利用大语言模型从云事件报告中进行结构化信息抽取与分析（进行中的论文）

Xiaoyu Chu
x.chu@vu.nl 阿姆斯特丹自由大学 荷兰

Shashikant Ilager
s.s.ilager@uva.nl 阿姆斯特丹大学 荷兰

Yizhen Zang
y.zang-3@student.tudelft.nl
代尔夫特理工大学 荷兰

Sacheendra Talluri
s.talluri@vu.nl 阿姆斯特丹自由大学 荷兰

Alexandru Iosup
a.iosup@vu.nl 阿姆斯特丹自由大学 荷兰

arXiv:2603.16818v1 [cs.PF] 17 Mar 2026

摘要

事件管理对于维持云计算服务的可靠性和可用性至关重要。云服务提供商通常会向公众披露事件报告，总结故障和恢复过程，以帮助最大限度地降低其影响。然而，这类报告往往篇幅冗长且结构松散，难以理解、分析和用于长期提升可信赖性。大语言模型的兴起为应对这一挑战带来了新的机遇，但目前对如何实现这一目标的研究仍不充分。本文探讨了利用前沿大语言模型从非结构化云事件报告中抽取关键信息。首先，我们从 3 家领先的云服务提供商（AWS、Azure 和 GCP）收集了 3,000 多份事件报告，并对这些样本进行人工标注。随后，我们设计并比较了 6 种提示策略，以抽取和分类不同类型的信息。我们考察了 6 种大语言模型，包括 3 种轻量级模型和 3 种最先进的模型（SotA），并评估了它们在不同数据集、模型、提示词和提取字段上的准确率、延迟和词元成本。研究得出了以下主要发现：(1) 大语言模型的元数据抽取准确率较高，具体取决于数据集，为 75%–95%。(2) 除分类任务外，少样本提示通常能提高元数据字段的准确率；由于输出词元更少，其延迟也更低，但需要多输入 1.5–2× 个词元。(3) 轻量级模型（例如 Gemini 2.0、GPT 3.5）在准确率、成本和延迟之间实现了良好的权衡；最先进的模型准确率更高，但成本和延迟显著增加。本研究提供了工具、方法和见解，助力借助大语言模型准确、高效地抽取事件报告信息。FAIR 数据和代码已公开发布于 https://github.com/atlarge-research/llm-cloud-incident-extraction。

关键词

事件报告、云计算、信息抽取、大语言模型（LLM）、根因分析、面向 IT 运维的人工智能（智能运维）

ACM 引用格式：
Xiaoyu Chu、Shashikant Ilager、Yizhen Zang、Sacheendra Talluri 和 Alexan- dru Iosup。2026。利用大语言模型从云事件报告中进行结构化信息抽取与分析（进行中的论文）。载于第 *17* 届 *ACM/SPEC* 性能工程国际会议伴会（*ICPE Companion '26*），*2026*年*5*月*4*日至*8*日，意大利佛罗伦萨。ACM，纽约，美国，8 页。https://doi.org/10.1145/3777911.3801103

1 引言

事件管理是云计算运营中一项重要且复杂的流程 [1, 18, 21]。云事件可能由多种原因引发，包括硬件或软件故障、部署和运行时错误、网络中断或安全漏洞 [12, 22]。此类事件可能中断用户服务，影响系统的可靠性。为了尽可能降低影响，云事件发生后，云运营商会发布公开事件报告 [4, 10, 23]，告知用户事件管理流程和根本原因。然而，这些报告内容复杂、篇幅冗长，且结构过于松散，难以理解，也不便用于长期分析和缓解。因此，尽管每起事件都会公开相应报告，但目前仍缺乏公开数据集，也缺少对这些事件的长期分析。如今，大语言模型已成为强大的工具，为解决此类问题带来了新的机遇。大语言模型已被有效应用于性能工程，例如日志解析 [3, 16], 和根因分析 [1, 7, 11, 21, 28]。因此，我们提出利用高级大语言模型，从公开的云事件报告中提取和分析关键信息。

我们设计了一套工作流，用于将大语言模型应用于报告数据抽取。我们的工作针对这一工作流过程中的三项主要挑战：首先，**目前没有公开可用的云事件报告标注数据集，也没有用于自动提取和分析报告关键信息的开源工具**。以往关于云事件和中断分析的研究依赖人工或基于规则的数据抽取 [12, 25], 由于报告篇幅较长且内容复杂，这种方式非常耗时且耗费人力。例如，在我们收集的 Azure 和 GCP 数据集中，报告平均长度超过 500 个词（见表1）。其次，**缺乏用于评估**

CCS 概念

•**计算机系统组织** → **可靠性**。

本作品采用 Creative Commons 署名 4.0 国际许可协议授权。*ICPE Companion '26*，意大利佛罗伦萨 © 2026 版权所有归所有者/作者所有。
ACM ISBN 979-8-4007-2326-1/2026/05
https://doi.org/10.1145/3777911.3801103

表1：云事件报告摘要。

ID	Name	时期	#行数	# 已标注	平均词数
1	AWS	2016年 - 2022年	774	150（19%）	151
2	Azure	2019年 - 2024年	127	95（75%）	575
3	GCP	2016年 - 2021年	2,186	215 (10%)	533
总计	3 个来源	2016年 - 2024年	3,087	460（15%）	—

将大语言模型应用于事件报告数据提取时，其性能和成本的评估方法及系统性评估尚不完善。尽管大语言模型已广泛用于信息抽取 [24, 30], 但它们在事件报告提取中的设计和应用仍不明确。具体而言，如何构建数据结构、如何设计提示词，以及如何将大语言模型应用于云事件分析，目前都尚未得到充分理解。第三，**针对事件报告开展长期统计特征分析的研究有限。**以往研究分析过云服务的事件报告 [12, 25], 但这些分析通常基于已标注数据集，且仅涵盖有限的字段，因此难以推广到长期数据集或多样化的事件报告。

针对这些挑战，我们的主要贡献如下：

(1) **数据集与工具箱构建（第 2 节）：**本研究首次提供开源数据集与工具箱，用于系统地探索和评估大语言模型在云事件报告结构化信息提取与分析方面的表现。我们收集了 3,000 多份云事件报告，并为报告数据评估整理了带注释子集。我们还开发了一个工具箱，用于利用大语言模型进行报告数据抽取、评估和分析。

(2) **大语言模型的适配与评估（第 3 节和第 4 节）：**本研究展示了最先进的大语言模型（如 GPT-4o）在事件信息抽取与分析方面的实用性。我们系统地评估了 6 个不同规模的大语言模型，并采用多种提示策略，将它们在不同评估指标下的准确率和成本进行比较。基于该数据集，我们在元数据抽取方面取得了 75%–95% 的准确率。我们还根据提示词和大语言模型的延迟与成本，提供了选择建议。

(3) **开放科学：**我们遵循开放科学的 FAIR原则，并发布数据集[1] 和软件[2]，以支持结果复现和后续研究。

2 数据收集与准备

2.1 数据集收集

云服务提供商经常通过各自的网络平台 [在线发布事件报告 4, 10, 23]不过，这些报告的结构各不相同，各云服务提供商公开信息的详尽程度也有所不同。为收集多样化的事件报告数据，我们从 3 家主要云服务提供商获取报告：AWS、Azure 和 GCP，涵盖 2016年 至 2024年。我们总共收集了约 3,000 份事件报告。每份报告都包含重要信息，例如服务名称、位置、用户症状和根本原因。

为收集数据，我们抓取并归档了 3 家云服务提供商的事件报告网页。随后，我们按照第2.3节所述的方法论处理数据。此外，我们

[1]Zenodo：https://zenodo.org/records/[2]
GitHub：https://github.com/atlarge-research/llm-cloud-incident-extraction

表2：提取字段摘要。图例：●表示从报告中完全提取，●表示由大语言模型推断。精确匹配 =完全匹配，词元级F1 = 词元级 F1，BS = BERTScore。

ID	提取字段	Type	数据集 AWS	数据集 Azure	评估指标 GCP 精确匹配	评估指标 TK	评估指标 BS
1	服务名称	实体	●	●	●	✓	
2	位置	实体	●	●	-	✓	
3	开始时间	实体	●	●	●	✓	
4	结束时间	实体	●	●	●	✓	
5	时区	实体	●	●	●	✓	
6	服务类别	类别	●	●	●	✓	
7	根因类别	类别	-	●	-	✓	
8	用户症状类别	多类别	●	●	●		✓
9	用户症状	text	●	●	●		✓
10	根本原因	text	-	●	-		✓

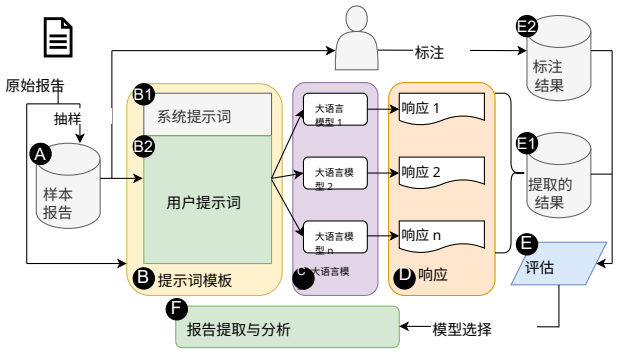


图1：基于大语言模型从云事件报告中进行数据抽取的工作流（结构化方法）概览。

人工标注了约 15% 的报告，作为评估的真实标注。尽管各提供商的标注数据比例有所不同，但最终样本量足以得出具有统计意义的结论；更多详情见第2.3节。表1汇总了本研究的数据集。

2.2 数据处理

在此阶段，我们处理原始数据集，以为提示构建做好准备。表2展示了识别出的用于提取的目标字段。这些字段被划分为实体、类别和自由文本片段，因为每种类型都需要不同的评估策略，具体内容将在第3.4节中详细讨论。因此，我们的数据处理包括以下步骤：(1) 原始数据集：我们从每个以 HTML 格式从外部抓取的文件中提取信息，并为每个运营商将其保存为 *Parquet* 文件。(2)清理后的数据集：由于每个供应商的报告结构不同，我们分别对其进行了处理，以筛除无效指标（例如空值和重复项）。之后，我们获得了结构相同的清理后的数据集，可轻松合并以进行统一分析。(3) 样本数据集：由于规模原因，对整个数据集进行标注并不现实，因此我们采用 *K-means* 聚类来选择具有代表性的样本数据集进行数据标注。(4)已标注数据集：我们从样本数据集中手动标注了表2所列的字段。遵循严格的数据标注流程（见第2.3节），我们获得了已标注数据集。

2.3 数据标注

数据标注是至关重要的一步，因为需要真实标注数据来与大语言模型提取的结果进行比较，并评估其准确率。由于没有可用于评估的公开标注云事件数据集

表3：模型配置。模型类型：L–轻量级模型；S–最先进水平。标有（*）的模型不提供时间指纹，但这些模型的运行日期见第4节。价格：T=词元。

模型别名	模型类型／名称	价格 [$/10⁶T] 输入	输出
GPT 4o	S / gpt-4o*	2.50	10.00
GPT 3.5	L / gpt-3.5-turbo*	0.50	1.50
Claude Sonnet 4	S / claude-sonnet-4-20250514	3.00	15.00
Claude 3.5	L / claude-3-5-haiku-20241022	0.80	4.00
Gemini 2.5	S / gemini-2.5-pro*	1.25	10.00
Gemini 2.0	L / gemini-2.0-flash*	0.10	0.40

表4：包含不同组件的六种提示策略。缩写：ZS=零样本；FS=少样本。

标签	组件
完整零样本	任务 + 思维链 + 类别 + 格式
Full-FS	任务 + 思维链 + 类别 + 示例 + 格式
基础零样本	任务 + 格式
基础少样本	任务 + 示例 + 格式
思维链零样本	任务 + 思维链 + 格式
Categ-ZS	任务 + 类别 + 格式

，我们创建了自己的标注数据集。为确保标注的准确性和一致性，我们选取了 3 名具有计算机系统背景的研究人员担任标注员。标注流程如下：首先，要求标注员 1 和 2 按照提供的说明，独立标注样本数据集。完成标注后，他们会比较结果以检查是否一致。如有分歧，他们会进行讨论以达成共识。如果无法达成一致，则由标注员 3 介入并作出最终决定。

3 大语言模型实验设置 3.1 使用大语言模型进行报告提取

图1展示了本研究的方法概述。完成数据收集和标注后，流程进入使用大语言模型从样本数据集（ ）中提取报告的阶段。首先，我们设计了一个提示词模板（ ），其中包含有关云事件的一般背景信息，以及分步骤的数据抽取说明，涵盖零样本和少样本设置。图2展示了一个完整模板示例，其中包含五个组成部分。根据云运营商的不同，模板也会略有变化，因为各运营商的报告在结构和可用信息方面有所不同。

对于每份报告，我们将其内容作为用户提示词的一部分插入模板中（ ）。我们为大语言模型设置系统提示词（ ），使其扮演系统操作员的角色，以帮助从云事件报告中执行数据抽取。随后，将完整提示词（连同系统提示词）发送给一组候选大语言模型（ ），并以 JSON格式收集模型的响应（ ）。我们通过将提取结果（ ）与标注数据集（ ）进行比较，评估其准确率（ ）；标注数据集作为基准

任务描述（任务）

逐步分析事件报告，抽取结构化信息。

思维链指令（思维链）

遵循以下推理步骤：1. 确定服务名称和服务位置。2. 从 {service_category_lst} 中选择一个最相关的服务类别。3. 提取描述用户症状的相关句子。然后，从 {user_symp_lst} 中选择一个或多个与提取出的症状最匹配的类别。4. 确定开始时间、结束时间和时区。时间格式为"HH:MM:SS"（24 小时制）。

分类指令（类别）

用户症状类别的定义为：{user_symp_instruction}。

回答格式（格式）

最后，将提取的信息以 JSON 对象形式返回，键为：*service_name, location, service_category, start_time, end_time timezone user_symptom* ,,, *user_symptom_category.*

少样本示例（示例）

以下是一些报告内容示例（标记为 **Q**）以及提取的信息（标记为 **A**）。**Q：**标题：**亚马逊 CloudWatch (爱尔兰)** 状态：[已解决] CloudWatch 指标延迟说明：(...) 太平洋标准时间下午 2:56，在 **上午 10:26至 下午 02:40** 太平洋标准时间期间，**我们遇到了 CloudWatch 日志事件处理延迟加剧的问题，影响了 EU-WEST-1 区域的指标筛选提取和日志订阅。**(...)

A："服务_名称"：**"亚马逊 CloudWatch"**，"位置"：**"爱尔兰"**，"服务_类别"：**"管理"**，"开始_时间"：**"10:26:00"**，"结束_时间"：**"14:40:00"**，"时区"：**"太平洋标准时间"**，"用户_症状_类别"：**"延迟"**，"用户_症状"：**"在 EU-WEST-1 区域，我们在指标筛选器提取和日志订阅的 CloudWatch 日志事件处理中遇到了更严重的延迟。"** *(未完待续.....)*

图2：用于从 AWS 报告中提取数据的 Full-FS 用户提示词模板，包含五个组件：任务、思维链、类别、格式**和**示例。**颜色：蓝色=表示已提取，红色=表示已分类。**

真实情况。我们还分析并评估了候选大语言模型的延迟和成本。最后，我们为每个运营商选出表现最佳的大语言模型，并开展事件报告提取与分析（ ）。

表5：GPT-3.5 在 AWS 数据集上采用不同提示策略时的抽取准确率（"准确率"）。准确率指标精确匹配、词元级F1、BS 见表2。对于每个字段，加粗 mygreen 文本表示表现最佳者，加粗 myred 文本表示表现最差者。

字段 / 提示策略	Acc.	完整零样本	Full-FS	基础零样本	基础少样本	思维链零样本	类别零样本
服务名称	EM	86.00	**100.00**	**52.00**	**100.00**	83.33	60.67
位置	EM	48.00	**96.67**	38.00	83.33	57.33	**44.67**
开始时间	EM	86.00	**91.33**	**70.00**	89.33	83.33	72.00
结束时间	EM	83.33	**86.00**	**64.00**	**86.00**	78.67	71.33
时区	EM	**98.67**	**98.67**	**98.67**	**98.67**	**98.67**	**98.67**
服务类别	EM	77.33	71.33	64.67	76.0	**80.00**	**61.33**
用户症状类别	TK	88.50	**88.94**	9.76	50.19	**9.33**	90.00
用户症状	BS	84.33	92.90	84.79	**93.79**	83.09	**81.63**
总体	**平均值**	71.08	**79.23**	**49.74**	73.60	61.44	62.44

3.2 环境设置与大语言模型

表3列出了我们在实验中选取用于比较的模型。我们从热门大语言模型提供商（OpenAI、Claude 和 Gemini）中选择了模型。为估算各模型的成本，表3还列出了官方网站公布的输入词元和输出词元价格。

3.3 提示工程

每个大语言模型在提示词的引导下从输入数据中抽取信息，而提示词会显著影响抽取结果的准确率和成本。为设计准确且有效的提示词，我们采用思维链（*CoT*）和上下文学习（*ICL*）提示技术，这些技术在大语言模型下游应用中日益流行 [15]。思维链提示词引导大语言模型逐步推理，从而获得更准确的答案 [14, 29]。相比之下，上下文学习直接在输入提示词中提供少量演示示例，使大语言模型能够执行多样化的下游自然语言处理任务 [6, 19]。

提示组件：为研究不同提示组件如何影响报告数据抽取，我们借鉴先前工作 [11] 定义了五个组件：任务描述（Task）、思维链指令（CoT）、分类指令（Category）、上下文示例（Examples）和回答格式（Format）。任务描述指示模型扮演系统操作员的角色，并从云事件报告中抽取信息。（思维链）该指令提供逐步指导，以促进对信息的推理。对于上下文示例，提供数据集中的两个样本来说明预期答案。最后，回答格式规定了模型应返回其输出的 JSON结构。

六种提示策略：通过以不同方式组合上述提示组件，我们构建了六种提示策略，如表4所示。其中，任务和格式是所有提示词共有的基础要素，其余组件则经过选择性添加，以评估它们各自及组合后的效果。

示例提示策略：图2展示了我们用于抽取报告信息的完整提示，其中包含少样本示例（Full-FS）。完整提示包含本研究考察的全部五个组成部分：任务、思维链、类别、示例和格式。相比之下，其他提示模板所包含的组成部分较少。

3.4 评估指标

准确率：大语言模型会抽取表2所列的所有字段，但不同字段的准确率各不相同。我们首先使用完全匹配（*EM*）评估实体和类别 字段的抽取准确率，该指标用于衡量抽取结果是否与标注的真实标注完全一致 [20]。对于用户症状类别这一多分类字段，我们采用分类任务的评估指标词元级 *F1*（*TK*）。该指标根据重叠词元计算精确率和召回率的调和平均值 [17, 24]。对于用户症状 和根本原因这类句子级文本字段，由于需要评估语义准确性，因此我们使用*BERTScore*（*BS*） [1, 32]，基于预训练BERT模型衡量语义相似度。

4 大语言模型的性能分析

4.1 提示策略比较

我们对第 3.3 节中介绍的六种大语言模型提示策略进行了实验比较。每种策略都以 GPT-3.5 为底层模型，在 AWS 数据集（表1）上进行评估。表5展示了各提示策略达到的准确率，各策略的平均值采用表格底行所报告的算术平均值计算。

发现 1： 正如预期，组件最丰富的策略 *Full-FS* 总体准确率最高（*79.23%*），并且在 *8* 个字段中的 *6* 个字段上是六种策略中表现最佳（单项最佳）的策略。上下文示例是效果最好的单一提示组件：将其添加到基础零样本后，基础少样本的准确率达到 *73.60%*；准确率比基础零样本提高近 *24%*，与组件更丰富的完整零样本相当。同样，将思维链（思维链零样本）和分类指令（*Categ-ZS*）添加到基础零样本后，准确率分别提高 *11.7%* 和 *12.7%*。

4.2 抽取准确率

表 6 报告了不同模型和提示方法之间完全匹配的总体准确率。

发现 2： 少样本提示通常能提高大多数字段的准确率，平均提升幅度最高可达 *17.34%*。然而，它并不能提升类别抽取的性能。在某些情况下，使用示例的模型甚至表现更差。

表 6：所选（元数据）字段的完全匹配准确率 [%]。对于每一行，mygreen =表示最佳（最高），myred = 表示最差（最低）。

AWS	GPT 3.5		GPT 4o		Claude 3.5		Claude 4		Gemini 2.0		Gemini 2.5	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few
服务名称	84.67	100.00	100.00	100.00	76.67	100.00	88.00	100.00	79.33	100.00	98.67	100.00
位置	48.00	96.67	83.33	98.67	48.67	96.67	55.33	97.33	76.67	96.67	78.00	96.00
开始时间	88.00	91.33	95.33	96.00	84.67	94.67	95.33	95.33	88.67	94.67	95.33	95.33
结束时间	83.33	86.00	87.33	88.00	66.00	86.00	85.33	86.67	83.33	88.67	86.67	88.00
时区	98.67	98.67	98.00	98.67	97.33	98.00	98.67	98.67	97.33	98.67	98.67	98.67
服务类别	73.33	73.33	90.00	68.00	85.33	87.33	88.00	89.33	84.00	86.00	90.67	90.67
平均值	79.33	91.00	92.33	91.56	76.44	93.78	85.11	94.56	84.89	94.11	91.34	94.78

Azure	GPT 3.5		GPT 4o		Claude 3.5		Claude 4		Gemini 2.0		Gemini 2.5	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few
服务名称	56.84	63.16	55.79	61.05	66.32	67.37	61.05	55.79	56.84	65.26	57.89	52.63
位置	63.16	67.37	65.26	64.21	64.21	69.47	66.32	70.53	52.63	67.37	60.00	65.26
开始时间	97.89	100.00	98.95	94.74	67.37	96.84	98.95	100.00	68.42	100.00	98.95	98.95
结束时间	92.63	96.84	94.74	93.68	65.26	93.68	93.68	96.84	64.21	93.68	95.79	96.84
时区	97.89	97.89	97.89	95.79	97.89	97.89	97.89	98.95	86.32	98.95	98.95	98.95
服务类别	64.21	58.95	67.37	67.37	63.16	63.16	64.21	66.32	60.00	65.26	62.11	62.11
根因类别	61.05	63.16	67.37	64.21	65.26	71.58	63.16	66.32	61.05	73.68	67.37	70.53
平均值	76.24	78.20	78.20	77.29	69.92	80.00	77.89	79.25	64.21	80.60	77.29	77.90

GCP	GPT 3.5		GPT 4o		Claude 3.5		Claude 4		Gemini 2.0		Gemini 2.5	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few
服务名称	83.72	89.77	85.12	86.05	73.02	88.37	83.72	87.91	59.53	89.77	79.07	89.30
开始时间	25.58	37.21	47.44	55.35	33.49	44.65	64.65	49.77	45.12	46.98	57.67	47.44
结束时间	32.09	44.19	78.60	84.19	66.98	80.47	85.12	86.98	73.95	77.67	85.12	88.37
时区	74.42	77.67	77.67	87.91	73.49	74.42	89.77	82.79	54.88	77.67	86.98	88.84
服务类别	61.86	53.02	61.86	40.93	64.19	56.74	62.33	61.86	62.33	67.91	62.33	64.19
平均值	55.53	60.37	70.14	70.89	62.23	68.93	77.12	73.86	59.16	72.00	74.23	75.63

总体而言，少样本提示在大多数情况下都能提高准确率。在我们的实验中，GCP 的模型-领域组合有 76.67%（23/30）取得提升，AWS 为 75.00%（27/36），Azure 为 64.29%（27/42）。然而，它对服务类别和根类别抽取的效果较差，某些情况下甚至会出现明显下降。例如，GPT-4o 在 AWS 上的服务类别准确率下降了 22.00%，在 GCP 上下降了 20.93%。一种可能的解释是，这些分类任务中的模式有限，因此在使用高级推理模型（如 GPT-4o）时，少样本示例可能导致过拟合。

发现3： 轻量级模型有时表现优于高级模型，尤其是在使用少样本示例时。

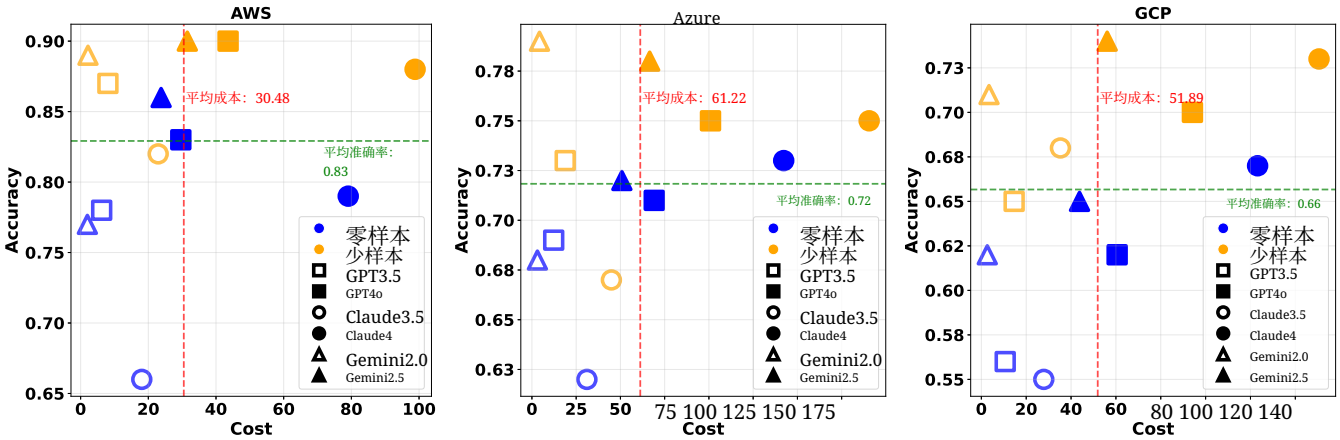
高级模型并不一定能确保更好的性能。例如，AZURE 的平均准确率最高值为 80.60%，由 Gemini 2.0 少样本提示取得；相比之下，Gemini 2.5 少样本提示的准确率为 77.90%。就平均准确率而言，对于 Azure，采用少样本学习的轻量级模型可能优于不采用少样本学习的高级模型。相反，对于 GCP，无论是否采用少样本提示，高级模型的准确率都高于通用模型。这表明，最佳模型与提示策略因数据集而异，因此仅选择单一模型或固定组合并不足够。

发现4： 少样本思维链的延迟更低，尽管它需要多出 *1.5–2×* 倍的输入词元，这可能是因为输出词元更少。

尽管少样本思维链提示词需要多出 1.5–2× 倍的输入词元，但其延迟可能更低，因为少样本示例引导的结构化输出减少了输出词元的数量，而输出词元是影响延迟的主要因素 [2]。例如，在 GCP 上，Claude 4 少样本方案耗时 10.00 秒，比零样本方案（11.08 秒）缩短了 1.08 秒。

发现5： 成本最高的模型比成本最低的模型贵 *50–60* 倍。

不同模型的平均成本差异显著，在各数据集中，成本最高的模型比成本最低的模型高出 50–60× 。例如，在 Azure 上，Claude 4 少样本的成本为 190.54 (10^{-4} \$)，比 Gemini 2.0 零样本和少样本分别高出 61.5× 和 44.6× ，而后者的成本分别仅为 3.10 (10^{-4} \$) 和 4.27 (10^{-4} \$)。然而，Gemini 2.0 少样本的准确率更高，延迟也低得多。这表明，Gemini 2.0 等轻量级模型能够以更低的成本和延迟实现有竞争力甚至更高的准确率，这凸显了



(a) AWS。 (b) Azure。 (c) GCP。

图 3：模型选择中准确率与成本之间的权衡。

这凸显了在实际应用中选择大语言模型时，考虑性能与成本权衡的重要性。

4.3 对模型选择的启示

发现 6： *Gemini 2.0* 和 *GPT 3.5* 等轻量级模型在各数据集中实现了准确率、成本和延迟之间的良好平衡。*Gemini 2.5* 和 *GPT-4o* 等高级模型能够达到更高的准确率，但成本和延迟也显著更高；而 *Claude* 则需要进一步优化（例如进行微调）才能有效使用。

图 3 展示了不同数据集、模型和提示词下准确率与成本之间的关系。位于左上角附近的模型表示其准确率更高、成本更低。绿色线表示平均准确率，红色线表示平均成本。总体而言，在使用大语言模型进行事件报告提取时，我们建议采用少样本提示词，并从 Gemini 2.0 和 GPT 3.5 等轻量级模型入手。如果需要更高的准确率，可考虑使用 Gemini 2.5 和 GPT-4o 等高级模型。Claude 系列模型仍需进一步优化，才能更好地应用于此任务。

5 有效性威胁

我们的工作展示了如何准确、有效地利用大语言模型从云事件报告中抽取信息。然而，我们的方法存在一些威胁和局限性：

(1) 模型的准确率 受所选少样本示例的数量和内容限制。未来需要进一步研究如何优化少样本提示词的设计，以提高准确率；

(2) 数据标注的质量 受特定字段细分方式的限制，例如服务类别、用户症状类别等。此外，这种"封闭世界"的分类方法可能无法捕捉不符合预设类别的复杂且不断演变的事件类型，从而可能遗漏关键细节。

(3) 当前事件分析的深度 受到公开事件报告可用性的限制。虽然报告呈现了许多数值结果并进行了可视化总结，但由于缺乏详细信息，深入的相关性分析或因果洞见仍然有限。例如，AWS 和 GCP 不会在其报告中披露详细的根因信息，这限制了我们分析的深度。

6 相关工作

事故管理与分析： 云事件管理与分析对于理解和提升云计算服务的可靠性至关重要，因为它们能提供有价值的见解，帮助开展根本原因检测和缓解 [1, 18, 21]。此前的研究涵盖了对云服务 [12, 25, 26] , 、大语言模型服务 [9, 31], 中停机和事件报告的实证特征分析，以及用于自动化此类分析的框架 [5]。然而，这些研究要么依赖人工或基于规则的信息抽取方法，要么仅关注事故元数据字段，而未深入分析用户影响和根因分类等内容。人工智能的兴起为解决这类问题提供了方案。已有研究探索了利用机器学习从事故记录中自动抽取结构化信息的方法，例如 SoftNER [24], 、贝叶斯网络 [8],和 CUA [25]。然而，系统探索大语言模型应用的研究仍然缺乏。

面向性能工程的人工智能： 人工智能可以为系统性能工程提供强有力的洞见 [13]。目前，两个主要应用领域尤为突出：用于日志解析的 LLM [3, 16], ，以及用于根因分析（RCA）的 LLM [1, 7, 21, 27, 28], ，后者已展现出良好的性能。例如，RCACopilot [7]将语言模型集成到诊断工作流中；TAMO [27] 则将日志、追踪数据和指标整合到一个工具辅助的大语言模型智能体中，以克服根因分析在上下文和模态方面的局限。然而，这些研究主要依赖遥测数据和内部监控数据，而非公开事件报告。据我们所知，我们是首个构建事件报告标注数据集、系统比较大语言模型及其提示词在结构化信息提取方面表现，并对多种事故特征开展纵向分析的研究。

7 结论与未来工作

准确、高效地从云计算事故报告中进行数据抽取与分析，对于提升云计算服务的可信赖性至关重要。为应对这一挑战，我们提出了一种

方法论，展示如何利用大语言模型进行结构化数据抽取。

在这项工作中，我们从三家云运营商处收集了 3,000 份事件报告，并从中标注了 460 份用于评估。我们提出了五种提示组件，并设计了六种策略。随后，我们使用轻量级和高级大语言模型，构建数据提取流水线，从文本事件报告中抽取十种类型的信息。之后，我们评估并比较了六种大语言模型的准确率、延迟和成本，为根据实际报告抽取中的不同需求选择提示词和模型提供参考。总体而言，我们总结了六项关键发现，并提供开放源代码的研究材料，供系统研究人员、云工程师和服务用户使用，以便更好地了解和改进云事件管理。

我们的未来工作包括：（1）优化提示词评估。还需进一步研究提示设计，并通过受控实验评估不同组件。（2）主动事件预测。抽取的信息可进一步用于预测事件持续时间和根本原因，从而更好地展现其在事件缓解中的下游效用。（3）高级大语言模型技术。可应用微调和检索增强生成（RAG）等更先进的大语言模型技术，利用历史模式进一步提高复杂字段的抽取准确率。

致谢

本研究部分由EU MSCA CloudStars (101086248)和Horizon Graph Massivizer (101093202)资助，并获得荷兰国家增长基金6G旗舰项目Future Network Services的支持。我们承认仅使用ChatGPT进行语法和清晰度方面的改进；本文内容为原创，完全由作者撰写。

参考文献

[1] Toufique Ahmed、Supriyo Ghosh、Chetan Bansal、Thomas Zimmermann、Xuchao Zhang和Saravan Rajmohan。2023年。使用大语言模型为云事件推荐根本原因和缓解步骤。载于第*45*届*IEEE/ACM*国际软件工程大会，*ICSE 2023*，澳大利亚墨尔本，*5*月 *14 20, 2023* 1737 1749 10 1109 48619 2023 00149 -。IEEE，–。 https://doi.org/。/ICSE。。OpenAI API。。延迟优化。https://platform.openai.com/docs/guides/latency-optimization，访问日期：2025-08-18。[3] Merve Astekin、Max Hort和Leon Moonen。2024年。大语言模型用于日志解析的比较研究。载于第*18*届*ACM/IEEE*实证软件工程与测量国际研讨会论文集（西班牙巴塞罗那）*(ESEM '24)*。计算机协会，美国纽约州纽约市，234 244 10 1145 3674805 3686684 –。https://doi.org/。/。Microsoft Azure。。Azure状态历史记录。https://azure.status.microsoft/en- us/status/history/，访问日期：2025-09-09。[5] Sándor Battaglini-Fischer、Nishanthi Srinivasan、Bálint László Szarvas、Xiaoyu Chu和Alexandru Iosup。2025年。FAILS：大语言模型服务事件自动收集与分析框架。载于第*16*届*ACM/SPEC*性能工程国际会议伴随会议（加拿大安大略省多伦多）*(ICPE '25)*。计算机协会，美国纽约州纽约市，187–194。https: //doi.org/ 10 1145 3680256 3721320 。/。Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever和Dario Amodei。2020年。语言模型是少样本学习者。载于神经信息处理系统进展*33*：神经信息处理系统年会*2020*，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，Hugo Larochelle、Marc'Aurelio Ranzato、Raia Hadsell、Maria-Florina Balcan和Hsuan-Tien Lin（编）。 https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[7] Yinfang Chen、Huaibing Xie、Minghua Ma、Yu Kang、Xin Gao、Liu Shi、Yunjie Cao、Xuedong Gao、Hao Fan、Ming Wen、Jun Zeng、Supriyo Ghosh、Xuchao Zhang、Chaoyun Zhang、Qingwei Lin、Saravan Rajmohan、Dongmei Zhang 和 Tianyin Xu。2024年。基于大语言模型的云事件自动根因分析。载于《第十九届欧洲计算机系统会议论文集》，EuroSys 2024，希腊雅典，2024年4月22日至25日。ACM，674–688。https://doi.org/10.1145/3627703.3629553[8] Yujun Chen、Xian Yang、Qingwei Lin、Hongyu Zhang、Feng Gao、Zhangwei Xu、Yingnong Dang、Dongmei Zhang、Hang Dong、Yong Xu、Hao Li 和 Yu Kang。2019年。云服务系统的宕机预测与诊断。载于《万维网会议论文集》，WWW 2019，美国加利福尼亚州旧金山，2019年5月13日至17日，Ling Liu、Ryen W. White、Amin Mantrach、Fabrizio Silvestri、Julian J. McAuley、Ricardo Baeza-Yates 和 Leila Zia（编）。ACM，2659–2665。https: //doi.org/ 10 1145 3308558 3313501 . / . Xiaoyu Chu、Sacheendra Talluri、Qingxian Lu 和[9] Alexandru Iosup。2025年。面向大语言模型的公共服务宕机与事件实证特征分析（ICPE '25）。计算机协会，美国纽约州纽约市， 69 80 10 1145 3676151 3719372 – 。https://doi.org/ . / . Google Cloud[10]。. Google Cloud 服务健康状态。https://status.cloud.google. com/summary，访问日期：2025-09-09。[11] Drishti Goel、Fiza Husain、Aditya Singh、Supriyo Ghosh、Anjaly Parayil、Chetan Bansal、Xuchao Zhang 和 Saravan Rajmohan。2024年。使用大语言模型进行云事件管理的X生命周期学习。载于《第32届ACM软件工程基础国际会议论文集》，FSE 2024，巴西 Porto de Galinhas，2024年7月15日至19日，Marcelo d' Amorim（编）。ACM， 417 428 10 1145 3663529 3663861 – 。 https://doi.org/ . / . Haryadi S. Gunawi[12]、Mingzhe Hao、Riza O. Suminto、Agung Laksono、Anang D. Satria、Jeffry Adityatama 和 Kurnia J. Eliazar。2016年。云为何停止计算？数百起服务宕机事件的经验教训。载于《第七届ACM云计算研讨会论文集》，美国加利福尼亚州圣克拉拉，2016年10月5日至7日，1 16 10 1145 2987550 2987583。ACM，– 。 https://doi.org/ . / .[13] 2025 Lizy K. John。. 面向性能工程的人工智能与面向人工智能的性能工程。载于《第16届ACM/SPEC性能工程国际会议论文集》（加拿大安大略省多伦多）（ICPE '25）。计算机协会，美国纽约州纽约市，1–2。https://doi.org/10.1145/3676151.3720528[14] Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yutaka Matsuo 和 Yusuke Iwasawa。2022年。大语言模型是零样本推理器。载于《神经信息处理系统进展》第35卷：神经信息处理系统年度会议 2022，NeurIPS 2022，美国路易斯安那州新奥尔良，2022年11月28日至12月9日。 http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html[15] Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。2023年。预训练、提示与预测：自然语言处理提示方法的系统综述。《ACM计算调查》55卷，9期，文章195（2023年1月），35页。https://doi.org/10.1145/3560815[16] Yilun Liu、Shimin Tao、Weibin Meng、Jingyu Wang、Wenbing Ma、Yuhang Chen、Yanqing Zhao、Hao Yang 和 Yanfei Jiang。2024年。使用提示策略和大语言模型进行可解释的在线日志分析。载于《第32届IEEE/ACM程序理解国际会议论文集》（葡萄牙里斯本）（ICPC '24）。计算机协会，美国纽约州纽约市， 35 46 10 1145 3643916 3644408 – 。 https://doi.org/ . / . Christopher D. Manning[17]、Prabhakar Raghavan 和 Hinrich Schütze。2008年。《信息检索导论》。剑桥大学出版社，美国。[18] Victor Ion Munteanu、Andrew Edmonds、Thomas Michael Bohnert 和 Teodor-Florin Fortis。2014年。云事件管理：挑战、研究方向与架构方法。载于2014年IEEE/ACM第七届效用与云计算国际会议。 786 791 10 1109 2014 128 – 。https://doi.org/ . /UCC. . Alec Radford[19] 等。2019. 语言模型是无监督多任务学习器。https://cdn.openai.com/better-language-models/language_models_ are_unsupervised_multitask_learners.pdf[20] Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev 和 Percy Liang。2016年。SQuAD：用于文本机器理解的100,000+ 个问题。载于《2016年自然语言处理实证方法会议论文集》。计算语言学协会，美国得克萨斯州奥斯汀，2383–2392。 https://doi.org/ 10 18653 1 16 1264 . /v /D - Devjeet Roy、Xuchao Zhang、Rashi Bhave[21]、Chetan Bansal、Pedro Henrique B. Las-Casas、Rodrigo Fonseca 和 Saravan Rajmohan。2024年。探索用于根因分析的基于大语言模型的智能体。载于《第32届ACM软件工程基础国际会议论文集》（巴西 Porto de Galinhas）（FSE 2024）。计算机协会，美国纽约州纽约市， 208 219 10 1145 3663529 3663841 – 。https://doi.org/ . / . Amrita Saha[22] 和 Steven C. H. Hoi。2022. 从云服务事件调查中挖掘根因知识以实现智能运维。载于《第44届国际软件工程会议论文集：软件工程实践》（宾夕法尼亚州匹兹堡）（ICSE-SEIP '22）。计算机协会，美国纽约州纽约市， 197 206 10 1145 3510457 3513030 – 。https://doi.org/ . / . Amazon Web Services[23]。. AWS 服务健康状态。https://health.aws.amazon.com/health/status，访问日期：2025-09-09。 2025

[24] Manish Shetty、Chetan Bansal、Sumit Kumar、Nikitha Rao、Nachiappan Nagappan 和 Thomas Zimmermann。2021年。从云服务事件中提取神经知识。载于第*43*届*IEEE/ACM*软件工程国际会议：软件工程实践，*ICSE*（*SEIP*）*2021*，西班牙马德里，*2021*年*5*月*25*日至*28*日。IEEE，218–227。https://doi.org/10.1109/ICSE-SEIP52600.2021.00031[25] Sacheendra Talluri、Dante Niewenhuis、Xiaoyu Chu、Jakob Kyselica、Mehmet Cetin、Alexander Balgavy 和 Alexandru Iosup。2025年。云正常运行时间档案：网络、云和游戏服务的开放可用性数据。*CoRR* abs/2504.09476（2025年）。https://doi.org/10.48550/ARXIV.2504.09476 arXiv:2504.09476[26] Sacheendra Talluri、Leon Overweel、Laurens Versluis、Animesh Trivedi 和 Alexandru Iosup。2021年。云故障用户报告的实证特征分析。载于*2021*年*IEEE*自主计算与自组织系统国际会议（*ACSOS*）。158–163。https://doi.org/10.1109/ACSOS52086.2021.00039[27] Qi Wang、Xiao Zhang、Mingyi Li、Yuan Yuan、Mengbai Xiao、Fuzhen Zhuang 和 Dongxiao Yu。2025年。TAMO：借助工具辅助的大语言模型智能体和多模态观测数据进行细粒度根因分析。*CoRR* abs/2504.20462（2025年）。https://doi.org/10.48550/ARXIV.2504.20462 arXiv:2504.20462[28] Zefan Wang、Zichuan Liu、Yingying Zhang、Aoxiao Zhong、Jihong Wang、Fengbin Yin、Lunting Fan、Lingfei Wu 和 Qingsong Wen。2024年。RCAgent：利用自主智能体和工具增强型大语言模型进行云根因分析。载于第*33*届*ACM*信息检索国际会议论文集与知识管理，*CIKM 2024*，美国爱达荷州博伊西，*2024*年*10*月*21*日至*25*日，Edoardo Serra 和 Francesca Spezzano（编）。ACM，4966–4974。https://doi.org/10.1145/3627673.3680016[29] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。2022年。思维链提示能够激发大语言模型的推理能力。载于 神经信息处理系统进展第*35*卷：神经信息处理系统*2022*年度会议，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月9日。http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html[30] Derong Xu、Wei Chen、Wenjun Peng、Chao Zhang、Tong Xu、Xiangyu Zhao、Xian Wu、Yefeng Zheng、Yang Wang 和 Enhong Chen。2023年。用于生成式信息抽取的大语言模型：综述。*CoRR* abs/2312.17617（2023年）。https://doi.org/10.48550/ARXIV.2312.17617 arXiv:2312.17617[31] Haoran Yan、Yinfang Chen、Minghua Ma、Ming Wen、Shan Lu、Shenglin Zhang、Tianyin Xu、Rujia Wang、Chetan Bansal、Saravan Rajmohan、Chaoyun Zhang 和 Dongmei Zhang。2025年。生成式人工智能云服务中生产环境事件的实证研究。*CoRR* abs/2504.08865（2025年）。https://doi.org/10.48550/ARXIV. 2504.08865 arXiv:2504.08865[32] Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q. Weinberger 和 Yoav Artzi。2020年。BERTScore：使用 BERT 评估文本生成。载于 第*8*届国际学习表征会议，*ICLR 2020*，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。https://openreview.net/forum?id=SkeHuCVFDr