击次数，转换为标准化分数（1-5 分）
	对话连贯性	持续对话过程中的逻辑连贯度	话题延续度 指代消解准确率	1.统计相邻对话轮次中核心实体/话题关键词的重叠率： 延续度=(含共同关键词的轮次对数)/总轮次对数×100% 2.测试模型解析代词（如“它”、“他们”）指代对象的正确率： 准确率=(正确解析的指代词数)/总指代词数×100%
应用性	上下文感知能力	有效利用历史上下文的能力	窗口内召回率 意图继承度	1.要求模型复现前 N 轮对话的特定信息（如日期/地名）： 召回率=(正确复现的信息数)/测试信息总数×100% 2.评估模型在多轮对话中维持核心意图的能力（如订酒店全程保持"预订"意图）： 继承度=(意图未偏移的轮次)/总相关轮次×100%
	个性化程度	根据用户历史	用户画像匹配度 跨对话偏	1.对比模型回应内容与预设用户画像（如兴趣/职业）的契合程度： 计算回应关键词与画像标签的余弦相似度（0-

		行为和偏好调整回应内容的能力	好一致性	1） 2.当用户透露新偏好时，后续推荐相关项的准确率： 准确率=(符合偏好的推荐数)/总推荐数×100%
	主动性	主动引导对话或推荐的能力	非请求推荐占比 问题预判准确率 打断纠正率	1.统计模型主动提出建议的频次： 占比=(主动推荐轮次数)/总轮次数×100% 2.评估模型预测用户潜在需求的正确性（如用户问航班时主动提示天气）： 准确率=(预判被用户认可的次数)/总预判次数×100% 3.当用户纠正模型时，立即调整策略的比例： 纠正率=(成功调整的打断次数)/总打断次数×100%
	领域适应性	模型快速适配特定领域的能力	1.微调效率提升率 2.少样本学习准确率	1.模型在少量领域数据微调后性能的提升效率：提升率=(微调后性能-基线性能)/微调所需数据量 2.模型仅用极少量样本（如 5~10 条）解决领域任务的能力：准确率=模型在目标领域少样本测试集上的正确预测比例

参考文献

[1]中国信息通信研究院.大模型基准测试体系研究报告[R].(2024-06).

[2]中移智库.“弈衡”多模态大模型评测体系白皮书[R].(2024-10).

[3]中华人民共和国国家市场监督管理总局，国家标准化管理委员会.人工智能大模型第 2 部分：评测指标与方法：GB/T45288.2-2025[S].(2025-02-28).

[4]上海人工智能实验室，上海财经大学，上海库帕思科技有限公司.金融大模型应用评测报告摘要版[R].(2024-12).

[5]哈尔滨工业大学.ChatGPT 调研报告[R].(2023-03-06).

[6]亚信科技，清华大学.AIGC 赋能通信行业应用白皮书(2023)[R].(2023-03).

[7]IDC 国际数据公司.2022 中国大模型发展白皮书[R].(2023-02).

[8]中国移动通信集团有限公司，中国电子技术标准化研究院，等．通用大模型评测标准[EB/OL]．(2024-01)．

中国移动通信集团有限公司

地址：北京市西城区金融大街 29 号

邮编：100033

联系电话：010-52686688

传真：010-52616230

网址：https://www.10086.cn