FineSurE: Fine-grained Summarization Evaluation using LLMs

摘要

本文复现并改进了 FineSurE 框架，旨在对大语言模型生成的文本进行细粒度评估。通过在 FRANK 和 REALSumm 数据集上的实验，验证了 FineSurE 在忠实性、完整性和简洁性三个维度上的卓越性能，尤其在与人类评估的一致性方面展现了出色的表现。复现过程中还探讨了不同 LLMs 在 FineSurE 框架下的表现，发现模型选择和调优对评估任务的成功至关重要。适当调优的开源模型能够显著提高评估性能，为 FineSurE 框架的广泛应用提供了新的可能性。尽管 FineSurE 框架展现出显著的优势，但仍存在挑战和未来工作的方向，包括多模型融合与优化、数据集的扩展与多样性以及增强评估指标等。

关键词： FineSurE；文本摘要评估；自动化评估；大模型

1 引言

随着自然语言处理（NLP）技术的不断发展，文本摘要作为其中的重要任务之一，已成为研究的热点。文本摘要任务旨在从长篇文章中提取出关键信息，并生成简洁、准确的摘要 [8]。随着大语言模型（LLMs）的应用，文本摘要的质量得到了显著提升，但在实际应用中，如何准确评估生成的摘要质量仍然面临诸多挑战 [10, 15]。传统的自动化评估方法，尤其是基于相似性的指标如 ROUGE，长期以来被广泛应用于文本摘要的评估。然而，研究表明，这些方法与人工评估的相关性较弱，导致其在准确评估摘要质量方面存在不足 [14]。尤其是当生成摘要的内容较为复杂时，ROUGE 等传统评估方法无法提供足够的细粒度分析，忽略了生成文本中的细节问题。与之相比，人类评估虽然能够提供更加细致的质量判断，但其高昂的成本和耗时性使得其在大规模评估中难以应用。

为了解决这一问题，自动化评估器成为近年来研究的重点。自动化评估方法不仅能够显著减少人工干预，提高评估效率，而且能够通过更为精细的评估维度提供更加准确的质量反馈 [4]。当前的研究尝试了多种方法，例如基于神经语言推理（NLI）和问答（QA）的自动评估框架 [2]，并证明大语言模型在模拟人类评估方面具有潜力 [14]。

然而，尽管这些方法取得了一定的进展，仍然存在粗粒度评估和评估维度不明确的问题。具体来说，粗粒度评估通常是在摘要层面评估忠实性、连贯性和相关性等维度 [4, 5]。这种 Likert 量表评分方法缺乏关于生成摘要中错误的详细信息，例如没有列出有质量问题的摘要句子的数量，或未具体说明每个句子中存在的错误类型。另一方面，由于“所有句子的集体质量”和“从源文本中选择重要内容”的定义不明确阻碍了对连贯性和相关性的评估 [2, 21]。因此，亟需开发一个更精确的评估框架，清晰定义评估维度，生成更详细的评估结果。

针对这些问题，复现论文提出了一种新的自动化评估框架—FineSurE（Fine-grained Summarization Evaluation）。该框架通过从更细粒度的角度出发，综合考虑忠实性、完整性和简洁性三个维度来评估生成摘要。与传统的粗粒度评估方法不同，FineSurE 方法通过两个核心步骤：事实检查和关键信息对齐，能够在句子和关键信息层面精确地识别出摘要中的质量问题。这一框架能够对生成文本进行更为细致的分析，避免了传统方法在评估维度上存在的模糊性问题，并且能够提供更为细致和可操作的评估结果。

复现论文的研究意义在于：首先，针对 LLM 生成的摘要存在的信息遗漏、冗长等问题，提出了更为精确的评估维度；其次，提出了基于关键信息对齐的新型评估框架—FineSurE，该框架能够利用 LLM 生成关键信息，将其对齐至摘要句子，并自动分类摘要中的错误，提供比传统评估方法更为详细的质量反馈；最后，通过与基于相似性、NLI、QA 和 LLM 的方法进行了全面的实验对比，验证了 FineSurE 框架在与人工评估相关性上的优势，为进一步优化文本摘要评估提供了新的思路和方法。

2 相关工作

在自然语言处理领域，自动化评估由语言模型生成文本的质量已成为一个重要的研究方向。近年来，研究者们针对文本生成质量评估提出了多种自动化评估方法，应对不同任务中的评估需求。主要包括基于相似性、基于神经语言推理（NLI）、基于问答（QA）和基于大语言模型（LLM）的评估方法。

2.1 基于相似性的评估方法

基于相似性的自动化评估方法主要依赖于生成文本与参考文本之间的 n-gram 重叠情况，常用的评估指标包括 ROUGE [13]、BLEU [18] 和 METEOR [1]。这些方法通过计算生成文本和参考文本之间的词汇重合度来评估文本质量。然而，传统的相似性评估方法通常只能通过精确匹配来判断文本质量，缺乏人类评估所需的多维度分析。为了解决这一问题，一些研究者利用上下文嵌入（如 BERTScore [26]、MoverScore [27]、BARTScore [25]）通过计算生成文本与参考文本之间的语义相似度来提高评估的准确性。尽管这些方法在一定程度上改进了评估的效果，但它们依然局限于单一维度的评分，难以全面评价文本的多维度质量。

2.2 基于神经语言推理（NLI）的评估方法

随着 NLI 任务的发展，研究者们开始探索通过事实检查来评估生成文本的质量。NLI 方法通过从输入文本中提取相关证据来支持生成文本中的观点 [6]。例如，DAE [7] 提出了基于依赖关系的蕴含推理方法，更精细地评估生成文本的忠实性；SummaC [12] 通过将输入文本划分为句子单元并聚合句子对之间的 NLI 得分，提出了一种轻量级的 NLI 模型来评估摘要的质量。尽管这些方法在提高忠实性评估性能方面表现出色，但它们的评估仍主要集中在忠实性这一维度，缺乏对文本质量其他方面的综合评价。

2.3 基于问答（QA）的评估方法

基于问答的评估方法则通过生成参考文本的合理问题，并结合生成的文本对这些问题进行回答来评估文本的质量 [20]。QAGS [23] 和 QAFactEval [2] 通过问答任务提高了忠实性评估的准确性，并且在文本摘要任务中优于传统的基于相似性和 NLI 的方法。UniEval 提出了一个统一的评估框架，能够通过问答任务评估文本生成的多维度质量，包括忠实性、连贯性、相关性和流畅性四个维度。这些方法通常需要训练神经模型来生成问题及其对应答案，因此其计算成本较高。

2.4 基于大语言模型（LLM）的评估方法

随着大语言模型（LLMs）的发展，研究者们开始尝试将 LLM 作为无需参考文本的自动化评估工具。近期的研究表明，LLM 在文本生成质量评估中展现出了巨大的潜力 [22]。有些研究通过编辑文本 [11]、原子事实 [16] 以及外部知识库 [3] 来评估忠实性，并尝试评估文本生成的多维度质量 [14]。虽然 LLM 在评估忠实性方面取得了一定的进展，但目前大多数方法仍集中在忠实性评估上，缺乏对摘要生成质量的全面、多维度分析。

与前述工作不同，复现论文提出了一种新型的细粒度评估框架——FineSurE，专门针对 LLM 生成的文本进行细粒度评估。这些细粒度的评估维度使得 FineSurE 能够在句子级别和关键信息层面提供更加详细的评估结果，从而解决了现有评估方法中存在的粗粒度评估和评估维度模糊的问题。通过与现有基于相似性、NLI、QA 和 LLM 的方法进行比较，FineSurE 框架在与人工评估的相关性上表现出色，展示了其在文本摘要质量评估中的优势。

3 本文方法

3.1 本文方法概述

本文提出了一个 FineSurE 框架，该框架旨在对大语言模型（LLMs）生成的文本进行细粒度评估，聚焦于忠实性（Faithfulness）、完整性（Completeness）和简洁性（Conciseness）三个关键维度。FineSurE 框架通过事实检查（Fact Checking）和关键信息对齐（Keyfact Alignment）两个任务，结合 LLM 的自动化能力，解决了现有评估方法在处理信息丢失、冗长和幻觉问题时的不足，框架图如图 1所示。FineSurE 不仅能准确评估生成文本的质量，还能提供详细的错误分类信息。整体框架通过定制化的提示工程和关键信息提取，优化了评估流程，提升了评估结果的细粒度和准确性。错误说明。

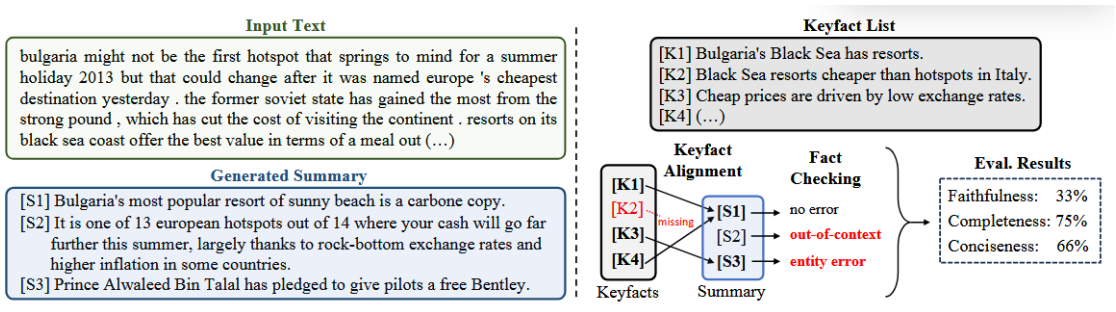

图 1. FineSurE 框架

3.2 评估维度

LLMs 在摘要生成任务中显著提高了生成质量，但仍然面临着幻觉、信息丢失和冗长等问题 [9,19]，因此论文对传统的“忠实性”评估进行扩展，增加对“完整性”和“简洁性”两个评估维度的考量，以全面评估生成文本的质量。

忠实性评估生成的文本是不是忠实于输入文本，不篡改原始信息，也不添加任何无法从输入文本直接推导的外部信息。完整性评估生成文本是否包含包含输入文本的所有关键信息。而简洁性评估生成文本是否避免生成输入文本的关键信息之外的内容，保持摘要的简洁。

此外，根据 [17] 的定义，原文将错误类型划分为七大类：外部错误（如“上下文错误”）、以及内在错误（包括“谓词错误”、“实体错误”、“情境错误”、“共指错误”、“话语连接错误”和“语法错误”）。

3.3 评估流程

在此部分，原论文设计了一套评估流程，包括事实检查和关键信息对齐两个核心任务。、每个任务都采用定制化的 Prompt，并输出 JSON 格式，提高指令执行成功率并便于解析。

事实检查任务将事实检查转化为分类任务如图 2所示，错误类别包括七种事实错误类型，以及“其他错误”和“无错误”两个额外类别。给定输入文本和生成摘要，LLM 应当输出每个句子的错误类型，并将其分类为九个类别之一，同时提供简明的错误说明。

关键信息对齐任务验证生成摘要的每个关键信息是否能够从生成文本中推导出，并标记出所有与关键信息匹配的摘要句子的行号。任务的输出包括每个关键信息的二元标签和相关句子行号的列表。

评估指标通过事实检查和关键信息对齐的结果计算得出。忠实性得分基于标记为“无错误”的摘要句子比例. 而完整性得分反映关键信息在摘要中的包含程度；简洁性得分则衡量与关键信息匹配的摘要句子占所有句子的比例，具体如公式（1-3）所示。

$$Faithfulness(D,S) = \frac{|S_{fact}|}{|S|}. \tag{1}$$

$$Completeness(K,S) = \frac{|\{k \mid (k,s) \in E\}|}{|K|}. \tag{2}$$

$$Conciseness(K,S) = \frac{|\{s \mid (k,s) \in E\}|}{|S|}. \tag{3}$$

与现有评估方法 [14,21,24] 相比，FineSurE 能提供更详细的句子错误类型信息以及关键信息与摘要句子之间的对齐关系。

3.4 提示工程和关键信息提取

FineSurE 框架中的提示工程采用多种策略，包括基本提示、指令式提示、分类提示、推理提示和证据映射提示。其中，忠实性评估建议结合分类和推理提示进行，而完整性和简洁性评估推荐使用指令式提示。这种设计既能引导 LLM 准确分类错误类型，又能确保关键信息对齐的高效性，具体如图 3所示。

关键信息列表对完整性和简洁性评估至关重要。尽管人工生成的关键信息列表在准确性上具有优势，但在某些情况下，获取人工关键信息存在挑战。FineSurE 提供了一种基于 LLM 自动提取关键信息的解决方案，通过定制化提示实现完全自动化，以适应不同领域的需求和场景。

```
You will receive a transcript followed by a corresponding summary.
Your task is to assess the factuality of each summary sentence
across nine categories:
* no error: the statement aligns explicitly with the content of
the transcript and is factually consistent with it.
* out-of-context error: the statement contains information not
present in the transcript.
* entity error: the primary arguments (or their attributes) of
the predicate are wrong.
* predicate error: the predicate in the summary statement is
inconsistent with the transcript.
* circumstantial error: the additional information (like location
or time) specifying the circumstance around a predicate is wrong.
* grammatical error: the grammar of the sentence is so wrong that
it becomes meaningless.
* coreference error: a pronoun or reference with wrong or non-
existing antecedent.
* linking error: error in how multiple statements are linked
together in the discourse (for example temporal ordering or
causal link).
* other error: the statement contains any factuality error which
is not defined here.

Instruction:
First, compare each summary sentence with the transcript.
Second, provide a single sentence explaining which factuality
error the sentence has.
Third, answer the classified error category for each sentence in
the summary.

Provide your answer in JSON format. The answer should be a list
of dictionaries whose keys are "sentence", "reason", and
"category":
[{"sentence": "first sentence", "reason": "your reason",
"category": "no error"}, {"sentence": "second sentence", "reason":
"your reason", "category": "out-of-context error"}, {"sentence":
"third sentence", "reason": "your reason", "category": "entity
error"},]

Transcript:
{input text}

Summary with N sentences:
{summary sentence 1}
{summary sentence 2}
        ...
{summary sentence N}
```

图 2. 事实检查任务的 Prompt

4 复现细节

4.1 与已有开源代码对比

在本研究的复现过程中，我们参考了 FineSurE 框架的开源代码，该代码库可在 GitHub 上的 FineSurE-ACL24 中找到。我们复现了开源代码中的事实检查和关键信息对齐模块，并对其进行了整理和优化，使得模块内容更加健壮。具体而言，事实检查模块能够将问题转化为多分类任务，并利用大语言模型（LLMs）识别和分类摘要中每个句子的错误类型。关键信息对齐模块则通过关键信息匹配来解决对齐问题，并输出每个关键信息的二元标签以及匹配的摘要句子行号。此外，我们还复用了开源代码中的多维度评估逻辑，包括忠实性、完整性和简洁性的计算方法。

```
You will receive a summary and a set of key facts for the same
transcript. Your task is to assess if each key fact is inferred
from the summary.

Instruction:
First, compare each key fact with the summary.
Second, check if the key fact is inferred from the summary and
then response "Yes" or "No" for each key fact. If "Yes", specify
the line number(s) of the summary sentence(s) relevant to each
key fact.

Provide your answer in JSON format. The answer should be a list
of dictionaries whose keys are "key fact", "response", and "line
number":
[{"key fact": "first key fact", "response": "Yes", "line number":
[1]}, {"key fact": "second key fact", "response": "No", "line
number": []}, {"key fact": "third key fact", "response": "Yes",
"line number": [1, 2, 3]}]

Summary:
[1] {summary sentence 1}
[2] {summary sentence 2}
          ...
[N] {summary sentence N}

M key facts:
{keyfact 1}
{keyfact 2}
   ...
{keyfact M}
```

图 3. 关键事实对齐任务的 Prompt

尽管我们参考了 FineSurE 的开源代码，但我们的工作在以下几个方面具有显著的创新和改进：

模型适配与优化：我们对开源代码进行了适配和优化，使其能够与我们选择的开源模型 qwen2.5 系列更好地协同工作，从而提高了评估的准确性和效率。此外，为了全面测试 FineSurE 框架的表现，我们还对不同参数量的 qwen2.5 模型进行了测试，进一步验证了框架在多种配置下的适用性和稳定性。

评估指标的增强：我们在原有评估维度的基础上，加入了额外的评估指标，例如错误类型的详细分布和关键信息的覆盖率，这些新指标为评估结果提供了更加丰富和多维的视角，使得评估结果更加全面和精确。

4.2 实验环境搭建

我们在 NVIDIA A100 GPU 配置的服务器上进行了所有实验，依赖于 vscode 编译器，使用 python3.10。开源模型 qwen2.5 等模型从 huggingface 上下载，模型部署和推理使用 vllm 框架，以便于集成和使用 LLMs。

为了确保实验的高效性和稳定性，所有实验均在配备 NVIDIA A100 GPU 的服务器上进行。此配置提供了强大的计算能力，能够支持大规模的模型训练和推理。实验环境使用 vscode 作为主要的开发编译器，Python 版本为 3.10，以保证与所用库和框架的兼容性。为了实现高效的模型推理和部署，我们选择了 vLLM 框架，它能够简化大语言模型的集成流程，并提高推理过程中的计算效率。

在模型方面，我们从 Hugging Face 平台下载了开源模型 qwen2.5、Mixtral 及其他相关模型。这些模型已预训练，并经过优化，适用于文本生成和评估任务。我们在部署过程中对模型进行了适配和调优，以确保其与 FineSurE 框架的兼容性，尤其是在处理大规模文本数据时，能够保持较高的评估准确性和推理速度。

4.3 创新点

模型适配与优化：我们对开源代码进行了深入的适配和优化，使其能够与我们选择的开源模型 qwen2.5-32b-instruct 更好地协同工作。通过这种优化，我们显著提高了评估的准确性和效率，尤其是在处理大规模数据集时，性能得到了有效提升。

数据集扩展：除了使用现有的 FRANK 和 REALSumm 数据集，我们还尝试扩展了数据集，包含了更多领域的数据，以测试 FineSurE 框架在不同领域下的适用性。这一扩展使得框架的评估能力得到了更全面的验证。

评估指标的增强：我们引入了额外的评估指标，如错误类型的详细分布和关键事实的覆盖率，为评估结果提供了更丰富的视角。

5 实验结果分析

在本节中，我们将详细分析实验结果，对实验内容和实验复现与改进结果进行阐述，来评估 FineSurE 框架的性能。

5.1 不同评估方法的比较

我们基于 FRANK 数据集对 FineSurE 框架的忠实性评估进行了测试。实验结果表明，FineSurE 在所有评估级别上都显著优于基于相似性、自然语言推理 (NLI) 和问答 (QA) 的评估方法。现有方法都没有提供句子级别的评估结果，而是依赖于摘要级别的评分，而 FineSurE 有能力评估每个句子是否包含事实错误，其实验结果与人类句子级别的判断显著一致。

通过使用开源模型 qwen2.5-32b-instruct 进行实验测试，如表1所示：我们可以看到实验结果接近甚至优于原论文中的性能表现，在句子级别的平均准确率达到了 87.3%，优于原论文使用的 GPT-4 专有模型的 86.4%。同时，我们观察到 FineSurE 在零样本学习环境下表现出色，进一步验证了其框架的广泛适用性。

在 REALSumm 数据集上，与仅提供单一综合分数的基于相似性的评估器相比，UniEval 和 G-Eval 产生四个不同的分数，分别用于评估忠实性、连贯性、相关性和流畅性。我们选取其中的连贯性和相关性分数来计算与人类分数的完整性和简洁性的相关性，前者代表关键信息的覆盖性，后者则反映信息密度。如表2和表3所示，FineSurE 在完整性和简洁性评估结果上与人类评估高度一致，并显著优于其他评估方法。此外，FineSurE 还提供了关键事实级别的评估结果，揭示了摘要中每个关键事实的匹配情况，这为进一步衡量摘要质量提供了重要参考。

表 1. FRANK 数据集的忠实性评估表现

Direction	**Meth**	**Sentence**	**Summary**		**System**
		bAcc (↑)	**Pearson Corr (↑)**	**Spearman Corr (↑)**	**Rank Corr (↑)**
Similarity	ROUGE1	-	0.314 (0.00)	0.312 (0.00)	0.866 (0.00)
NLI	SummaC-Conv	-	0.819 (0.00)	0.805 (0.00)	0.863 (0.00)
QA	QAFactEval	-	0.833 (0.00)	0.804 (0.00)	0.911 (0.00)
LLM(qwen2.5)	GEval	-	**0.839 (0.00)**	0.824 (0.00)	0.943 (0.00)
	FineSurE	**87.3%**	**0.839 (0.00)**	**0.838 (0.00)**	**0.950 (0.00)**

表 2. REALSumm 数据集的完整性评估表现

Direction	**Method**	**Summary**		**System**
		Pearson Corr (↑)	**Spearman Corr (↑)**	**Rank Corr (↑)**
Similarity	ROUGE-1	0.465 (0.00)	0.432 (0.00)	0.501 (0.00)
QA	UniEval	0.128 (0.00)	0.176 (0.00)	0.321 (0.07)
LLM(qwen2.5)	G-Eval	0.488 (0.00)	0.346 (0.00)	0.802 (0.00)
	FineSurE	**0.661 (0.00)**	**0.648 (0.00)**	**0.932 (0.00)**

5.2 不同模型的比较

我们比较了不同 LLMs 在 FineSurE 框架下的表现，并分析了模型选择对框架评估性能的影响。

如表4所示，qwen2.5-32b-instruct 模型在成功率上可与专有 LLMs 相媲美，而其他开源模型，如 Mixtral 模型的成功率较低。通过分析失败案例，我们发现主要问题集中在两方面：其一是输出格式不正确，某些模型未能生成符合预期的 JSON 格式输出，或生成了错误的 JSON 结构。其二是输出信息不完整，部分模型仅输出了几行句子或关键事实，未能覆盖任务需求。这些问题导致评估系统难以正确解析和利用数据，进而影响了框架的整体评估性能。这表明模型的选择对评估任务及 FineSurE 框架的稳定性至关重要。

表 3. REALSumm 数据集的完整性评估表现

Direction	Method	Summary		System
		Pearson Corr (↑)	Spearman Corr (↑)	Rank Corr (↑)
Similarity	ROUGE-1	0.391 (0.00)	0.367 (0.00)	0.342 (0.00)
QA	UniEval	0.084 (0.00)	0.119 (0.00)	-0.166 (0.33)
LLM(qwen2.5)	G-Eval	0.310 (0.00)	0.266 (0.00)	0.524 (0.00)
	FineSurE	**0.485 (0.00)**	**0.443 (0.00)**	**0.909 (0.00)**

此外，我们还探讨了开源 LLM 的参数规模对评估结果与人类评分一致性的影响。如表5和表6所示，实验结果显示，随着模型参数规模的增加，FineSurE 在评估任务中的表现显著提升。即便是开源模型，通过适当的指令调优，其在 FineSurE 任务中的性能也能显著改进。例如，Mixtral-8x7b 在指令调优后表现大幅提升，进一步说明了模型优化对评估性能的积极作用。

表 4. 不同 LLM 模型的忠实性评估表现

Type	LLM	Sentence	Summary		System	SucRate
		bAcc (↑)	PearCor(↑)	SpearCor(↑)	RankCorr(↑)	
Open-source	qwen2.5-14b-i	84.00%	0.800 (0.00)	0.807 (0.00)	0.900 (0.00)	97.50%
	qwen2.5-32b-i	**87.30%**	**0.839 (0.00)**	0.838 (0.00)	**0.95 (0.00)**	**98.40%**
	Mixtral-v0.1	48.70%	-0.0018 (0.00)	0.039 (0.00)	-0.383 (0.00)	59.60%
	Mixtral-v0.1-i	80.70%	0.717 (0.00)	0.717 (0.00)	0.833 (0.00)	83.90%
profietary	GPT-4-turto	86.30%	0.823 (0.00)	**0.841 (0.00)**	**0.95 (0.00)**	97.90%

表 5. 不同 LLM 模型的完整性的评估表现

Type	LLM	Summary		System
		Pear Corr(↑)	Spearman Corr(↑)	Rank Corr(↑)
Open-source	qwen2.5-14b-instruct	0.627 (0.00)	0.611 (0.00)	0.941 (0.00)
	qwen2.5-32b-instruct	0.648 (0.00)	0.661 (0.00)	0.932 (0.00)
	Mixtral-8x7B-v0.1	0.167 (0.00)	0.148 (0.00)	0.455 (0.01)
	Mixtral-8x7B-Instruct-v0.1	0.510 (0.00)	0.497 (0.00)	0.634 (0.00)
profietary	GPT-4-turto	**0.688 (0.00)**	**0.676 (0.00)**	**0.944 (0.00)**

表 6. 不同 LLM 模型的简洁性的评估表现

Type	LLM	Summary		System
		Pear Corr(↑)	Spearman Corr(↑)	Rank Corr(↑)
Open-source	qwen2.5-14b-instruct	0.450 (0.00)	0.420 (0.00)	0.900 (0.00)
	qwen2.5-32b-instruct	0.485 (0.00)	0.443 (0.00)	**0.909 (0.00)**
	Mixtral-8x7B-v0.1	0.096 (0.00)	0.104 (0.00)	0.264 (0.00)
	Mixtral-8x7B-Instruct-v0.1	0.380 (0.00)	0.360 (0.00)	0.687 (0.00)
profietary	GPT-4-turto	**0.510 (0.00)**	**0.461 (0.00)**	0.881 (0.00)

6 总结与展望

本文复现并改进了 FineSurE 框架，深入评估和分析其在大语言模型生成文本细粒度评估中的性能。实验结果表明，作为一种细粒度的文本摘要评估工具，FineSurE 能够有效地在句子级别和关键事实级别提供详细评估，显著优于传统的基于相似性、NLI 和 QA 的评估方法。通过在 FRANK 和 REALSumm 数据集上的实验，我们验证了 FineSurE 在忠实性、完整性和简洁性三个维度上的卓越性能，尤其在与人类评估的一致性方面，展现了出色的表现。

此外，我们还探讨了不同大语言模型（LLMs）在 FineSurE 框架下的表现，并对开源模型如 qwen2.5 进行了适配和优化。实验结果表明，模型选择和调优对评估任务的成功至关重要，且适当调优的开源模型能够显著提高评估性能。这为 FineSurE 框架的广泛应用提供了新的可能性，表明框架具备在不同场景下灵活适配的潜力。

尽管 FineSurE 框架在本研究中展现出了显著的优势，但仍有一些挑战和未来工作的方向值得探索：

多模型融合与优化：复现过程中，我们主要使用了 qwen2.5 等开源模型，未来可以通过多模型融合的方式来提升框架的整体表现，尤其是在不同语言环境和任务场景下的泛化能力。通过引入模型集成、混合推理等技术，可以进一步提高框架的稳定性和准确性。

数据集的扩展与多样性：目前的实验主要在比较小范围的领域进行，未来可以扩展到更广泛的领域，如医疗、法律等专业领域，以测试和改进 FineSurE 在不同领域的适用性和准确性。

增强评估指标：虽然当前框架在忠实性、完整性和简洁性方面已经做出了有效的评估，但仍然可以加入更多维度的评估指标，如流畅性、社会偏见等，进一步提升评估结果的多维度性和全面性。

参考文献

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

[3] Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281*, 2023.

[4] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*, 2023.

[5] Mingqi Gao and Xiaojun Wan. Dialsummeval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, 2022.

[6] John Glover, Federico Fancellu, Vasudevan Jagannathan, Matthew R Gormley, and Thomas Schaaf. Revisiting text decomposition methods for nli-based factuality scoring of summaries. *arXiv preprint arXiv:2211.16853*, 2022.

[7] Tanya Goyal and Greg Durrett. Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478*, 2020.

[8] Som Gupta and Sanjai Kumar Gupta. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65, 2019.

[9] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[10] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*, 2019.

[11] Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. Summedits: measuring llm ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, 2023.

[12] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. Summac: Revisiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

[13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[14] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[15] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[16] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

[17] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*, 2021.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[19] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

[20] Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*, 2021.

[21] Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. Large language models are not yet human-level evaluators for abstractive summarization. *arXiv preprint arXiv:2305.13091*, 2023.

[22] Xiaoming Shi, Jie Xu, Jinru Ding, Jiali Pang, Sichen Liu, Shuqing Luo, Xingwei Peng, Lu Lu, Haihong Yang, Mingtao Hu, et al. Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation. *arXiv preprint arXiv:2308.07635*, 2023.

[23] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*, 2020.

[24] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*, 2023.

[25] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

[26] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[27] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*, 2019.