ICS 35.240
CCS L70

团 体 标 准

T/ISC XXX—XXXX

人工智能 大规模预训练模型总体技术要求及评估方法

Artificial intelligence—Technical requirements and testing methods for large scale pre-trained model

（征求意见稿）

2024-10-20

XXXX－XX－XX 发布 XXXX－XX－XX 实施

中 国 互 联 网 协 会 发布

目 次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国信息产业商会团体标准专业委员会提出并归口。

本文件起草单位：

本文件主要起草人：

本文件及其所代替文件的历次版本发布情况为：

——

引　言

近年来，随着数据资源的不断丰富以及深度学习计算能力的不断提升，大模型具备了强大的表征能力，在自然语言处理、图像识别、语音识别等领域都展现出了强大的性能和广泛的应用价值，正逐渐成为人工智能领域的研究热点和产业应用焦点。随着产业界的深入参与和合作，大模型的研发和应用从聚焦基础设施构建、数据获取、模型训练及部署等方面技术突破和工程化落地，到关注领域应用、行业应用等方面大模型服务能力评估。然而，针对整个大模型系统，业界还缺乏规范性的技术要求及评估体系，使得不同的研究机构和企业对大规模预训练模型系统难以开展标准化的评估。

制定一个统一、科学的大模型评估标准可以提供统一的标准化的功能与性能指标用于评估不同供应商研发生成的大模型，保证评估结果的客观性、准确性和可靠性，为人工智能大模型的采购方和应用方提供可参考的评估依据。促进我国人工智能产业生态的健康发展和行业的良性竞争。

人工智能 大规模预训练模型总体技术要求及评估方法

1 范围

本文件规定了大规模预训练模型系统参考架构及技术要求、大规模预训练模型系统角色、大规模预训练模型能力评估方法。

本文件适用于人工智能大规模预训练模型的设计、研发、评估和应用。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867-2022 信息技术 人工智能 术语

AIIA/PG 0074-2022 大规模预训练模型技术和应用评估方法 第4部分：可信要求

AIIA/PG 0141-2024 人工智能开发平台通用能力要求 第4部分：大模型技术要求

3 术语和定义

3.1

大规模预训练模型 large-scale pre-trained model

使用大规模数据集通过训练得到的，具备大参数量的深度学习模型。

3.2

大规模预训练模型系统 large-scale pre-trained model system

围绕大规模预训练模型构成的整个生态系统，包括基础设施、数据、模型、行业应用等子系统。

3.3

推理 inference

从给定的前提进行论证并得出结论。

[来源于：GB/T 41867-2022]

3.4

微调 fine-tuning

围绕大规模预训练模型构成的整个生态系统，包括基础设施、数据、模型、行业应用等子系统。

[来源于：GB/T 41867-2022]

3.5

理解能力 comprehension capability

大规模预训练模型在处理和分析大量数据时，对信息的深入洞察和解读能力。

3.6

生成能力 generative capability

大规模预训练模型基于已学习的知识和模式，自主产生新的内容或结果的能力。

3.7

推理能力 reasoning capability

大规模预训练模型理解和利用相关的证据或逻辑来推导结论或做出决策的能力。

3.8

知识集成能力 knowledge integration capability

一种基于事实证据的支撑，完成知识密集型任务的能力。

3.9

专业能力 professional capability

大规模预训练模型作为领域专家，用于解决特定领域任务的能力。如医疗领域、教育领域、法律领域等。

3.10

行业 industry

大规模预训练模型被应用的特定领域的行业，如通信、医疗、金融等。

3.11

场景 scenario

大规模预训练模型被应用到的具体环境和上下文，如推荐系统、自动驾驶、问答系统等场景。

3.12

任务 task

被调度的训练或推理对象。

3.13

模态 modal

指一切表达或感知事物的方式，每一种信息的来源或者形式，都可以称为一种模态，如文本、图像、音频、视频等。

3.14

多模态 multi-modal

多种感官信息在一起协同作用。

3.15

自动化评估 automated evaluation

指使用计算机算法和预定义评估指标来自动执行模型评估任务，无需或几乎无需人工干预。

3.16

人工评估 human evaluation
指由人类评估者（如人类专家）手动执行模型评估任务。

4 符号和缩略语

下列符号和缩略语适用于本文件。
AI：人工智能（Artificial Intelligence）
BLEU：双语评估替补（Bilingual Evaluation Understudy）
CCL：集合通信库（Collective Communication Library）
CI：清晰度指数（Clarity Index）
CPU：中央处理器（Central Processing Unit）
CV：计算机视觉（Computer Vision）
EER：等错误率（Equal Error Rate）
FPGA：可编程门阵列（Field-Programmable Gate Array）
FID：FID分数/弗雷歇感知距离（Fréchet Inception Distance）
GPU：图形处理器（Graphics Processing Unit）
HBM：高带宽内存（High Bandwidth Memory）
IB：无限带宽（InfiniBand）
IoU：交并比（Intersection over Union）
NPU：神经网络处理器（Neural Processing Unit）
NLP：自然语言处理（Natural Language Processing）
NI：自然度指数（Naturalness Index）
PER：因素识别错误率（Phone Error Rate）
RDMA：远程直连内存访问（Remote Direct Memory Access）
RoCE：基于聚合以太网的RDMA（RDMA over Converged Ethernet）
ROUGE：基于召回率的摘要评估方法（Recall-Oriented Understudy for Gisting Evaluation）
WER：语音识别错误率（Word Error Rate）

5 概述

本文件的标准化对象包括大规模预训练模型及其系统。其中第六章规范了大规模预训练模型系统的参考架构和系统角色；七章在此基础上规范了大规模预训练模型系统的基础设施层、数据层、模型层、应用层、系统安全层的技术要求；第八章对大规模预训练模型的能力评估方法进行规范，包括评估框架、评估指标、评估数据集及评估流程，第八章的评估方法仅针对大规模预训练模型本身，不包含大规模预训练模型系统的其他组成部分。

6 大规模预训练模型系统参考架构

6.1 参考架构

大规模预训练模型系统参考架构包括基础设施层、数据层、模型层、应用层、系统安全层，如图1所示。

a) 基础设施层：用于支撑大规模预训练模型系统运行，包括软件部分和硬件部分。软件部分主要包括深度学习框架、分布式训练框架、运行加速库、通信库等。硬件部分主要包括算力资源、存储资源、网络资源。

b) 数据层：用于实现大规模预训练模型数据接入与处理，包括数据接入、数据预处理、数据集构造及数据集管理。

c) 模型层：用于支撑大规模预训练模型训练、微调，验证并进行模型部署、推理和汇聚。包括训练微调部分、部署推理部分和模型纳管部分，训练微调部分包括模型预训练、模型微调、模型测试验证。部署推理部分包括模型压缩、模型部署、模型推理。模型纳管部分包括集合通用大模型、行业大模型、专业大模型。

d) 应用层：用于支撑大规模预训练模型不同领域的应用，包括通用应用部分和行业应用部分。通用应用部分主要包括NLP、CV、语音、多模态等通用领域应用，行业应用部分主要包括通信、政务、医疗、能源等行业领域应用。

e) 系统安全层：用于支撑大规模预训练模型系统的安全可信与合规，主要包括基础设施安全可信、数据安全可信、模型安全可信、服务安全可信、内容安全可信等。

图1 大规模预训练模型系统参考架构

6.2 系统角色

6.2.1 供给者

供给大规模预训练模型系统运行的基本生产要素，包括供给数据、供给算力、供给通用和行业大规模预训练模型。

6.2.2 汇聚者

汇聚大规模预训练模型系统中的关键生产要素，包括汇聚国产训练推理芯片、国产训练框架等软硬件基础设施、汇聚大数据、汇聚大规模预训练模型训练、微调、部署推理、评估等服务。

6.2.3 运营者

管理和优化大规模预训练模型的运营流程，通过运营管理、服务优化、性能监控、资源调配、风险管理等方式确保大规模预训练模型系统在通用 NLP、CV、语音等通用应用场景和通信、政务、医疗、能源等行业应用场景的高效赋能与持续创新。

7 大规模预训练模型系统技术要求

7.1 基础设施层

7.1.1 硬件

7.1.1.1 算力资源

应支持至少二种类型计算芯片作为算力基础设施，如CPU、GPGPU、DSA、NPU、TPU、AISC、FPGA 等。

7.1.1.2 存储资源

a) 应支持至少一种存储介质，如 SSD、机械硬盘、HBM 等；
b) 应支持至少一种存储接口，如 SATA、PCIe 等；
c) 应支持至少一种存储方式，如对象存储、多级存储、文件存储、块存储等；
d) 应支持至少一种存储连接方式，如 SAN、NAS、DAS 等。

7.1.1.3 网络资源

a) 应支持至少一种网络通信标准，如 InfiniBand、RoCE 等；
b) 宜支持 RDMA 网络通信技术；
c) 宜支持至少二种卡间通信协议，如共享内存、PCIe、NVLink、MTLink 等。

7.1.2 软件

7.1.2.1 深度学习框架

a) 应支持至少一种深度学习框架，如 PyTorch、TensorFlow、Caffe、MindSpore、PaddlePaddle、MXNet 等；
b) 宜支持至少一种神经网络交换格式，如 ONNX 等。

7.1.2.2 分布式训练框架

a) 应支持至少一种跨深度学习框架通用分布式训练框架，如 DeepSpeed、Megatron-LM、Colossal-AI、BMTrain、AscendSpeed 等；
b) 应支持至少一种深度学习框架内嵌式分布式训练框架，如 PyTorch、TensorFlow、MindSpore、PaddlePaddle 等。

7.1.2.3 运行加速库

宜支持提供的运行加速库支持至少一种模型开发任务，如模型训练、模型微调、模型压缩、模型推理等。

7.1.2.4 通信库

a) 宜支持提供的通信库支持至少一种模型开发任务，如模型训练、模型微调、模型压缩、模型推理等。
b) 宜支持集合通信能力。

7.2 数据层

7.2.1 数据接入

宜支持 AIIA/PG 0141-2024 中 6.1.1 节数据接入技术要求。

7.2.2 数据预处理

宜支持 AIIA/PG 0141-2024 中 6.1.2 节数据预处理技术要求。

7.2.3 数据集构造与管理

宜支持 AIIA/PG 0141-2024 中 6.1.3 节数据集构造与 6.1.4 节数据集管理技术要求。

7.3 模型层

7.3.1 训练微调

7.3.1.1 模型预训练

a) 应支持至少一种预训练方法，如从头预训练、继续预训练等；
b) 应支持训练中的断点处理，如断点保持、断点续训、断点重训等；
c) 应支持至少一种分布式训练方法，如数据并行、模型并行（流水线并行，张量并行）、混合并行、MOE 并行等；
d) 应支持至少一种训练优化技术，如 ZeRO、混合精度训练等。

7.3.1.2 模型微调

应支持至少一种模型微调方法，如全参微调、低参微调（Lora）、指令微调等。

7.3.1.3 模型测试验证

a) 宜支持包含不同数据分布、场景和类别的训练数据集、微调数据集、评估数据集，以验证模型的泛化能力；
b) 宜根据模型的任务类型选择合适的评估指标，如准确率、召回率、F1 分数、AUC-ROC、BLEU 等中的一项或多项；

c) 宜支持与相关领域内的基准模型进行对比，以衡量模型性能优劣；
d) 宜支持根据测试结果对模型进行调优，如调整超参数、优化模型结构等，以提高模型性能；
e) 宜支持错误分析与诊断，对于测试中出现的错误和异常，应进行详细的分析和诊断。

7.3.2 推理部署

7.3.2.1 模型压缩

a) 应支持至少一种模型压缩方法，如低比特量化、感知量化训练、训练后量化等；
b) 宜支持至少一种模型压缩效果显示，如压缩比显示、压缩前后精度差异显示、压缩前后性能差异显示等。

7.3.2.2 模型部署

a) 应支持至少一种模型部署格式，如以镜像方式进行部署等；
b) 宜支持至少一种模型部署方式，如云端部署、边端部署等；
c) 宜支持分布式模型部署。

7.3.2.3 模型推理

a) 应支持至少一种推理加速框架，如 TensorRT-LLM、DeepSpeed-MII、Triton、vLLM 等；
b) 应支持至少一种推理优化技术，如缓存优化、并行化推理、异步化推理等；
c) 宜支持对推理服务反馈数据的回流，以用于循环迭代。

7.3.3 模型纳管

a) 应支持纳管至少两种大模型，包括通用大模型、行业大模型、专用大模型；
b) 注：通用大模型可支持语言、视觉、语音等不同模态通用任务。行业大模型，包括例如通信、政务、医疗等行业。专用大模型，包括例如办公大模型、客服大模型等。
c) 宜支持汇聚模型的二次开发需求；
d) 宜支持对多个模型进行协同调度和推理；
e) 宜支持对模型服务进行监控，如基础资源监控（如计算、存储、网络等）、故障异常监控等；
f) 宜支持至少一种模型服务监控方式，如可视化方式、后台日志方式等。

7.4 应用层

7.4.1 通用应用

7.4.1.1 单模态

a) 应具备 NLP 领域能力。如：
   1) 应具备文本理解能力，并支持至少三种通用文本理解任务，如文本分类、命名实体识别、信息抽取、文本问答、代码理解等；
   2) 应具备文本生成能力，并支持至少三种通用文本生成任务，如摘要生成、机器翻译、文本改写、代码生成等；
   3) 应具备文本推理能力，并支持至少一种通用文本推理任务，如逻辑推理、数学推理、任务分解等。
b) 应具备 CV 领域能力，如应具备视觉理解能力，并支持至少三种通用图像理解任务，如图片分类、图片分割、目标检测、视频分类、行为识别等。

c) 应具备语音理解能力，并支持至少两种通用语音理解任务，如声纹识别、环境音分类等。

7.4.1.2 多模态

a) 应具备多模态理解能力，并支持至少五种多模态理解任务，如图文检索、图片回答、视觉空间关系、视觉语言推理、视觉蕴含、视频检索、视频问答等图文理解任务，文音检索等文音理解任务，以及有声视频检索，有声视频问答等图文音理解任务；
b) 应具备多模态生成能力，并支持至少五种多模态生成任务，如文本生成图片、图片生成文本、文本生成视频、视频生成文本等图文生成任务，文本生成有声视频、视频生成文本等图文音生成任务，以及语音合成、语音识别、语音翻译等文音生成任务。

7.4.2 行业应用

7.4.2.1 通信

a) 应支持至少两项通信行业通用场景，如客服场景、运营场景、知识问答场景、数据分析场景等；
b) 应支持至少三项通信行业专业场景，如网络规划、网络建设、网络维护、网络优化、网络运营、网络资源管理等 。

7.4.2.2 政务

a) 应支持至少两项政务行业通用场景，如一网统管、一网通办、一网协同；
b) 应支持至少三项政务行业专业场景，如行政服务、公安服务、人社服务、财税服务、市场监管、经济监管等。

7.4.2.3 医疗

应支持至少三项医疗行业场景，如医疗知识查询、医疗文档理解、健康问答、智能导诊、、查房问诊、医生助手等。

7.4.2.4 能源

a) 应支持至少两项能源行业通用场景，如客服场景、运营场景、知识问答场景、数据分析场景、会议纪要场景、公文写作场景等；
b) 应支持至三项能源行业专业场景，如矿山安全生产、发电厂安全生产、工人排班调度、煤炭营销规划等 。

7.5 系统安全

宜支持AIIA/PG 0074-2022中给出的基础设施安全可信、数据安全可信、模型安全可信、服务安全可信、内容安全可信等技术要求。

8 大规模预训练模型能力评估方法

8.1 评估框架

大规模预训练模型能力评估框架主要由评估对象、评估能力和任务、评估指标 3 个维度组成，如图 2 所示，其中

a) 评估对象：指评估的大规模预训练模型类型，如语言大模型、视觉大模型、语音大模型、多

模态大模型等不同模态模型或通信大模型、政务大模型等不同行业模型；

b) 评估能力和任务：指评估的大规模预训练模型的能力项及任务项，能力项主要包括理解能力、生成能力，推理能力等，每个能力项包含多个任务，如理解能力包含文本分类、命名实体识别等任务，生成能力包含机器翻译、摘要总结等任务、推理能力包括逻辑推理、数学推理等任务。每个任务项由相应评估指标进行评估；

c) 评估指标：指对大规模预训练模型进行能力评估的关键指标，分为客观指标和主观指标。客观指标指基于模型预测结果与实际值之间的量化差异来评估模型性能的指标，不受人的主观感受或偏见影响，能够提供客观、准确的模型性能评估，如准确率、F1 值、BLEU、ROUGE 等。主观指标指基于评估者个人感受、评价或专业判断来评估模型性能的指标，受到评估者主观因素（如经验、知识、偏好等）的影响，如相关性、完整性、有效性、连贯性等。具体指标见 8.2 节。

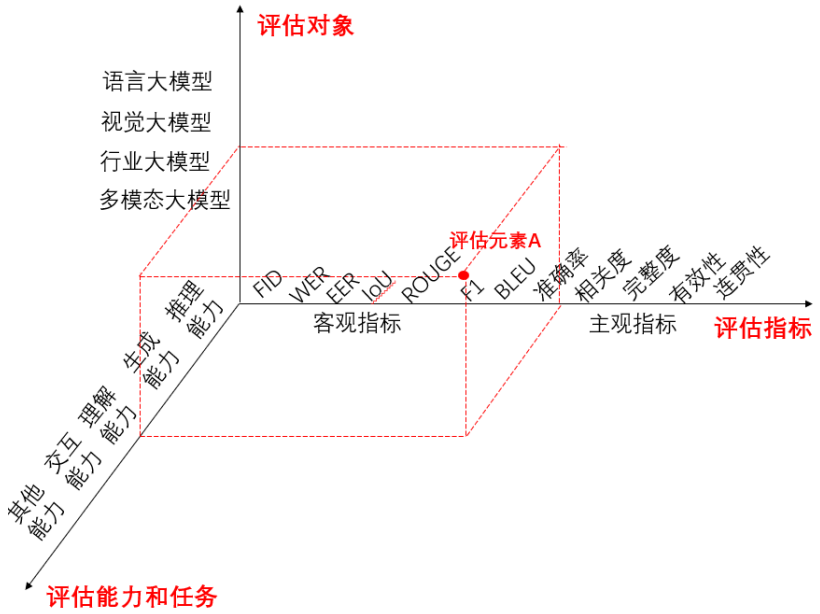

图 2 评估框架

整体评估范围是由上述三个维度确定，整体评估结果是由评估范围内的所有评估元素的评估结果综合确定。具体评估方式主要包括利用工具进行自动化评估，基于人类评估者进行人工评估、针对不同版本模型的对比评估以及针对用户反馈分析（例如点赞、点踩）的用户反馈评估等。

如图 2 评估元素 A 对应某行业大模型将准确率作为评估指标对理解能力的评估结果。具体行业大模型的评估案例见附录 C。

8.2 评估指标

大规模预训练模型能力评估维度主要分为理解能力、生成能力、推理能力及专业能力，理解能力主要包括文本理解、图像理解、音频理解 3 个单模态理解能力，以及图文理解、文音理解 2 个多模态理解能力。生成能力主要包括文本生成单模态生成能力，以及图文生成、文音生成 2 个多模态生成能力。推理能力主要包括文本推理单模态推理能力。

能力评估维度下对应具体评估任务与评估指标，详细说明见表 1：

表 1 评估指标

能力项	能力子项	任务项	任务项说明	评估指标

理解能力	文本理解	文本分类	将文本划分为不同的类别或标签。可以应用于垃圾邮件过滤、情感分析、新闻分类等应用场景。	准确率
		命名实体识别	识别文本中的实体，如人名、地名、组织机构、日期等。这对于信息提取和语义分析非常重要。	准确率
		文本问答	从给定的文本或知识库中提取相关信息，经过分析和推理，生成准确、简洁的答案回答用户提出的问题。可以应用于如智能客服、搜索引擎等应用场景。	准确率
	图像理解	图片分类	指模型能够理解图片的语义内容，并输出其对应的类别标签。	准确率
		图片分割	把图片分成若干个特定的、具有独特性质的区域并提取感兴趣目标的技术和过程。	准确率
		目标检测	在图片中检测和定位特定的目标物体。	IoU
	音频理解	声纹识别	是把声信号转换成电信号，再用计算机进行识别，包括说话人辨认和说话人确认。	准确率
	图文理解	图文检索	指模型能够根据给定的图片/文本检索到与之最匹配的文本/图片构成配对。	准确率
		图片问答	指模型能够回答针对图片的文本问题。	准确率
		视觉空间关系	指模型能够基于图片内容正确判断文本中所描述的对象间位置关系。	准确率
	文音理解	文音检索	指模型能够根据给定的音频/文本检索到与之最匹配的文本/音频构成配对。	准确率
生成能力	文本生成	机器翻译	模型能够理解文本指令，将文本从一种语言翻译成另 一种语言。	BLEU、ROUGE、主观指标等
		摘要总结	模型能够理解文本并根据输入内容生成相应摘要总 结。	准确率、ROUGE、主观指标等
		文本改写	模型将文本从一种表述方式改写成另一种表述方式。	METEOR、主观指标等
		代码生成	模型能够理解文本指令，生成符合其要求的编程代码。	准确率
	图文生成	文生图	模型能够理解文本指令，生成符合其要求的图片。	FID、CLIPScore、AestheticScore、主观指标等
		图生文	指模型能够对图片的内容进行概括总结，生成合理的 文本描述。	BLEU、主观指标等
	文视频生成	文生视频	模型能够理解文本指令，生成符合其要求的视频。	主观指标
		视频生文	模型能够理解视频内容，并生成符合要求的文本形式描述，如视频摘要等。	BLEU、ROUGE、主观指标等
	文音生成	语音翻译	模型能够理解输入语音及其语言，并将其翻译为指定语言所对应的语音。	WER、PER
		语音识别	模型能够理解输入的语音，并将其转录为对应的文本。	WER、EER
		语音合成	模型可以根据指定文本生成对应的语音。	CI、NI、SNR、主观指标等

推理能力	文本推理	逻辑推理	指模型根据已有的事实或知识，如上下文信息、常识、定理等，完成数学、符号、逻辑推理过程，并形成合乎逻辑结果。	准确率
		数学推理	指把表示关系的运算方法、逻辑术语运用于研究对象，得到数学的结论或者验证数学的结论。	准确率
行业能力	根据行业应用场景并结合上述通用能力项及评估指标要求确定。 本文件不对行业能力作规范性要求，附录 C 中给出了通信等行业大模型评估案例。			

具体客观和主观评估指标定义以及主观指标的评分方法参见附录 A。

8.3 评估数据集要求

大规模预训练模型的能力评估，依托于高质量的评估数据集。需要从评估数据集的全面性、多样性、均衡性等方面综合考察以选择合适的评估数据集。

a) 评估数据集的全面性：指数据集覆盖广泛的主题和领域。其中评估文本理解能力的数据集需要包含多领域文本、多语言文本等；评估图像理解能力的数据集需要包含多场景图像、多风格图像等；评估音频理解能力的数据集需要包含多音色音频等；评估图文理解能力的数据集需要包含多问题类型等；评估文音生成能力的数据集需要包含多领域及语言文本、多音色及音调音频等；评估文本推理能力的数据集需要包含多推理形式等。

b) 评估数据集的多样性：指数据集具有多种问题形式，包括选择问答题、开放问答题、半开放问答题、填空题、判断题等多种形式的题目。

c) 评估数据集的均衡性：指数据集中不同难度、不同类别等的数据分布均衡，以避免由数据分布不均而导致得评估结果偏差。其中评估文本理解能力的数据集需要保持问题难度与类型等的均衡；评估图像理解能力的数据集需要保持不同类型对象、不同信息量图像等的均衡；评估音频理解能力的数据集需要保持不同质量音频、不同音色音频等的均衡；评估文本推理能力的数据集需要保持推理问题类型等的均衡。

d) 评估数据集的质量：指数据集需要经过清洗和标注。其中数据清洗需要针对缺失数据、异常数据、重复数据和敏感数据进行处理。数据标注则需要确保文本类数据具有语言、主题、情感倾向、命名实体等准确标注；确保图像类数据具有类别、对象位置、语义分割等准确标注；确保音频类数据具有所属语言、情感倾向等准确标注。

业界常用评估数据集参见附录 B。

8.4 评估流程

针对大规模预训练模型能力评估，给出统一和规范的评估流程，如图 3：

图 3 评估流程

a) 确定评估对象：确定评估的大规模预训练模型类型，比如语言大模型、视觉大模型、多模态大模型等通用大模型以及通信大模型、政务大模型、客服大模型等行业大模型。

b) 确定评估能力项及任务项：根据 a）中选择的大规模预训练模型类型，确定评估的模型能力

项以及相应的任务项。比如理解能力以及理解能力包含的文本分类、命名实体识别等任务，生成能力以及生成能力包含的机器翻译、摘要总结等任务、推理能力以及推理能力包含的逻辑推理、数学推理等任务等。

c) 确定评估指标：根据 b）中确定的评估任务项，选择合适的评估指标，比如文本分类、命名实体识别等任务对应的准确率指标，机器翻译、摘要总结等任务对应的 BLUE、ROUGE 等客观指标以及完整度、连贯性等主观指标，逻辑推理、数学推理等任务对应的准确率指标等。

d) 确定评估数据集和评估方式：根据 b）中确定的评估任务项，选择合适的评估数据集，比如理解任务对应的 CLUE、GLUE、DROP 等数据集、生成任务对应的 APPS、CodeXGLUE 等数据集、推理能力对应的 MATH、GSM8K 等数据集、以及综合任务对应的 MMCU、MMLU 等数据集。当缺少选定任务的数据集时，需专门构建相应评估数据集。根据 c）中确定的评估指标选择合适的评估方式，包括准确率、F1 值、BLUE、ROUGE 等客观评估指标对应的自动化评估方式，完整度、有效性、连贯性等主观评估指标对应的人工评估方式，GSB 指标对应的内部对比评估方式，以及用户点赞点踩指标对应的用户反馈评估方式等。

e) 输出评估结果：根据 d）中确定的评估数据集和评估方式，对 c）中确定的评估指标进行评估，得到 b）中确定的任务项即评估元素的评估结果，根据任务的不同优先级或比重确定任务评估结果的权重， 将不同评估元素的评估结果经过加权平均等方式综合处理并输出，得到 a）中确定的大规模预训练模型在 b）中确定的能力项上的评估结果。

f) 通过对客观指标的计算可以直接得到客观指标评估结果，主观指标的评估结果可以通过主观指标评分方法得到，具体评估方法参见附录 A.3。如需对主客观指标评估结果进行综合处理，可以将主观指标评估结果和客观指标评估结果进行加权求和得到综合评估结果。

附 录 A
（资料性）
评估指标

A.1 客观评估指标

A.1.1 准确率（Accuracy）

指正确分类的样本数与样本总数之间的比例，计算公式为：

$$A_{cc} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad \text{.....（1）}$$

其中
TP--真正的正样例数量；
TN--真正的负样例数量；
FP—错误的正样例数量；
FN—错误的负样例数量。
模型回答正确的测例占所有测例的比例。

A.1.2 F1值

指精确率（precision）和召回率（recall）的调和平均值，其中，精确率是模型判断为正样例中，真正为正样例的比例，要求模型做出正例判断时尽可能准确，召回率是所有实际为正样例的样本中，被模型正确识别为正样例的比例，要求模型尽可能找出所有的正例。计算公式为：

$$P = \frac{TP}{TP+FP} \times 100\% \quad \text{.....（2）}$$

$$R = \frac{TP}{TP+FN} \times 100\% \quad \text{.....（3）}$$

$$F1 = \frac{2\times P\times R}{P+R} \times 100\% \quad \text{.....（4）}$$

其中
TP--真正的正样例数量；
TN--真正的负样例数量；
FP—错误的正样例数量；
FN—错误的负样例数量。

A.1.3 BLEU

一种衡量机器翻译任务中模型生成的译文（reference）与参考译文（candidate）之间相似程度的指标。主要基于准确率，其取值范围为[0,1]，若模型生成的译文和参考译文完全匹配，BLEU是1；反之若完全不匹配，则为0。 计算公式为：

$$BLEU = BP \times e^{\sum_{n-1}^{N} w_n \log P_n} \quad \text{.....（5）}$$

$$BP = \begin{cases} 1, & c > r \\ e^{\frac{1-r}{c}}, & c \le r \end{cases} \quad \text{.....（6）}$$

式中BP是“过短惩罚”函数，其取值范围是(0,1]，模型生成内容越短，越接近0。c表示模型生成译文，r表示参考译文。w_n表示权重，通常取值1/n，P_n表示 n-gram 精度。1-gram 准确率可用于衡量模型生成译文单词的准确性，更高阶 n-gram 的准确率可用来衡量模型生成译文句子的流畅性。

A.1.4 Rouge

一种用于评估文本摘要任务中模型生成摘要质量的指标，主要关注模型生成译文是否捕捉到参考译文的信息。计算方法是评估参考译文中有多少n-gram出现的模型生成的译文中。

相较BLUE指标更着重信息的完整性，偏向召回率。

A.1.5 FID

一种用于评估生成模型生成图像质量的指标。FID衡量了生成图像在特征空间中的分布与真实图像在特征空间中的分布之间的距离。具体来说，FID通过计算生成图像和真实图像在预训练的深度网络的特征表示之间的均值和协方差差异来得出。

FID不需要人工标注的数据，是一种自动化的评估方法。

A.1.6 WER

是语音识别和自然语言处理中的一个重要评价指标，用于衡量自动语音识别系统生成的转录文本与参考文本之间的差异。WER 衡量的是两个文本序列中词错误的数量占参考文本中词的总数的比例。计算公式为WER= (插入错误的词数 + 删除错误的词数 + 替换错误的词数) / 参考文本中的词数。WER 越低，表示自动语音识别系统的性能越好，转录文本与参考文本之间的差异越小。

A.1.7 EER

用于声纹识别任务，表示在一个分类或识别系统中，当误报率和拒真率相等时的错误率。误报率是指系统将非目标对象错误地识别为目标对象的比例，而拒真率则是指系统将目标对象错误地识别为非目标对象的比例。

A.1.8 IoU

一种计算机视觉任务中用于评估目标检测、图像分割等任务性能的常用指标。在目标检测任务中，IoU衡量的是预测边界框与真实边界框之间的重叠程度，计算公式为预测边界框与真实边界框的交集面积除以它们的并集面积，IoU的值域为[0,1]，其中1表示完全重叠，0表示没有重叠。在图像分割任务中，IoU用于评估预测分割结果与真实分割结果之间的重叠程度。

A.1.9 CLIPScore

评估图文一致性的指标。将输入的图像和 Prompts 放入到 CLIP 特征提取器中获取 embedding，然后计算两者的余弦距离来评估图像和文字的相似程度。

A.1.10 AestheticScore

基于 CLIP 模型的线性估计器，旨在预测图像的审美质量。使用 LAION AI 训练的 aesthetic-predictor进行图像美学质量打分，得到分值。

A.2 主观评估指标

A.2.1 相关度

指回答与对话上下文的关联程度。

A.2.2 完整度

指生成的回答是否有信息缺失遗漏。

A.2.3 有效性

生成回答的有用程度。

A.2.4 连贯性

回答是否符合对话流程。

A.2.5 图文一致性

是评估图片在多大程度上与文字描述内容一致。

A.2.6 图片质量

是在不考虑图文一致性的情况下，评估生成图片的清晰度、色彩、构图等影 响主观感受的因素。

A.2.7 总体印象

评测人员对视频的直观感觉和总体印象。

A.2.8 真实性

视频是否像 AI 生成的，若像则是不真实。若视频是展现现实场景，则看是否与真实世界相符合；若是展现超现实场景，则看是否符合对动画、科幻电影等超现实场景的认知。

A.2.9 视频质量

视频内容在清晰度、流畅度等方面的表现程度。

A.2.10 美学质量

视频内容在布局构图、色彩搭配、艺术性、和谐性、景深和细节呈现等方面的综合表现。

A.3 主观指标评分方法

根据人工评估的指标维度，由参与者以分数的形式来进行评分。评估方法为按照指标维度对数据集中每条数据分别评分，并计算得到最终平均得分结果。以A 1.2.1~A 1.2.4评估指标为例，具体评分方法见表A.1：

表A.1 主观评估指标评分方法

分数	总体	相关度	完整度	有效性	连贯性
5分	回答正确且质量 高，结果真实，无冗余，非常符合用户期望。	生成的内容与prompt内容高度切合，没有不相关内容。	生成的内容完全和用户的意图对应，无任何信息缺失遗漏。	生成的内容全部有用，不存在重复冗余等影响有效性的内容。	回答对话流程连贯，回答内 容之间的连接质量非常高， 完全没有内 容 的 任 意 堆砌。

4分	大部分回答正确，结果真实，存在部分非关键错误，正确部分符合用户期望。	生成的内容与prompt内容的切合度在80%以上，存在少量不相关内容。	生成的内容有部分存在信息的缺失遗漏，对整体内容理解影响较小。	生成的内容80%以上有用，存在少量无用信息。	回答对话流程连贯性一般，回答内容之间的连接质量 一般，存在部分信息内容的堆砌。
3分	大部分回答不正确或结果不真实，存在部分关键错误，只有很少一部分符合用户期望。	生成的内容与prompt内容的切合度在60%以上，存在较多的不相关内容。	生成的内容有60%的信息缺失，对整体内容理解影响较大。	生成的内容60%以上有用，存在较多的无用信息。	回答对话流程连贯性较差，回答内容之间的连接质量较差，存在大部分信息内容的堆砌。
2分	有结果，但回答基本错误或回答相关度很低。	生成的内容与prompt几乎无关，好像理解用户意图又好像不理解，乱说。	生成的内容有80%的信息缺失，只有少数部分可以理解。	生成的内容80%以上无用，存在少量有用信息。	回答对话流程不连贯，回答 内容个别部分之间存在连接性，但绝大部分信息内容任意堆砌。
1分	结果为空、完全错误或回答无关。	生成的内容与prompt要求完全没有相关性，脱离用户意图。	生成的内容信息缺失严重或为空，导致无法理解。	生成的内容无用或几乎无用。	回答内容之间完全没有连接性可言，信息内容任意堆砌。

附 录 B
（资料性）
评估数据集

B.1 通用评估数据集

B.1.1 综合能力评估数据集

B.1.1.1 MMCU

MMCU是用于衡量中文大规模预训练模型处理多任务准确度的数据集，包含了来自医学、法律、心理和教育领域的单/多项选择问题。这些问题是由专业人员从免费提供的在线资源中手动收集而来，包括大学医学考试、国家统一法律职业资格考试、心理咨询师考试、心理学专业研究生入学考试和中国高考等。

MMCU通过计算模型在所有任务上的zero-shot和few-shot准确率来评估模型性能。

B.1.1.2 MMLU

MMLU是用于衡量模型处理多任务准确度的数据集，与MMCU类似。该数据集涵盖了人文学科、社会科学、自然科学和其他重要领域，工具有57个任务，15908个多选问题。

B.1.2 理解能力评估数据集

B.1.2.1 CLUE

CLUE是一个中文语言理解的评估基准，涵盖了多种不同难度、不同大小形式的句子分类和阅读理解任务。其提供了多种任务下的评估数据集，以及一个由语言学家开发的评估数据集（包含多种语言现象）。

B.1.2.2 GLUE

GLUE类似于CLUE，也是一个语言理解的评估基准。其也提供了一个数据集，用于评估模型在各种语言现象方面的表现。在该数据集上采用R_3（一种Matthews相关系数的推广）进行评估。

B.1.2.3 DROP

DROP是一个复杂英文阅读理解基准数据集，需要对段落内容进行离散推理。

B.1.2.4 SQuAD

SQuAD斯坦福问答数据集，是一个用于评估模型阅读理解能力的数据集。其包含107785个问题-答案对，涵盖了536篇文章。

B.1.2.5 RACE

RACE是一个用于评估模型阅读理解能力的数据集。该数据集从中国初高中学生的英文考试中收集而来，包含近28000篇文章和近100000个由人类专家（英语教师）生成的问题。特别地，RACE中需要推理的问题的比例要比其他阅读理解基准数据集大得多。

B.1.2.6 DuoRC

DuoRC是一个评估模型阅读理解能力的数据集。该数据集包含186089个问题-答案对，这些问答对是从7680对电影情节中创建的，每对情节来自于同一部电影的两个版本（从一个版本的情节中创建问题，并从另一个版本中提取或合成答案）。DuoRC从设计上确保了问题和其对应答案在片段之间几乎没有词汇重叠。此外，由于这两个版本具有不同的情节细节、叙述风格、词汇等，因此从第二个版本回答问题需要更深入的语言理解和融入外部背景知识。

B.1.2.7 WDW

WDW是一个评估模型阅读理解能力的数据集，包含超200,000个填空多项选择问题。WDW是通过LDC English Gigaword新闻语料库构建的填空式数据集，其选择两篇描述同一事件的新闻文章，将其中一篇生成段落，另一篇则生成问题。

B.1.2.8 TriviaQA

TriviaQA是一个评估模型阅读理解能力的数据集。该数据集包含650,000个问题-答案-证据三元组，其中问题和相应答案-证据句子之间具有相当大的语法和词汇变异性，并且需要更多的跨句子推理来找到答案。

B.1.2.9 WIKIQA

WIKIQA是一个用于开放域问答的数据集。该数据集是以一种自然而现实的方式构建，其包含3,047个问题，最初是从Bing查询日志中采样得到的。

B.1.3 生成能力评估数据集

B.1.3.1 APPS

APPS是UCB开发的，用于评估模型代码生成能力的数据集。APPS包含从不同开放访问编码网站（如Codeforces、Kattis等）收集的10,000个平均问题长度为293.2个单词的编程问题，这些问题涵盖了各个难度级别，包括简单的入门问题、面试级别的问题和编程竞赛。

APPS使用“测试用例平均值”和“严格准确性”这两个指标评估模型的表现。

B.1.3.2 CodeXGLUE

CodeXGLUE是微软亚洲研究院开发的针对代码理解和生成的基准数据集。CodeXGLUE包括14个数据集以及10个任务，涵盖了以下场景：（1）代码-代码：克隆检测、缺陷检测、填空测试、代码补全、代码修复和代码到代码翻译；（2）文本-代码：自然语言代码搜索、文本到代码生成；（3）代码-文本：代码摘要；（4）文本-文本：文档翻译。

B.1.4 推理能力评估数据集

B.1.4.1 C-Eval

C-Eval是一个评估基础模型高级知识和推理能力的综合性中文评估数据集。它包含13948个具有四个难度级别（初中、高中、大学和专业）的多项选择问题，涵盖了从人文到科学到工程的52个不同学科领域。

C-Eval通常使用准确率（Accuracy）作为评估模型的指标。

B.1.4.2 GSM8K

GSM8K是OpenAI开发的，用于评估模型数学推理方面能力的数据集。GSM8K包含8.5K个高质量、语言多样的小学数学问题。这些问题通常需要2到8个步骤来解决，主要涉及加减乘除等基本运算来得到最终答案。

评估指标通常采用准确率。

B.1.4.3 MATH

MATH是UCB开发的，用于评估模型解决数学问题的能力。MATH包含了125000个具有挑战性的数学问题，这些问题来自MAC、AIME竞赛。由于MATH具有较大的调整性，因此模型可能首先需要在数学基础知识方面进行充分的训练。

评估指标通常采用准确率。

B.1.4.4 HotpotQA

HotpotQA是一个大规模的问答数据集，用于评测模型多跳推理及为答案提供解释的能力。该数据集包含113,000个基于Wikipedia的问答对，这些问答对具有以下四个关键特点：（1）问题需要查找和推理多个支持文档才能回答；（2）问题多样化，不受任何现有知识库或知识架构的约束；（3）提供句子级的支持性事实以进行推理；（4）提供一种新型的事实比较问题，用于评测模型提取相关事实并进行必要比较的能力。

B.1.4.5 LogiQA

LogiQA是一个用于评估模型在阅读理解中逻辑推理能力的数据集。LogiQA来自专家撰写的用于测试人类逻辑推理的问题集合，包括8,678个问答实例，涵盖了范畴推理、条件推理、析取推理和联合推理。

B.1.4.6 PIQA

PIQA是一个用于评测模型物理常识推理能力的数据集。该数据集包括超过16,000个多项选择问题，采用准确率指标进行评估。

B.1.4.7 MuTual

MuTual是一个用于评估模型对话推理能力的数据集。该数据集包括8,860个手动注释的对话，来源于中国学生的英语听力考试。

B.1.4.8 CMMU

CMMU是中文多模态多题型理解及推理数据集，从中国教育体系规范指导下的全国小学、初中、高中考试题中抽取并制作了 3603道题目，题型包括单选题、多选题、填空题，并采用多重评测手段避免模型“随机猜对答案”。按照学段来划分，小学题目有250道，初中和高中分别为1697和1656道，其中，小学只包含了数学一门学科，初中和高中包含了七门学科。

B.1.5 其他能力评估数据集

B.1.5.1 HalluDial

一个大规模的对话层级自动幻觉评估基准，旨在评估大语言模型在对话中识别幻觉的能力及其产生幻觉的倾向。为了全面理解大语言模型在面对对话层级幻觉时的表现，HalluDial设计了自发性幻觉和诱导性幻觉两类场景，涵盖了事实性幻觉和忠实性幻觉两类主要类型。HalluDial数据集包含18,357

轮对话，共有146,856条数据样例和相应的幻觉评估结果；评估结果包含幻觉检测、幻觉定位以及佐证检测结果的解释说明。

B.2 行业评估数据集

行业评估数据集包含行业敏感信息，多为闭源或私有，很难从公开渠道获取。中国移动持续推进体系化网络数据集开放，围绕网元智能、运维智能、服务智能三大领域，开放22项2亿规模网络智能精品数据集，包括感知、诊断、预测、决策、通用AI、网络大模型等能力领域，支撑行业网络AI能力研发。

附 录 C
（资料性）
通信行业评估示例

C.1 确定评估对象

选取通信行业大模型作为评估对象。通信行业大模型支持的场景主要分为通用场景和专业场景，通用场景主要包括客服场景、营销场景、知识问答场景等，专业场景主要包括网络规划、网络建设、网络维护、网络优化、网络运营、资源管理等，其中，每个场景包含相应任务项，每个任务项对应相应评估指标。具体场景及任务项介绍参见表C.1。

C.2 确定评估能力项、任务项及评估指标

针对通用或专业场景中的某些场景，选取该场景下的若干任务项，用于对通信行业大模型进行评估。如选取表C.1中客服场景意图识别任务和业务分类任务作为评估能力项与任务项，并选取准确率作为评估指标。

表C.1 通信行业大模型评估指标

场景类型	能力项	任务项	任务项说明	评估指标
通用场景	客服场景	意图识别	通过对用户查询的理解和分析，智能地识别用户的真实意图，为通信行业智能客服提供精确的服务导向。	准确率
		业务分类	自动将用户咨询的内容分类到相应的业务领域中，以便提供快速而准确的业务处理路径。	准确率
	营销场景	营销规则制定	根据市场响应和内部策略，智能制定营销规则。	BLUE、ROUGE
		智能推荐	利用用户行为和偏好数据，向客户推荐最适合其需求的产品或服务。	BLUE、ROUGE
		用户内容定制	分析用户特征和历史互动，提供定制化的内容和服务。	BLUE、ROUGE
	知识问答场景	通信信息检索	自动从海量的通信数据中检索用户所需的具体信息，快速准确地提供查询结果。	准确率
		通信文本摘要生成	从长篇通信文档中提取关键信息，生成准确的文本摘要。	BLUE、ROUGE
		通信文档知识问答	对用户提出的网络运维场景等通信知识文档进行深度理解并提供精准的答案。	BLUE、ROUGE
专业场景	网络规划	覆盖规划	通过分析地理、用户分布，智能地规划无线网络布局，例如，基站位置、天线方位角倾角等。	BLUE、ROUGE
		容量规划	根据用户需求，查询设备利用率、资源利用率等历史数据，为网络容量分析提供依据。	BLUE、ROUGE
		切片自动勘察	自动勘测网络切片需求，分析最佳部署策略，确保网络资源的高效利用。	BLUE、ROUGE

	网络建设	现场入网验收	根据网络建设相关管理规定，针对于现场验收标准问题给出相关回答。	BLUE、ROUGE
		工程方案设计	根据网络建设领域相关企业标准，对设计方案的合理性相关审核问题给出回答。	BLUE、ROUGE
	网络维护	故障隐患识别	根据各专业相关规定，对各类故障隐患定义、分类、阈值等相关问题给出相应回答。	准确率
		故障定界定位	查询故障相关指标数据，为故障定界定位提供依据。	准确率
		故障处理方案设计	根据历史故障运维案例，为故障提供相关的处理方案回答。	BLUE、ROUGE
	网络优化	网络质差识别	根据各专业相关规定，对质差相关信息给出专业解答。	准确率
		网络优化方案设计	根据历史案例及规范规定，给出网络覆盖、容量等质量优化相关的方案解答。	BLUE、ROUGE
	网络运营	投诉预测	在用户投诉到达前，自动分析网络运行数据，预测并处理可能的投诉原因。	准确率
		投诉分析	对用户投诉内容进行深入理解，提取问题核心。	BLUE、ROUGE
	资源管理	哑资源数据采集	对现网的无源设备进行采集、录入并在资管系统中进行增删改查等管理能力。	准确率
		哑资源识别	可实现对一些无源设备、器件等网络资源进行识别。	准确率

基于选取的客服场景能力项，意图识别和业务分类任务项，以及准确率评估指标，确定评估框架如图4所示，其中评估元素A表示通信行业大模型在客服场景下针对意图识别任务的准确率，评估元素B表示通信行业大模型在客服场景下针对业务分类任务的准确率。

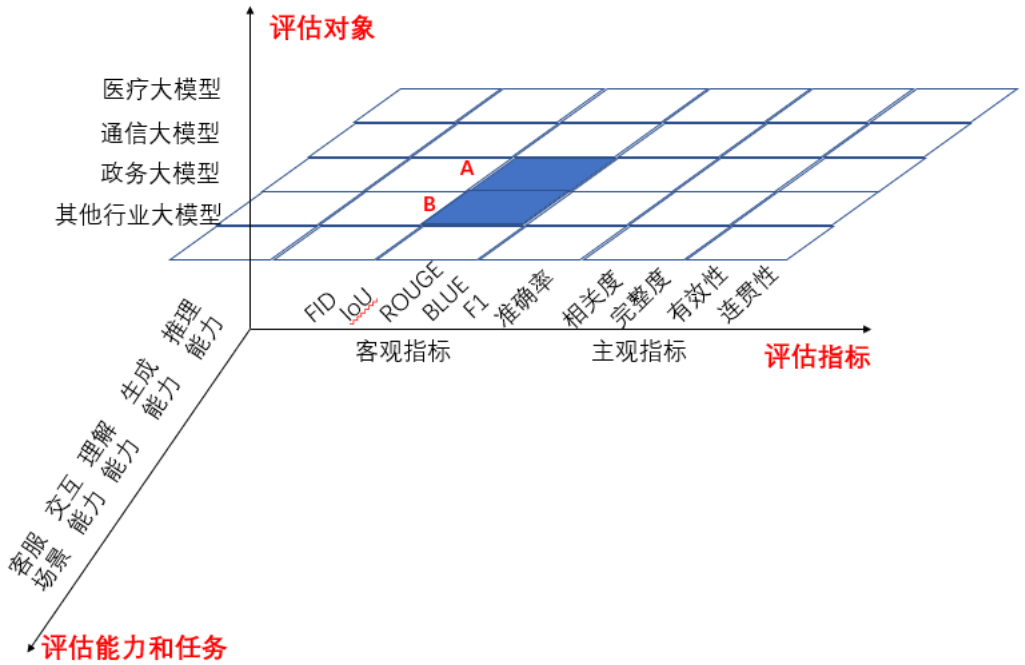

图4 通信行业大模型评估框架

C.3 确定评估数据集和评估方式

为简化说明，本节以C.1.2中确定的客服场景意图识别任务为例。选择意图识别数据集作为评估数据集并基于准确率评估指标采用自动化评估方式进行评估。意图识别数据集包含上千条评估样本，每条评估样本由用户输入问题和候选意图列表组合成的prompt以及期望结果构成。将评估样本的prompt部分输入到模型，并根据模型从候选意图列表中选择的意图识别结果与期望意图识别结果进行比较，相同为1，不同为0，判断模型在该评估数据集上的评估结果。通过对意图识别数据集的所有1352条评估样本进行评估，最终正确完成意图识任务的评估样本为1150条，因此得到通信行业大模型在该意图识别评估数据集上的准确率为85%。同理对客服场景业务分类任务通过任务分类数据集进行评估，得到评估结果为93%。

C.4 输出评估结果

将C.1.3中确定的意图识别任务的评估结果85%通过百分制转化为评估得分85分，将业务分类任务的评估结果93%通过百分制转化为评估得分93分，根据业务需求及专家经验，分别确定任务加权值为0.6和0.4后，通过加权平均方法：

$$加权平均值 = \frac{\sum(x_i \times w_i)}{\sum w_i} \quad \text{……（7）}$$

其中

x_i是第i个任务项得分；

w_i是第i个任务项加权值。

得到通信行业大模型在客服场景下的综合得分为$\frac{（85\times0.6+93\times0.4）}{（0.6+0.4）} = 88.2$，评估结果参见表C-2所示：

表C.2 通信行业大模型客服场景评估结果

能力项	任务项	评估指标与评估结果	评估得分	加权值	综合得分
客服场景	意图识别	准确率（85%）	85分	0.6	88.2
	业务分类	准确率（93%）	93分	0.4	

参 考 文 献

[1] Zhao, Wayne Xin et al. “A Survey of Large Language Models.” *ArXiv* abs/2303.18223 (2023): n. pag.