E-Commerce Letters 电子商务评论, 2026, 15(7), 939-948
Published Online July 2026 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2026.157837

Hans汉斯

# 面向电子商务知识服务的大语言模型评价指标研究

**邓星志**

贵州大学省部共建公共大数据国家重点实验室，贵州 贵阳

收稿日期：2026年6月5日；录用日期：2026年6月18日；发布日期：2026年7月29日

## 摘　要

**大语言模型在电商客服场景中展现出强大的语义理解与文本生成能力，为自动化客服提供了新的技术路径。然而，现有研究主要依赖文本相似度指标评估模型性能，这类指标难以有效反映电商客服回答中的业务规则正确性与知识完整性。针对上述问题，本文提出两种面向电商知识问答任务的专业评价指标：知识要素准确率与知识要素覆盖率。知识要素准确率用于衡量模型回答中知识要素的正确程度，知识要素覆盖率用于评估模型回答对关键业务知识的覆盖情况。通过对多个主流大语言模型在电商客服问答任务上的实验分析，结果表明与传统文本评价指标相比，本文提出的指标能够更加准确地反映模型在电商知识理解与业务规则表达方面的能力，为电商平台智能客服系统的评估与优化提供科学依据。**

## 关键词

**电子商务，评价指标，大语言模型，知识要素，智能客服**

# Research on Evaluation Metrics of Large Language Models for E-Commerce Knowledge Services

**Xingzhi Deng**

State Key Laboratory of Public Big Data, Guizhou University, Guiyang Guizhou

Received: June 5, 2026; accepted: June 18, 2026; published: July 29, 2026

## Abstract

**Large language models have demonstrated strong semantic understanding and text generation**

文章引用: 邓星志. 面向电子商务知识服务的大语言模型评价指标研究[J]. 电子商务评论, 2026, 15(7): 939-948.
DOI: 10.12677/ecl.2026.157837

**capabilities in e-commerce customer service scenarios, offering a new technical path for automated customer service. However, existing research mainly relies on text similarity metrics such as BLEU and ROUGE to evaluate model performance, which are inadequate in effectively reflecting the correctness of business rules and the completeness of knowledge in e-commerce customer service responses. To address these issues, this paper proposes two specialized evaluation metrics for e-commerce knowledge question-answering tasks: Knowledge Component Accuracy (KCA) and Knowledge Component Recall (KCR). KCA measures the accuracy of knowledge components in model responses, while KCR assesses the coverage of key business knowledge in model responses. Through experimental analysis of multiple mainstream large language models on e-commerce customer service question-answering tasks, the results show that the metrics proposed in this paper can more accurately reflect the models' capabilities in e-commerce knowledge understanding and business rule expression compared to traditional text evaluation metrics, providing a scientific basis for the evaluation and optimization of intelligent customer service systems in e-commerce platforms.**

## Keywords

**E-Commerce, Evaluation Indicators, Large Language Model, Knowledge Elements, Intelligent Customer Service**



## 1. 引言

随着互联网技术的快速发展，电子商务已成为现代商业体系的重要组成部分。根据中国互联网络信息中心(CNNIC)发布的统计数据，截至 2024 年，中国网络购物用户规模已突破 9 亿，电商平台日均处理咨询量达到数千万级别[1]。随着用户规模的持续增长，平台所需处理的客户咨询问题呈现快速增长趋势，涵盖订单查询、物流信息、售后服务、商品咨询等多种类型。在传统客服模式下，大量用户咨询主要依赖人工客服处理，这种方式不仅运营成本较高，而且在业务高峰期容易出现响应延迟、服务质量波动等问题，难以满足用户对即时性服务的期望。为提升服务效率并降低运营成本，越来越多的电商平台开始引入智能客服系统。智能客服通过自然语言处理技术实现自动化问答，可在一定程度上替代人工客服完成常见问题解答。近年来，大语言模型在自然语言处理领域取得显著进展，如 GPT 系列[2]、LLaMA 系列[3]、Qwen 系列[4]等模型在文本生成、对话理解等任务中表现出较高能力。这些模型具备较强的语义理解能力，能够根据用户输入生成较为自然、连贯的回答，因此逐渐被应用于电商客服场景中。

尽管大语言模型在生成能力方面表现优异，但如何准确评估其在电商客服场景中的回答质量仍然是一个关键问题。目前大多数研究仍采用 BLEU [5]、ROUGE [6]等传统文本生成评价指标。这类指标主要通过计算生成文本与参考答案之间的词汇或 n-gram 重叠程度来评估文本质量。然而，在电商客服问答任务中，用户问题往往涉及具体的业务规则，例如退货期限、退款条件或配送政策等。如果模型回答在业务规则上出现错误，即使文本表达与参考答案相似，仍会对用户造成误导，甚至引发客诉纠纷。

此外，在实际客服问答中，正确回答往往包含多个关键知识点，例如时间限制、适用条件以及操作步骤等。如果模型回答缺少部分关键知识点，用户可能无法获得完整信息，导致操作失败或体验下降。传统评价指标难以识别这类知识缺失问题，因此无法反映模型回答的业务质量。

针对上述问题，本研究提出了知识要素准确率(Knowledge Component Accuracy, KCA)与知识要素召

回率(Knowledge Component Recall, KCR)两个评估指标对模型性能进行量化评价。通过对客服知识进行结构化拆分，来衡量模型回答中的知识准确性与信息完整性。本文的主要贡献包括：

(1) 系统分析传统文本评价指标在电商知识问答场景中的局限性，指出其难以有效识别业务规则错误、知识缺失及语义表达差异等问题；

(2) 提出知识要素准确率与知识要素覆盖率两种新的评价指标，建立面向电商客服场景的专业评估体系。

## 2. 相关工作

### 2.1. 大语言模型

大语言模型是近年来自然语言处理领域的重要研究方向。随着计算能力的提升与训练数据规模的扩大，大语言模型在多个任务中取得了显著进展。OpenAI 提出的 GPT 系列模型是大语言模型的代表之一。GPT 模型基于 Transformer 架构[7]，通过自回归方式进行文本生成，在对话系统、文本生成和代码生成等任务中表现出较高性能。随后，多种开源大语言模型陆续出现，例如 Meta 发布的 LLaMA 系列模型，在较小参数规模下仍能够保持较好的性能表现，为学术界和工业界提供了重要的研究基础。在国内研究方面，智谱 AI 发布的 GLM 模型[8]在中文语言理解任务中取得了良好效果。此外，DeepSeek 等新兴模型也在多语言问答和知识推理任务中展现出较强能力[9]。

尽管这些模型在通用对话任务中表现出较高水平，但在特定行业场景中仍存在知识准确性不足的问题。研究表明，大语言模型在垂直领域应用中容易产生“幻觉”现象，即生成看似合理但事实错误的内容。因此，需要结合行业知识进行进一步评估与优化，建立适合特定场景的评价体系。

### 2.2. 电商智能客服

智能客服系统是电子商务平台的重要组成部分，其主要目标是通过自动化方式回答用户常见问题，从而降低人工客服压力。传统智能客服系统通常基于规则匹配或检索式问答技术实现，通过关键词匹配或知识库检索生成回复[10]。

近年来，随着深度学习技术的发展，基于生成式模型的客服系统逐渐应用于客服领域。Zhang 等[11]研究了基于预训练语言模型的客服对话生成方法，发现微调后的模型在电商场景下能够生成更加自然流畅的回复。但由于微调的模型存在着知识过时的风险，导致模型在回答问题错误。因此 Feng 等[12]提出了基于 RAG 增强型检索技术的电商本地知识库与轻量化本地部署大语言模型，通过外搭知识库的方式解决模型参数知识过时的问题。Deriu 等人[13]指出，现有客服系统评估多关注对话流畅度，而忽视了业务知识的准确性。因此，对模型回答进行准确评估具有重要意义，需要建立能够反映业务质量的评价体系。

### 2.3. 模型事实性评估方法

针对大语言模型在生成过程中产生的内容与既定事实不一致的问题，根据评估机制的不同，现有方法可分为基于规则匹配的自动评估、基于神经网络的语义评估、基于人工判断的真实性评估以及基于大语言模型的自动化评估四类[14]。

#### 2.3.1. 基于规则的自动评估指标

基于规则匹配评估指标因具有一致性、可预测性和易于实现的特点，被广泛应用于事实性评估。精确匹配(Exact Match)要求生成文本与参考答案逐字一致，常用于开放域问答任务。BLEU 指标通过计算生成文本与参考文本之间的 n-gram 共现频率评估短语级一致性，而 ROUGE 指标则基于召回率衡量最长公

共子序列和跳词二元组的匹配程度。METEOR [15]在 BLEU 基础上引入显式词匹配和调和平均，以解决召回率不足和高阶 n-gram 缺失问题。

#### 2.3.2. 基于神经网络的语义评估指标

神经网络评估指标通过学习评估模型来比较生成输出与参考文本。BERTScore 利用预训练 BERT 模型的上下文嵌入计算生成句与参考句的语义相似度，通过召回率、精确率和 F1 值量化匹配程度[16]。BLEURT [17]通过在大规模合成句对上预训练 BERT，并结合词汇和语义级监督信号，使其能够准确估计人工评分。BARTScore [18]将评估视为文本生成问题，使用预训练的序列到序列模型 BART 计算生成文本与参考文本之间的翻译概率，分数越高表明准确性和流畅性越好。

#### 2.3.3. 基于人工判断的真实性评估

人工评估在事实性评估中至关重要，能够捕捉自动化系统难以识别的语言细微差别。FActScore [19]针对长文本生成的事实精确性评估，将生成内容分解为原子事实(传达单一信息的短陈述)，逐一验证于可靠知识源，最终分数为被支持原子事实的百分比。该指标在传记生成任务中揭示，在商业大模型中，事实精确率为 42%至 71%，准确率随着实体稀有度增加而显著下降。

#### 2.3.4. 基于大语言模型的自动化评估

利用 LLM 自身进行评估具有效率高、可扩展性强、减少人工标注依赖等优势。GPTScore [20]利用 19 个不同规模的预训练模型的零样本指令遵循能力评判生成文本，支持自定义多维度评估。LLM-Eval [21]采用统一的提示模式在单次模型调用中评估对话质量的多个维度，实验表明其有效性和适应性优于传统评估方法。然而，此类方法存在验证不足导致的偏见风险，且评估范围常局限于开放域对话。

#### 2.3.5. 与本文方法的对比分析

上述方法在通用事实性评估方面取得了显著进展，但在电商客服这一垂直领域存在适配不足：基于规则的方法(如 BLEU、ROUGE)仅关注词汇层面相似度，无法评估业务知识的语义正确性；基于神经网络的方法(如 BERTScore)需要对模型进行训练，且评估质量受训练数据分布的影响；基于人工判断的方法(如 FActScore)依赖通用知识库(如维基百科)，缺乏对电商专有术语(如“满减”、“预售”、“定金膨胀”)和动态业务规则(如实时库存、促销政策)的覆盖；基于大语言模型的方法虽具备灵活性，但在垂直领域的专业知识覆盖和评估稳定性方面仍有局限。

针对上述问题，本文提出 KCA (Knowledge Component Accuracy)和 KCR (Knowledge Component Recall)两种评估指标。KCA 设计源于专业领域大模型评估中对“事实准确性”的核心要求，KCR 则源于对“知识完备性”的评估需求。两种方法针对电商客服场景进行了系统性适配，具体设计详见第 4 章。

## 3. 电商知识问答评价问题分析

### 3.1. 业务知识错误无法识别

在电商客服问答中，一些问题涉及明确的业务规则，例如退货政策或售后条件。如果模型回答出现业务规则错误，可能会误导用户，甚至引发客诉纠纷。然而，传统评价指标难以识别这种错误。例如，当用户提出“商品支持七天无理由退货吗？”时，平台实际政策可能是“部分商品支持七天无理由退货，定制商品除外”。如果模型回答为“所有商品都支持七天无理由退货”，虽然与参考答案在词汇层面较为相似，但在业务规则上明显错误。在这种情况下，BLEU 仍可能得到较高分数，从而导致评价结果偏离实际情况，如表 1 所示。

**Table 1.** Examples of business knowledge errors
**表 1.** 业务知识错误案例

| 项目 | 内容 |
|---|---|
| 用户问题 | 商品支持七天无理由退货吗？ |
| 参考答案 | 大部分商品支持七天无理由退货，但定制商品、鲜活易腐商品除外。 |
| 模型回答 | 所有商品都支持七天无理由退货，您可以在收货后七天内申请。 |
| BLEU 分数 | 0.68 |
| 问题分析 | 模型回答遗漏了例外情况，业务规则错误 |

### 3.2. 业务规则缺失

如表 2 所示，从表中可见，模型回答与参考答案的 BLEU 分数达到 0.68，表面相似度较高。但模型回答遗漏了“定制商品除外”这一关键限制条件，可能导致用户错误申请退货。传统指标无法捕捉此类业务规则错误。

**Table 2.** Business rule errors
**表 2.** 业务规则错误

| 项目 | 内容 |
|---|---|
| 用户问题 | 如何申请退款？ |
| 参考答案 | 在订单页面点击“申请退款”，需在收货后 7 天内提交，仅未确认收货订单支持。 |
| 模型回答 | 您可以在订单页面找到退款入口，点击申请即可。 |
| BLEU | 0.45 |
| 问题分析 | 模型回答缺少时间限制和适用条件，信息不完整 |

### 3.3. 语义表达差异

在自然语言中，同一业务规则可以通过多种方式表达。例如，参考答案可能为“退款需在 7 天内申请”，而模型回答为“用户可在收货后七日内提交退货申请”。虽然两者表达方式不同，但语义完全一致。然而，BLEU 等指标对词汇匹配较为敏感，在这种情况下可能给出较低分数，从而低估模型回答质量。这反映出传统指标在语义理解方面的局限性，无法准确判断不同表述方式下的知识等价性。

## 4. 电商知识问答评价指标设计

为了解决上述问题，本文提出一种基于知识要素的评价方法，通过对客服知识进行结构化拆分，并设计新的评价指标来衡量模型回答质量，如图 1 所示。本节详细阐述指标设计的理论依据、形式化定义及计算方法。

### 4.1. 指标设计的理论依据

本文指标设计参考了专业领域大模型评估的研究成果。在医疗领域，医学大模型的评估强调诊断建议的准确性与治疗方案的完整性[22]。在法律领域，专业评估方法将模型输出分解为可验证的知识单元，分别评估其正确性、完整度与相关度[23]。这些研究的核心设计理念包括：1) 知识要素导向：将模型输出分解为可验证的知识单元，分别评估每个单元的质量；2) 事实准确性优先：设置事实准确性作为核心评估维度，对虚构、错误的信息给予严厉扣分；3) 知识完备性要求：评估模型是否覆盖问题所需的全部关键知识点，遗漏重要信息将导致评分降级。参考上述专业领域评估方法的设计理念，本文针对电子商

务客服领域的知识结构与决策特点，提出知识要素准确率与知识要素覆盖率两个专业评估指标。电商客服知识具有明确的结构化特征，业务规则可被形式化为离散的知识单元，这为知识要素级别的评估提供了基础。

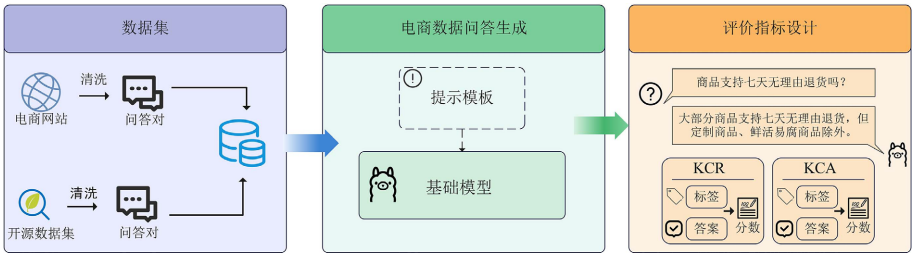


**Figure 1.** Model evaluation flowchart
**图 1.** 模型评价流程图

## 4.2. 知识要素定义

在电商客服知识中，许多业务规则可以被表示为结构化信息。本文将最小业务知识单位定义为“知识要素”。知识要素是电商客服知识中不可再分的最小语义单元，通常可表示为三元组结构$(e_h, r, e_t)$，其中$e_h$表示头实体(如商品、订单、用户等)，$r$表示关系属性(如退货期限、退款条件、配送方式等)，$e_t$表示尾实体(具体的属性值)。例如，在退货政策中可以提取以下知识要素：(商品，退货期限，7 天)，(商品，退货条件，未使用)，(商品，退货方式，在线申请)，(定制商品，退货政策，不支持七天无理由)。

通过将客服知识拆分为多个知识要素，可以更清晰地描述业务规则结构，为精细化评估奠定基础。知识要素抽取方法：对于简短明确的答案，采用基于正则表达式的数值与实体抽取方法。对于复杂结构化长文本答案，则通过大语言模型语义解析将其分解为独立原子事实陈述并映射为三元组集合，如图 2 所示。

```
你是一名知识抽取专家。请从以下{' 参考答案' if source_type == 'reference' else ' 模型回答'}中提取结构化的知识元素。
需要提取的知识元素类型：
1. entity（实体）：具体的产品、设备、部件、组织、人员等名词
   示例：手机屏幕、充电器、电源键、官方售后
2. attribute（属性）：描述实体特征的信息
   示例：价格300元、保修期1年、屏幕尺寸6.5寸
3. relation（关系）：实体之间的关联
   示例：手机-包含-屏幕、充电器-连接-手机
4. constraint（约束/限制）：条件限制、规则、政策
   示例：保修期内免费、需要购机凭证、非人为损坏
5. action（操作/行为）：建议的操作步骤或行为
   示例：强制重启、连接充电器、备份数据、前往维修点
6. condition（条件）：前提条件或触发条件
   示例：如果屏幕不亮、当电量耗尽时、在保修期内
7. numerical（数值信息）：具体的数字、范围、时间、金额
   示例：300-500元、10秒、7天、3个月
请以 JSON 格式返回，格式如下：
{{
 "knowledge_elements": [
   {{
    "type": "entity",
    "content": "手机屏幕",
    "value": null,
    "unit": null
   }},
   {{
    "type": "numerical",
    "content": "价格区间",
    "value": "300-500",
    "unit": "元"
   }},
   {{
    "type": "action",
    "content": "强制重启",
    "value": "10秒",
    "unit": "秒"
   }}
 ]
}}
注意：
尽可能提取所有关键信息，不要遗漏
数值信息要单独提取，包含单位和范围
如果某个字段没有值，使用 null
只返回 JSON，不要有其他内容"""
```

**Figure 2.** Prompt template for semantic parsing and knowledge element extraction of large models
**图 2.** 大模型语义解析抽取知识元素提示模板

### 4.3. 知识要素准确率

知识要素准确率(Knowledge Component Accuracy, KCA)的设计源于专业领域大模型[22]评估中对“事实准确性”的核心要求。在医疗领域评估标准中，诊断建议的准确性被置于首位，错误的医疗建议可能导致严重后果。同理，在电子商务客服领域，业务规则的准确性直接关系到用户体验与平台信誉。若模型推荐的退货政策与实际规则不符、退款条件描述错误或配送时间承诺失实，将导致用户操作失败、客诉增加甚至法律纠纷。因此，必须设计专门指标来量化模型生成内容中正确知识要素的比例，以直接反映模型抑制“幻觉”、生成可靠事实的能力。知识要素准确率定义为模型生成答案中正确知识要素占生成知识要素总数的比例，计算公式如公式(1)所示：

$$\mathrm{KCA} = \frac{\left|K_{pred} \cap K_{gt}\right|}{\left|K_{pred}\right|} \tag{1}$$

其中，$K_{pred}$ 表示模型生成答案中提取的知识要素集合，$K_{gt}$ 表示标准答案中的知识要素集合，$\left|K_{pred} \cap K_{gt}\right|$ 表示两者中一致的知识要素数量。若 $K_{pred} = 0$ (即模型未生成任何实质性知识要素)，则定义 $\mathrm{KCA} = 0$。

KCA 指标的核心价值在于惩罚错误知识的生成。当模型产生“幻觉”或业务规则错误时，错误知识要素会被计入分母但不计入分子，从而导致 KCA 分数下降。这一设计确保模型不能通过生成大量模糊或错误信息来获得高分，必须保证生成内容的准确性。

### 4.4. 知识要素覆盖率

知识要素覆盖率(Knowledge Component Recall, KCR)的设计依据源于专业领域评估中对“知识完备性”的要求。在法律领域评估标准[23]中，“完整度评分”明确要求模型应覆盖用户问题所需的全部关键法律知识点，遗漏重要条文或程序将导致评分降级。在电子商务客服领域，一个完整的决策建议通常由若干关键知识要素构成，包括时间限制、适用条件、操作步骤等核心信息单元。若模型仅回答部分内容(如只告知申请入口但未说明时间限制)，用户在实际操作中仍可能因信息缺失而导致申请失败。因此，必须设计指标来衡量标准答案中关键知识点被模型覆盖的程度，以反映模型对问题所需知识体系的掌握深度与回答全面性。知识要素覆盖率定义为标准答案中被成功召回的知识要素占标准知识要素总数的比例，计算公式如公式(2)所示：

$$\mathrm{KCR} = \frac{\left|K_{pred} \cap K_{gt}\right|}{\left|K_{gt}\right|} \tag{2}$$

KCR 关注模型是否遗漏重要知识。与 KCA 不同，KCR 的分母固定为标准答案的知识要素数量，因此该指标主要反映回答的完整性而非准确性。

## 5. 实验

### 5.1. 实验设置

实验数据来源于电商客服问答多轮数据集 [1]，包含 100 个典型用户咨询问题。为全面评估所提出指标的有效性，本文选取了 Qwen-7B、DeepSeek-V3、GLM-4 这三个主流大语言模型进行对比分析。模型推理在 1 张 NVIDIA A100-40GB 显卡上进行，采用批量推理方式提高效率。知识要素抽取使用 Qwen-32B 辅助完成，经人工校验确保准确性。实验同时计算传统指标(BLEU、ROUGE-1)以及本文提出的 KCA、KCR 和 KCR_fuzzy 指标，以便对比分析，值越大越好。

[1] https://www.modelscope.cn/datasets/xuri2004/dianshang_dataset

## 5.2. 实验结果

表 3 展示了各模型在电商客服任务上的实验结果，评测基于知识元素提取方法。如表 3 所示，本文提出的指标能够有效区分模型在知识覆盖和准确性方面的差异。

从传统指标来看，GLM-4 在 BLEU 和 ROUGE-1 指标上表现最优(分别为 82.18 和 83.45)，表明其生成文本与参考答案的表面相似度最高。然而，在 KCA 指标上，DeepSeek-V3 取得了 68.35 的最高分，说明该模型在业务知识准确性方面表现最佳，其生成的知识元素错误率最低。这一差异反映出传统指标与知识质量指标评估维度的不同：高流畅度并不等同于高准确性。

值得注意的是，所有模型在 KCR 指标上的得分(平均约为 58.67)均显著低于 KCA 指标(平均约为 60.82)，且普遍低于传统指标。这表明当前大语言模型在电商客服任务中普遍存在知识遗漏问题，难以完整覆盖参考答案中的全部关键知识元素(如检查步骤、约束条件、数值参数等)。KCR 指标有效揭示了这一被传统指标掩盖的短板。

**Table 3.** Experimental scores
**表 3.** 实验分数

| 模型 | BLEU | ROUGLE-1 | KCA | KCR |
|---|---|---|---|---|
| DeepSeek-V3 | 74.25 | 72.86 | **68.35** | 58.62 |
| GLM-4 | **82.18** | **83.45** | 62.18 | **64.45** |
| Qwen-7B | 76.42 | 77.18 | 55.42 | 52.18 |

## 5.3. 案例分析

在案例研究部分，本文选取了两个具有代表性的电商客服问答案例，展示不同指标对模型回答的评估差异，结果如表 4 所示，针对“退货政策咨询”这一具体案例，用户提出的问题聚焦于产品出现质量问题后的换新可行性。

**Table 4.** Case Study
**表 4.** 案例分析

| 案例 1：退货政策咨询 | |
|---|---|
| 用户问题 | 我的产品如果有质量问题我能直接换新吗？ |
| 标签 | 关于退换政策如下：1). 您的耳机购买时间在 1 年保修期内。2) 保留着原始购买凭证。根据三包规定，如果是非人为造成的性能故障，7 天内可退换，15 天内可换新。 |
| ChatGLM | 关于您提到的“直接换新”，这需要根据具体的故障原因以及您的购买时间来判定：1) 如如果是产品质量问题：且在保修期内，按照国家三包规定，提供免费维修或换新服务。2) 如果判定为非质量问题：或者是超出了保修范围，可能无法直接换新。(KCA = 0.35, KCR = 0.22, BLEU = 0.61) |
| DeepSeek | 关于能否直接换新，主要取决于以下几点：1) 购买时间：若在退换货期内(通常 7~15 天，以电商或店铺政策为准)，可凭断连问题申请退换。若在保修期内(通常 1 年)，需按保修流程处理。2) 故障原因判定：非人为因素(如内部元件故障、设计缺陷等)属于质量问题，可要求换新或维修。人为或环境因素(如进水、摔损、强干扰环境使用等)可能不在保修范围。容性问题(如与特定设备不匹配)需双方协商解决。品牌售后政策：不同品牌对“断连”问题的处理标准不同，建议优先联系官方客服或查看保修条款。(KCA = 0.94, KCR = 0.89, BLEU = 0.73) |
| 分析 | 由于 DeepSeek 回答的更准确且知识要素更多因此分数更高，ChatGLM 因为其缺失时间知识要素导致分数较低。 |

在对比两个模型的生成结果时，观察到显著的评估差异。DeepSeek 模型的回答在结构和内容上均表现出更高的专业水准，其不仅涵盖了购买时间、故障原因判定等关键知识要素，还进一步区分了非人为因素与人为因素的处理差异，从而得到了更高的分数。相比之下，ChatGLM 的回答虽然提及了保修期和维修服务，但缺失了“7 天/15 天”这一关键的时间界定要素，导致回答的分数不高。

从量化评估的角度来看，这种内容上的差异在不同指标上得到了鲜明的体现。DeepSeek 在知识准确性与知识覆盖率指标上分别获得了 0.94 和 0.89 的高分，而 ChatGLM 的对应得分仅为 0.35 和 0.22。这一巨大的分差验证了 KCA 与 KCR 指标的有效性，证明其能够精准识别模型回答中的业务知识错误与信息缺失问题。

## 6. 结论

本文针对电子商务智能客服场景中大语言模型评价问题进行了系统研究。通过分析传统文本生成评价指标在电商知识问答任务中的局限性，指出其难以识别业务规则错误、知识缺失等问题。这些局限性导致传统指标无法准确反映模型在电商客服场景中的真实表现，可能误导模型选型与优化方向。为解决上述问题，本文提出基于知识要素的评价方法，并设计了知识要素准确率(KCA)与知识要素覆盖率(KCR)两种指标，用于衡量模型回答中的知识正确性与信息完整性。实验结果表明，与传统评价指标相比，该方法能够更加准确地评估大语言模型在电商客服知识问答任务中的能力，有效识别业务错误与知识缺失问题。

## 参考文献


[1] 本刊编辑部. 第 57 次《中国互联网络发展状况统计报告》在京发布[J]. 中国教工, 2026(2): 66.

[2] Brown, T., Mann, B., Ryder, N., *et al*. (2020) Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, *NeurIPS* 2020, 6-12 December 2020, 1877-1901.

[3] Touvron, H., Lavril, T., Izacard, G., *et al*. (2023) LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971

[4] Bai, J., Bai, S., Chu, Y., *et al*. (2023) Qwen Technical Report. https://arxiv.org/abs/2309.16609

[5] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics*, Philadelphia, 6-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[6] Lin, C.Y. (2004) Rouge: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, Barcelona, 25 July 2004, 74-81.

[7] Vaswani, A., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[8] Zeng, A., Liu, X., Du, Z., *et al*. (2022) GLM-130B: An Open Bilingual Pre-Trained Model. https://arxiv.org/abs/2210.02414

[9] Liu, A., Feng, B., Xue, B., *et al*. (2024) DeepSeek-V3 Technical Report. https://arxiv.org/abs/2412.19437

[10] 王思宇, 邱江涛, 洪川洋, 等. 基于知识图谱的在线商品问答研究[J]. 中文信息学报, 2020, 34(11): 104-112.

[11] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D. and Weston, J. (2018) Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too? *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 2204-2213. https://doi.org/10.18653/v1/p18-1205

[12] 冯源, 钱松荣, 陆宇亮. 基于 RAG 本地电商知识库的 DeepSeek 电商模型构建与优化研究[J]. 电子商务评论, 2025, 14(5): 1346-1359.

[13] Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., *et al*. (2021) Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, **54**, 755-810. https://doi.org/10.1007/s10462-020-09866-x

[14] Wang, C., Liu, X., Yue, Y., *et al*. (2023) Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. https://arxiv.org/abs/2310.07521

[15] Banerjee, S. and Lavie, A. (2005) METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, June 2005, 65-72.

[16] Zhang, T., Kishore, V., Wu, F., *et al*. (2019) BERTScore: Evaluating Text Generation with BERT. https://arxiv.org/abs/1904.09675

[17] Sellam, T., Das, D. and Parikh, A. (2020) BLEURT: Learning Robust Metrics for Text Generation. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 7881-7892. https://doi.org/10.18653/v1/2020.acl-main.704

[18] Yuan, W., Neubig, G. and Liu, P. (2021) Bartscore: Evaluating Generated Text as Text Generation. *Advances in Neural Information Processing Systems* 34: *Annual Conference on Neural Information Processing Systems* 2021, *NeurIPS* 2021, 6-14 December 2021, 27263-27277.

[19] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P., *et al*. (2023) FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 12076-12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

[20] Fu, J., Ng, S., Jiang, Z. and Liu, P. (2024) GPTScore: Evaluate as You Desire. *Proceedings of the* 2024 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 6556-6576. https://doi.org/10.18653/v1/2024.naacl-long.365

[21] Lin, Y.T. and Chen, Y.N. (2023) LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. *Proceedings of the* 5*th Workshop on NLP for Conversational AI* (*NLP4ConvAI* 2023), Toronto, 14 July 2023, 47-58.

[22] Singhal, K., Tu, T., Gottweis, J., *et al*. (2023) Towards Expert-Level Medical Question Answering with Large Language Models. https://arxiv.org/abs/2305.09617

[23] 许建峰, 刘程远, 况琨, 何浩, 孙常龙, 李宝善, 等. 法律大模型评估指标和测评方法[J]. 中国人工智能学会通讯, 2024, 2(14): 10-22.