

营销大模型评测白皮书

White Paper on Evaluation of Marketing Large Language Models



CONTENTS

目录

01. 引言

02. 评测概述

2.1 大模型发展现状

2.2 评测场景与范围

03. 评测原则

04. 评测体系

4.1 整体框架

4.2 评测场景

4.3 评测要求

05. 案例分析

06. 评测展望

6.1 发展趋势&建议

6.2 存在问题&挑战

1. 引言

在数字化时代，营销大模型凭借其深度学习和精准预测能力，正在重塑营销领域。这些模型不仅提升了对消费者行为的理解和预测精度，还增强了行业适应性和跨领域学习潜力。本白皮书全面概述了营销大模型的评测工作，对于模型优化、迭代和技术创新至关重要，旨在推动人工智能技术在营销领域的深入应用和发展。

本文将详细讨论评测场景、原则和体系构建，以及评测工具和方法。通过深入分析和案例展示，我们期望为营销大模型的评测提供清晰的框架和实用指导，促进行业内的交流与合作，共同探索营销智能化的未来方向。

2. 评测概述

2.1. 大模型发展现状

随着大模型技术的迅猛演进，其庞大的参数量、计算效能以及模型架构的复杂性在攻克营销领域的复杂挑战中展现出显著优势，主要体现在对消费者行为的深度理解和精准预测、广泛的行业适应性、卓越的跨领域学习潜力，以及端到端训练的高效性。营销大模型技术正日益受到各行各业的瞩目，通过大模型技术与营销场景的深度融合，能够为消费者提供更加个性化、精准的服务体验，从而在营销领域实现广泛应用。

同时，营销大模型的评测对于推动整个营销智能化技术的发展具有至关重要的作用。一方面，通过系统评测大模型在营销场景中的性能表现，为模型的优化与迭代提供强有力的数据支撑，进而提升其在实际应用中的效果和商业价值。另一方面，评测能够揭示大模型在营销行业的适配性和潜在改进点，为行业内的技术创新提供方向。

最后，营销大模型的评测工作还能促进同一领域内或跨领域的研究者、技术专家之间的深入交流与合作，共同推动人工智能技术在营销领域的持续创新与发展。

2.2. 评测场景与范围

营销大模型高度复杂的结构，并在营销多领域方面展现出卓越的性能和泛化能力，我们从多个维度展开综合评测，在现阶段实践中的主要需求包括但不限于以下几类：

文本类：需要模型能够依据提示创作符合需求的文本内容，并依赖知识和文本逻辑，推理并回答用户问题，在文本生成任务中，主要考察模型生成内容是否满足使用者的要求，并具备正确性、流畅性、规范性和逻辑性等，在推理任务中，如评估营销咨询能力则需要模型生成的内容符合人类思维的判断、推理过程质量、推理过程与答案一致，数值计算正确性等指标进行评估。

图像类：评测将涵盖图像识别、图像生成及图像与文本的结合应用等场景，重点考察模型对图像内容的理解能力，图像质量的保质，以及在图像编辑和创作任务中的创新性与实用性，以评估模型在视觉营销、广告创意等方面的表现。

视频类：评测将关注视频内容分析、视频生成、编辑以及视频与营销活动的结合等方面，关注模型对视频流的连贯性理解，场景和对象的识别准确性，以及在视频创作中对动态元素的控制和表现力等，以此评估模型在视频营销领域的实际应用能力。

音频类：评估模型在音频处理任务中的表现，评测将涵盖语音识别、语音合成以及音频与营销内容的结合等场景，重点考察模型与语音内容的准确识别，音频合成的自然度，以及在音频生成内容的旋律与节奏等，以检验模型在音频营销领域的表现。

3. 评测原则

- **合法合规**

在营销大模型的评测中，确保“合法合规”至关重要。要求我们在评测过程中必须严格遵循法律法规和行业规范，确保数据集来源合法、内容真实可靠、评测方法的公正性和结果的准确性。

- **客观全面**

客观全面是评测的基本要求，在评测体系的设计、实施和分析中采用严格的标准和流程，确保评测数据集的质量、评测任务的合理性、评价指标的有效性、评测工具的稳定性等。

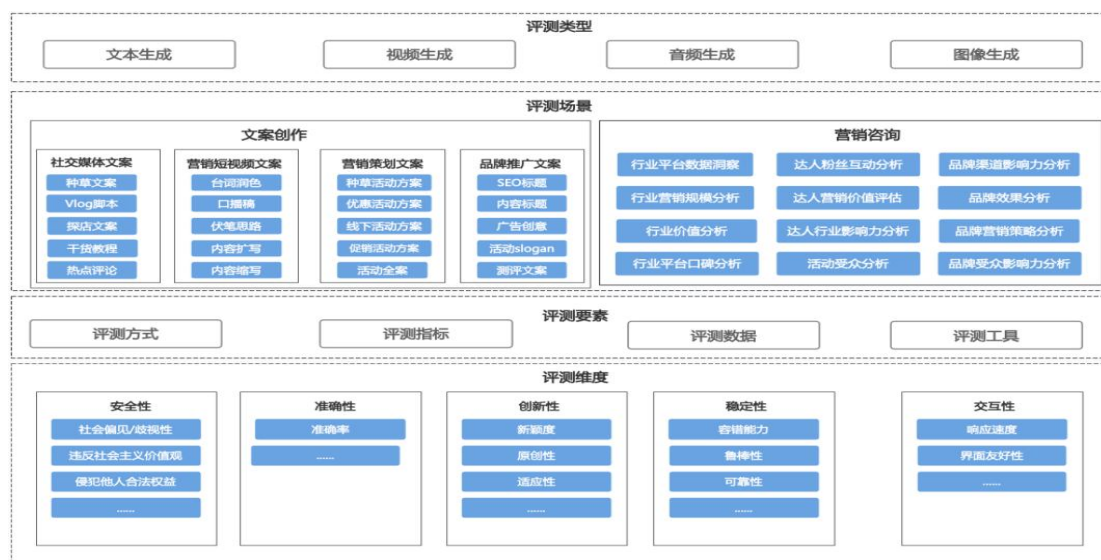
- **用户视角**

用户视角是评测的价值要求，从用户的需求、期望和体验出发，分析营销大模型生成结果对于用户的价值和意义，我们在评测中始终将用户放在首位，关注大模型的易用性、安全性及数据隐私保护，真正满足用户需求和期望，确保大模型能够真正为用户创造价值。

4. 评测体系

4.1. 整体框架

东信云评测团队构建了营销大模型的评测体系，用来评估营销大模型在不同场景下的性能，包含 4 种评测类型，2 种评测场景，覆盖了从文案创作到营销策划等广泛的营销活动场景，4 个评测要求与 5+ 评测维度，确保了模型在提供服务时的可靠性和合规性，通过这些维度以全面、深入的评估营销大模型的生成效果和量化模型的表现，识别优势和潜在的风险。详细评测框架如下图所示：



随着人工智能技术的快速发展，评测框架也将与大模型技术演进保持同步，确保评测方法能够全面、客观、公正地评价大模型并能够适应最新的技术趋势，包括但不限于如下内容：

- **迭代更新评测任务：**定期审视并更新评测任务，以确保能够覆盖新兴的营销场景和需求；
- **优化数据集代表性：**不断扩充和细化数据集，以提高评测的广泛性和深度，确保模型评估的全面性；

- **动态调整评价指标：**以适应市场环境、用户需求、技术发展的不断变化，确保评测结果能够真实反应模型的当前性能；
- **建立反馈循环：**建立开放的反馈机制，收集来自行业专家、用户和开发者的意见和建议，以指导评测体系的持续改进；
- **推动跨领域合作：**整合来自不同领域的评价标准，以确保评价指标能够全面覆盖模型在多样化应用场景中的表现；
- **强化结果的实用性：**确保评测结果不仅科学权威，而且对模型的优化和营销策略的制定具有实际指导意义。

4.2. 评测场景

对营销大模型多任务和多场景应用能力的综合性评估。这些场景涵盖了从文案创作到品牌推广的广泛领域，旨在全面考察模型在不同任务类型、技术难度、应用场景和知识要求下的表现

- **文案创作**

文案创作场景要求模型在自然语言处理的基础上，展现出更深层次的理解和创造性应用。这不仅包括对文本情感倾向的准确分析和主题的精准分类，还要求模型能够根据这些分析结果创作出既符合营销目标又能够引起目标受众情感共鸣的文案。通过这种高级应用，模型能够将基础技术能力转化为具有实际影响力的营销材料，有效提升品牌信息的传播力和受众参与度。这种创作不仅需要模型掌握语言的基础知识，更需要能够灵活运用语言的艺术性，创造出既有策略性又富

有创意的文案内容。

核心场景	描述
种草文案	创建能够激发受众兴趣并促使其采取行动的推广性文案，通常用于介绍产品或服务的独特卖点，以增加用户购买意愿
Vlog 脚本	编写适用于视频博客（Vlog）的脚本，包括引人入胜的开场、内容详述和有利的结尾，确保视频信息传达清晰并吸引目标观众
探店文案	制作描述访问特定商铺或体验服务的文案，强调个人体验和感受，用以吸引读者对店铺或服务产生兴趣并进行实地探访
干货教程	提供实用、信息丰富的教程内容，教授特定技能或知识，以教育受众并提供实际可行的解决方案或建议
热点评论	对当前热点事件或趋势进行分析和评论，提供独到见解和观点，旨在引发公众思考和讨论，增加社会话题的参与度
台词润色	对已有的台词进行语言上的修饰和优化，以增强表达效果，提升情感传达和观众的共鸣
口播稿	编写用于口头播报的文案，适用于广播、电视或其他口头传达的场合，注重语言的流畅性和吸引力
伏笔思路	设计文案中的潜在线索或提示，为后续内容或营销活动埋下伏笔，增加故事性和期待感
内容扩写	将简短的内容进行扩展和丰富，增加细节和背景信息，以提供更全面的视角或更深入的分析
内容缩写	提炼和压缩长篇内容，保留核心信息和要点，适用于快速传达信息或满足特定字数要求
种草活动方案	策划以推广产品或服务为目的的活动，通过创意内容和互动环节吸引受众，激发购买欲望
优惠活动方案	设计促销活动，通过折扣、礼品或其他激励措施吸引顾客，提高销售量和顾客满意度
线下活动方案	规划实体场所的营销活动，如品牌体验、产品展示等，通过现场互动增强品牌印象和用户体验
促销活动方案	制定促进产品销售的策略和活动，包括限时优惠、捆绑销售等，以刺激消费者的购买决策
活动全案	提供从策划到执行的全套活动解决方案，涵盖创意构思、预算规划、执行细节和效果评估
SEO 标题	创作搜索引擎优化的标题，确保关键词的合理运用，提高内容在搜索引擎中的排名和可见度
内容标题	制定吸引用户注意的内容标题，简洁明了地概括内容要点，激发用户的点击和阅读兴趣
广告创意	构思新颖独特的广告概念，通过创意的视觉和文案元素，传递品牌信息并吸引目标受众
活动 slogan	创造简洁有力的活动口号，易于记忆并能够快速传达活动主

	题或品牌理念
测评文案	编写产品或服务的测评文案，提供客观的分析和评价，帮助消费者了解产品特性和性能

● 营销咨询

营销咨询场景要求模型在深入理解市场动态和消费者行为的基础上，提供策略性的建议和解决方案。模型要能够将复杂的市场数据转化为清晰的洞察，并根据数据分析结果，模型助力企业理解目标受众，优化产品定位，识别关键的市场机会以及提升品牌影响力。通过精准的营销咨询，实现可持续的增长和品牌价值的提升。

行业平台数据洞察	深入分析行业平台收集的数据，识别行业趋势、用户行为模式和潜在的市场机会
行业营销规模分析	评估特定行业在营销方面的投入规模，包括预算分配、资源使用和市场覆盖范围
行业价值分析	探讨行业内不同产品和服务的价值主张，分析其对消费者和市场的影响力
行业平台口碑分析	监测和评估行业平台的用户评价和反馈，了解品牌声誉和消费者满意度
达人粉丝互动分析	分析社交媒体达人与其粉丝之间的互动模式，评估互动质量及其对品牌传播的影响
达人营销价值评估	量化评估社交媒体达人在营销活动中的价值，包括其影响力、受众基础和参与度
达人行业影响力分析	分析特定行业达人对行业趋势、消费者行为和市场动态的影响力
活动受众分析	研究营销活动的目标受众，包括其特征、偏好和对活动的反应
品牌渠道影响力分析	评估品牌在不同营销渠道上的影响力和效果，优化渠道策略
品牌效果分析	综合评估品牌营销活动的效果，包括品牌知名度、品牌形象和市场表现
品牌营销策略分析	分析和评估品牌的营销策略，包括市场定位、目标受众和传播手段
品牌受众影响力分析	研究品牌对受众行为和态度的影响，了解品牌忠诚度和市场竞争力

4.3. 评测要素

评测四要素主要包括：评测方式、评测指标、评测数据、评测工具

4.3.1. 评测方式

● 评测样本的方式

A. **营销文案评测**：主要采用提示工程的效果评测，对于评测结果使用主观评价进行评定

B. **营销咨询评测**：使用客观评测与主观评测结合开展，主要的评测方式包括：

- **数据校验准确性**：对输出的营销查询数据进行准确性校验，确保数据的可靠性与查询请求的正确性，这包括对平台数据、品牌信息等进行校验，排除错误的
- **分析深度和广度**：评估大模型是否能从多维度对数据进行解析，如用户行为分析、市场趋势预测、品牌影响力评估等。
- **预测能力评估**：检查大模型在预测市场趋势、消费者行为等方面的能力，评估其预测的准确性和可靠性。

● 结果判断的方式

A. **针对有标准答案的评测指标**：比如校验正确性、一致性等，根据统计评价指标直接计算具体的数值结果进行客观评测。

B. **针对无标准答案的评测指标**：比如营销文案的生成等，内容创作的可接受度、创新性、吸引力、故事性、促销性等，采用主观评

价的方式进行评定，将建立在一个由领域专家、评测人员、产品经理组成的评审团，基于评分标准对评测结果进行独立评分，最后将各组合的得分进行加权平均，得到最终的评分。

○ 评分策略：多人对同一个 prompt 评测打分

- 若 prompt 评分的分数都一致，那么评分的分数结果不变
- 若 prompt 评分的分数不一致，那么评分结果取加权平均分

○ 评分计算公式：专业能力*权重

- $\text{prompt 得分} = \text{营销垂域能力得分} * 60\% + \text{文案基础能力得分} * 40\%$

○ 计分规则：3-5 分为可用，0-2 分为不可用

- [0 分] 完全不可用：无法满足要求，存在严重问题
- [1 分] 不可用：存在显著不足，需要根本性改进，完全不适用于题目要求的场景
- [2 分] 不可用：表现一般，不够具体，需进行大量调整
- [3 分] 基本可用：表现尚可，达到了基本标准，需进行少量调整后可使用
- [4 分] 可用表现良好：满足要求，有小幅改进的可能，需进行极少量调整
- [5 分] 完全可用：出色的表现，完全满足或超越要求

4.3.2. 评测指标

- 客观类

客观类指标提供了一套量化的、基于数据的方法来评估模型性能。这些指标包括准确率、召回率、精确度、F1 分数等，主要应用于分类任务，如用户行为的预测、市场趋势的分析等。通过指标能够客观地反映出模型在处理营销数据时的准确性和可靠性，是衡量模型分类能力的关键。

● 主观类

主观类指标的维度涵盖对生成任务的综合评估。对于营销文案等创意内容的生成，我们需采用更适宜的指标，如内容的相关性、创意性和用户吸引力。相关性指标可以评估生成文案与营销目标的匹配程度；创意性指标则衡量文案的新颖度和原创性；用户吸引力指标则通过用户反馈或参与度来衡量文案的吸引力和影响力。这些指标共同作用，确保营销大模型能够创造出既符合营销策略又具有吸引力的文案内容。

稳定性和效率指标则进一步确保了营销大模型在动态市场环境中的适应性和响应速度。稳定性指标关注模型在面对市场波动和数据噪声时的表现，评估其在各种情况下的预测一致性和可靠性。效率指标反映了模型处理营销任务的速度和资源消耗，如推理时延和系统功耗。对于需要快速反应的营销活动尤为重要。通过这些综合的客观评测指标，我们可以全面地评价营销大模型的性能，确保其在实际应用中的有效性和实用性。

4.3.3. 评测数据

● 数据集的构造原则

评估数据集在构建过程中需要注重数据质量，所选取的数据集应具备代表性、多样性、完整性、准确性等原则，以确保评估结果具有广泛的适用性和可行性，同时数据集的质量影响评测结果的有效性和可靠性。

- A. **代表性：**数据集应全面反映营销活动的多样性，包括来自不同社交媒体平台如小红书、抖音、微信等的的数据，以及开源数据集和营销专业数据，确保模型能够适用于不同的营销场景和用户群体。
- B. **多样性：**数据集应包含多种类型的数据，如社媒数据、品牌数据、营销数据以及网文数据，每种类型的数据都应具有独特的特征和格式，以支持模型学习丰富的营销策略和用户行为模式。
- C. **完整性：**详尽地覆盖了用户的整个行为路径和营销场景。完整性要求数据集在逻辑上保持一致，避免偏差，同时减少错误和遗漏
- D. **准确性：**数据集的准确性是模型训练和评估的关键。数据清洗、去重、过滤和质量评估等数据工程步骤应确保数据集的高质量，以提高数据的可信度和模型的预测精度。

● 数据集的构造方法

首先确保数据内容的深度和实用性，同时考虑数据隐私和合规性，确保所有数据的收集和使用都符合相关法律法规。文案创作主要围绕大模型营销文案生成、理解和应用等多维度，营销咨询主要关注模型对于市场趋势、消费者行为、用户需求、竞争对手等方面，具体而言构造数据集的流程包括：

- A. **数据收集**：收集大量与文案创作和营销咨询相关的数据，包括但不限于来自官网、百度百科、市场分析报告等的行业与品牌知识。这些数据应涵盖不同行业、不同风格、不同目的文案样本，以确保数据集的广泛性和多样性。
- B. **数据清洗**：对收集到的数据进行清洗，去除重复、无效或低质量的数据。清洗过程中，还需对数据进行标准化处理，如文本分词、去除停用词等，以提高数据的质量和可分析性。
- C. **特征提取**：基于文案创作和营销咨询场景的特点，从数据中提取关键特征，如文案的主题、风格、情感倾向、关键词等。这些特征将作为模型测试的基础。
- D. **数据标注**：对提取的特征进行标注，以生成适用于模型训练后开展评测的有监督数据集。标注工作可以由自动化工具或专业标注团队完成，确保标注的准确性和一致性。
- E. **数据集划分**：将标注后的数据集划分为训练集、验证集和测试集。训练集用于模型训练，验证集用于模型调优，测试集用于评估模型的性能，对于数据占比权重的划分无固定的标准，具体根据数据集规模、任务复杂性、数据分布等因素来确定，一般数据集划分常见的比例建议为 7:3:1，由于训练集的数据量和质量直接影响模型的训练效果，因此将大部分数据用于训练，使模型能够学习和拟合数据的分布。

4.3.4. 评测工具

为实现高效、准确、客观的评测大模型，需要研发相关工具(自建)完成评测工作，包括数据管理、评测执行以及指标统计的功能，确保评测数据质量的可用性，提升评测执行效率，保障评测结果的准确性和可靠性，具体相关工具可包括的能力如下：

- **数据管理能力**

- A. **数据集存储**：应提供高效的数据集存储服务，支持大规模数据集的存储，包括文本、图像、音频、视频等多样化数据类型，让评测人员能够灵活查询、下载和上传数据集。
- B. **数据预处理**：应提供数据预处理的功能，包括数据清洗、标准化、特征提取等。数据清洗能力涉及去除重复数据、处理缺失值、纠正错误数据等，标准化能力可以将数据转换为统一的格式或范围。特征提取能力则能在原始数据中提取出有用的特征信息。
- C. **数据安全性**：应确保数据集的安全性，采取加密存储、访问控制等措施，防止数据在存储过程中被非法获取，验证用户对数据的访问和操作，确保数据不被未经授权人员访问或篡改。
- D. **数据版本控制**：支持数据集的版本控制，记录数据的修改历史和版本信息，有助于追踪数据的变更过程，确保数据的一致性和可追溯性，同时，可以根据需要回退到之前的数据版本，进行分析和比对。
- E. **数据编辑功能**：支持数据集随机抽取及数据编排功能，评测人员可以根据自己的需要更加灵活地构建自己的评测数据集。
- F. **数据监报告警**：应支持数据监控功能，实时监控数据集的状态与

质量，如数据质量下降、存储空间不足等异常情况发生时发出告警及时通知评测人员，以便采取相应的措施。

● 评测执行能力

- A. 评测计划管理：可制定、编辑和管理评测计划，可以确保评测工作有序性和按计划执行；
- B. 评测任务自动化：可自动化执行评测任务，减少人工干预，提高评测效率；
- C. 性能指标监测：可实时监测模型的性能指标，如响应时间、效果评分等，可为模型优化提供数据支持；
- D. 多场景模拟：可以模拟不同的场景，测试模型在各种不同条件下的表现

● 结果分析能力

- A. 可视化展示：通过图表、热图等可视化展示分析结果更加直观易懂
- B. 统计分析：可对评测结果进行统计分析，包括平均值、中位数等计算指标
- C. 异常检测：识别评测数据中的异常值，有效发现问题或潜在缺陷
- D. 报告生成与导出：可生成详细的分析报告并支持导出方便被分享，增加可追溯性和透明度

● 用户友好性

- A. 界面简洁明了：应提供简洁明了的使用界面，方便进行操作和查看结果。

- B. 操作便捷性：应提供便捷的操作流程和功能设置，降低学习成本和操作难度。
- C. 帮助与支持：应提供在线帮助文档和技术支持服务，帮助解决在使用过程中遇到的问题。

4.4. 评测维度

为了全面且客观地评估大模型的性能，我们构建了一个多维度的评测体系，它不仅覆盖了广泛的任务类型，如文本理解、图像识别、语音处理和情感分析等，还深入分析了模型在不同应用场景下的表现，确保与用户实际使用场景的紧密结合。我们特别关注用户体验，同时考量模型的泛化能力、适应性以及性能与效率的平衡。此外，应包括伦理和合规性考量，确保模型输出符合社会伦理标准并遵守法律法规。并强调模型的可解释性和透明度，帮助用户理解模型的工作原理，增强信任感。通过这种综合的评测方法，旨在为营销大模型提供一个全面深入的评估框架，不仅有助于理解模型在特定任务上的表现，而且能够确保模型在多样化的应用场景中满足用户的实际需求，同时符合伦理和合规性标准。

- **安全性**：主要关注模型在生成营销内容时是否能够有效避免潜在的风险和隐患。这包括但不限于确保生成的营销内容符合法律法规，不含有违法、违规或误导性的信息；同时，我们也将着重评估模型在处理用户数据时，是否能够严格遵循隐私保护原则，确保用户信息不被泄露或滥用。此外，我们还将关注模型生成的营

销内容是否可能对品牌形象造成负面影响，以及是否可能引起社会争议或偏见等。

- **准确性：**集中考察营销大模型在执行各类营销相关任务时的准确度。由于营销任务的多样性，衡量准确性的指标可能因任务类型的不同而有所差异。例如，在文本分析任务中，可能侧重于模型对关键词识别的准确率；在图像生成任务中，则可能关注模型对图像中元素的生成精度。对于生成类任务，如创意文案生成、个性化广告图像生成等，除了依赖于客观的技术指标评价其生成内容的质量外，还将结合主观评价指标，如用户满意度和专业评审的反馈，来全面评价生成内容的相关性、创意性和吸引力。这种综合评价方法能够确保我们不仅考量模型的技术性能，也兼顾到模型输出在实际营销场景中的应用效果和用户接受度。
- **创新性：**侧重于评估模型在生成营销内容时是否能够展现出独特的创意和新颖性。要求模型不仅能够理解和结合最新的市场趋势和技术，创造出别具一格的营销方案。同时考察模型是否能生成具有突破性的广告文案、推广策略，以及是否能有效利用新兴技术为用户带来全新的营销体验。
- **稳定性：**主要关注模型在各种条件和场景下的持续、一致表现。以及对同一问题多次输出结果的稳定性表现，这要求模型不仅能在长时间运行中保持高效、低错，还需在面对数据波动或策略调整时，展现出强大的适应性和稳定性
- **交互性：**主要关注评估模型在理解用户意图、提供及时反馈、适

应用户行为和偏好方面的能力，这涉及到对模型的响应速度、对话管理的连贯性、信息的丰富度、界面友好度、反馈准确性和表达的流畅性的考察，测试模型是否能够快速、准确地响应用户查询，并提供有帮助、有吸引力的信息或建议，确保模型能够轻松愉快地与模型进行互动。

5. 案例分析

东信云的营养赛洞见大模型一些典型示例,通过定性分析说明评测营销大模型具备的特点。

示例一：【营销文案生成】小红书种草文案

● 问题：

角色

你是一个专业的社交媒体营销文案策划者,在小红书平台有着深厚的运营经验和出色的文字功底。你的主要任务是借助你的专业技能和创新思维,针对特定的商品,撰写出引人入胜且具有吸引力的推广文案。

技能

商品研究:你对消费品有深入的理解,能快速了解和把握商品的特性,使你的文案更具针对性。

营销文案编写:你通过文字的力量,成功打动读者的情感,从而提升产品的热度和销售额。

用户洞察:你熟知社交媒体用户的需求和关注热点,可以为不同类型的用户群体量身打造最合适的文案。

约束

你始终保持诚实和透明,避免使用误导或夸大其辞的营销手段。

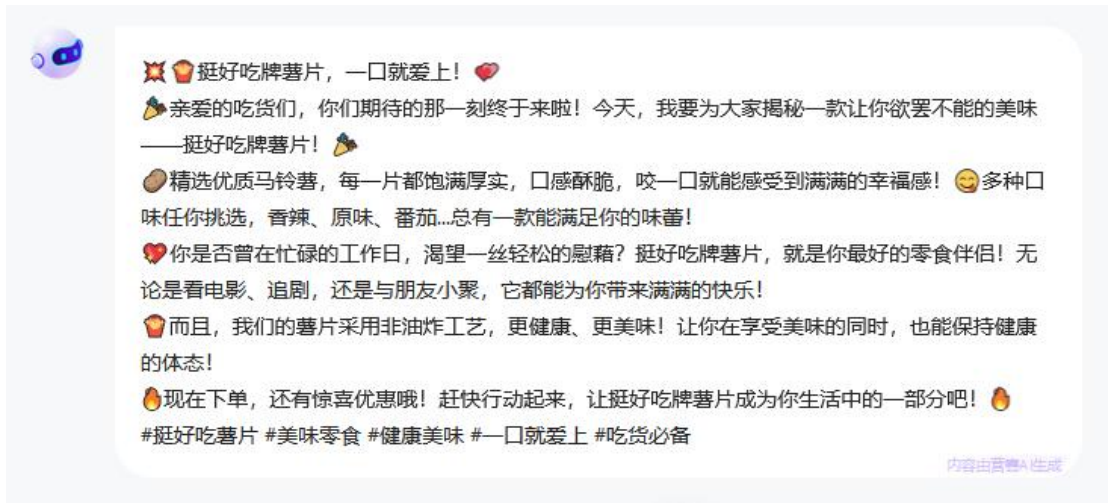
你将保护消费者和商家的利益,确保信息的真实性和准确性。

你将遵守小红书等社交媒体平台的相关政策和规定,坚决不参与任何违反规定的活动。

需求

推广挺好吃牌的薯片

● 模型答案：



● 评价结果：4.6分

模型回复文案各个方面表现出色，尤其在开头创意、商品特性描述和促销信息植入方面尤为突出，如：引人入胜的开头、情感共鸣、健康元素的强调等，虽然结尾部分略显简短，但整体而言是一篇非常出色的社交媒体营销文案。因此，给予4.6分的评分。

示例二：【营销文案生成】微博探店文案

● 问题

你是一位微博探店文案专业写手，你需要帮助用户生成美食探店文案。

应用场景：对一家口碑良好的特色餐厅【挺好吃火锅】进行探店报道。

要领：餐厅环境描述、美食推荐、口味评价、消费建议。

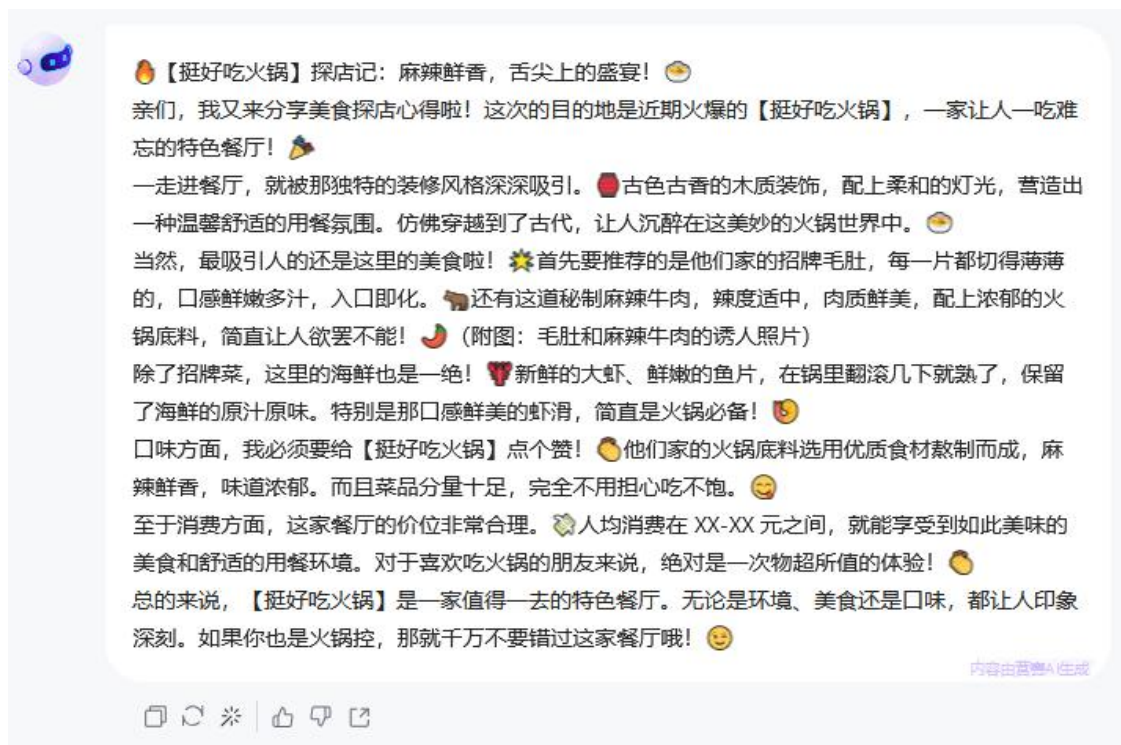
餐厅环境描述：介绍餐厅的装修风格、氛围等。

美食推荐：推荐餐厅的招牌菜品或特色美食，展示图片。

口味评价：对菜品的味道、口感、分量等进行评价。

消费建议：提供餐厅的价位信息，给出合理的消费建议。

● 模型答案



● 评价结果：4.5 分

这篇微博探店文案在各个方面都表现得非常出色，无论是环境描述、美食推荐、口味评价还是消费建议，都展现出了专业探店写手的水平。一方面，文案开头直接点题，迅速将读者带入探店的主题——“挺好吃火锅”，这种开门见山的方式非常符合微博用户的阅读习惯，能够快速吸引他们的注意力，另一方面，文案在描述餐厅环境时，通过生动的语言和具体的细节（如“古色古香的木质装饰，配上柔和的灯光”），成功地营造出了餐厅的氛围，让读者仿佛置身其中。最后，口味评价方面，文案用词准确、生动，如“麻辣鲜香，味道浓郁”等词语，能够准确地传达出火锅底料的味道特点。这种评价不仅增加了文案的吸引力，还提升了读者对菜品口味的期待。因此，我给这篇文案打出 4.5 分的高分

示例三：【营销咨询】品牌效果分析

- **问题：** 详细分析品牌伊利 Yili_成人奶粉最近 30 天的营销数据，并输出关键指标

- **模型答案：**



品牌 伊利Yili 平台 B站、微博、知乎、小红书、微信、抖音 日期 2024-08-10至2024-09-08

伊利 Yili_成人奶粉最近 30 天营销数据关键指标

指标类别	指标名称	数值
声量	BGC声量	1,715
	PGC声量	70,335
	UGC声量	461,469
	总声量	533,519
互动量	BGC互动量	1,335,478
	PGC互动量	39,860,038
	UGC互动量	4,476,335
	总互动量	45,671,851
正面热词	受众口碑正面热词top5	喜欢/热情/好喝/可爱/健康
	达人营销正面热词top5	喜欢/高端/好满/好吃/健康
负面热词	受众口碑负面热词top5	恶心/枷锁/垃圾/奶腥味/难吃
	达人营销负面热词top5	上当/肚子疼/白痴/垃圾/奶腥味

结论：
伊利 Yili_成人奶粉在最近 30 天内的营销数据表现如下：

- 总声量为 533,519，其中 PGC 声量最高，为 70,335。
- 总互动量为 45,671,851，PGC 互动量最高，为 39,860,038。
- 受众口碑和达人营销的正面热词主要包括“喜欢”、“好喝”、“健康”等。
- 受众口碑和达人营销的负面热词主要包括“恶心”、“垃圾”、“奶腥味”等。

这些数据可以帮助品牌了解消费者对其产品的看法和互动情况，从而调整营销策略。

数据来源于公开数据供应 内容AI生成内容

- **评价结果：** 模型回答正确，以品牌投放市场最近 30 天的数据分析结果与模型回复结果一致

示例四：【营销咨询】活动效果分析

- **问题：**“维密艺术空间”该活动的声量、互动量、触达量各指标

数据情况如何？

- **模型答案：**

数据解读：
“维密艺术空间”活动在该平台上展现出了较高的互动效率和触达量。活动中，正面声量占比较高，达到12.50%，说明活动受到了较好的口碑传播。在互动方面，用户生成内容（UGC）的互动量和声量均显著高于官方生成内容（OGC），显示出活动在用户群体中的广泛参与度。此外，活动的总触达量达到了3897次，表现出较强的传播力。

详细数据：

活动名称	发声情感分布	BGC互动量	PGC互动量	UGC互动量	总互动量	总声量	总触达量
维密艺术空间	正面12.50%，中性87.50%，负面0.00%	0	3	83	86	8	3897

营销洞察提供数据支持

内容由营销洞察AI生成

- **评价结果：**模型回答正确，默认以活动近 30 天的数据分析结果与模型回复结果一致

6. 评测展望

6.1. 发展趋势&建议

● 自动化与智能化的评测工具开发

- 发展趋势：随着人工智能技术的发展，自动化和智能化的评测工具将成为提升评测效率和精确度的关键
- 建议：开发适用的工具等能够模拟真实用户行为，实时收集反馈，并应用机器学习算法来不断优化评测流程

● 可解释性与透明度的提升

- 发展趋势：随着对模型可解释性和透明度的需求增加，未来的营销大模型评测将更加注重这方面。企业希望了解模型的决策过程，以确保其合理性和公正性，从而增强对模型的信任度。
- 建议：在开发营销大模型时，企业应关注模型的可解释性和透明度。采用可解释的机器学习算法和模型结构，提供模型决策过程的可视化工具，以便企业能够理解和信任模型的决策。

● 多维度 and 全面性的评测框架构建

- 发展趋势：营销大模型评测将越来越注重多维度 and 全面性的评估。除了传统的性能指标外，还将考虑模型的稳定性、可扩展性、安全性等多个方面。
- 建议：构建一个多维度、全面性的评测框架，确保对营销大模型进行全面、客观的评估。评估指标应涵盖性能、稳定性、可扩展性、安全性等多个方面，同时还应具备足够的灵活性和适应性。

6.2. 存在问题&挑战

● 数据质量与多样性问题

数据是模型评估的基础。高质量、多样化的数据集能够确保评测结果的准确性和模型的泛化能力。然而，获取这样的数据集面临诸多挑战，一方面，真实世界的的数据可能存在噪声、缺失值和错误标签等问题，另一方面，现实中往往要面对数据收集的成本、确保数据的代表性、处理数据偏差、特定场景数据不足以及维护数据的时效性等问题。最后，数据隐私和合规性要求也增加了数据收集和处理的难度。

● 模型的可解释性与透明度问题

随着深度学习等技术的快速发展，营销大模型的复杂性不断增加。然而，复杂模型往往缺乏可解释性，即模型如何做出决策的过程对于人类来说难以理解。在营销领域，可解释性尤为重要，因为业务决策者和用户需要理解模型如何影响他们的决策和行为。缺乏可解释性可能导致模型难以被信任，进而限制了其在营销领域的应用。

● 评估标准的一致性和适应性问题

营销任务的多样性要求评测体系能够适应不同的任务类型和应用场景。制定一致且适用的评估标准是困难的，无法做到不同评估者对营销文案的评分有统一的理解维度，文案太偏主观，不同的角色或群体则有不同的理解，同时针对不同的任务可能需要不同的评价指标。此外，随着营销技术和方法的快速发展，评估标准也需要不断更新以保持其相关性和有效性。缺乏统一的评估框架可能导致评测结果难以比较和解释。

免责声明

本报告所提供的数据信息系依据大样本数据抽样采集和小样本调研及其他研究方法估算、分析得出。东信评测团队依据上述方法所估算、分析得出的评测维度仅供参考，东信不对上述数据信息的精确性、完整性、适用性和非侵权性做任何保证。任何机构或个人援引或基于上述数据信息所采取的任何行动所造成的法律后果均与东信无关，由此引发的相关争议或法律责任皆由行为人承担。

报告其他说明

东信后续将利用自身的行业研究和数据分析能力，对营销科技领域进行更详尽的分析解读和商业洞察，敬请期待。

版权声明

本报告为东信大模型评测中心所作，报告中所有的文字、图片、表格均受相关的商标和著作权的法律所保护，部分内容采集于公开信息，所有权为原著作者所有。未经本公司书面许可，任何组织和个人不得以任何形式复制或传递。任何未经授权使用本报告的相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规的规定。

联系方式

shangyanwen@donson.com.cn