一种生成式 AI 驱动的声称检索系统，能够从多语言的社交媒体平台中检测和检索声称

Ivan Vykopal[1,2], **Martin Hyben**[2], **Robert Moro**[2], **Michal Gregor**[2] and **Jakub Simko**[2]
[1] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[2] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
{ name.surname } @kinit.sk

Abstract

在线虚假信息构成全球性挑战，对事实核查者提出了重大要求，他们必须高效验证声明以防止虚假信息传播。此过程中一个主要问题是对已经核查过的声明进行重复验证，这增加了工作量并延缓了对新出现声明的响应。本研究引入了一种方法，该方法检索先前已核查的声明，评估其与给定输入的相关性，并提供补充信息以支持事实核查者。我们的方法使用大型语言模型（LLMs）来过滤无关的事实核查，并生成简明的总结和解释，使事实核查者能够更快地评估声明是否已被验证。此外，我们通过自动和人工评估来评估我们的方法，其中人类与开发的工具交互，审查其有效性。我们的结果表明，LLMs 能够过滤掉许多无关的事实核查，从而减少工作量并简化事实核查过程。

1 引言

社交媒体的兴起加速了虚假信息的传播，带来了显著的社会、经济和公共健康风险 (Zubiaga et al., 2018) 。这一挑战进一步因虚假信息的多语言性而加剧，使得事实核查成为一个复杂且资源密集的任务。事实核查员常常难以在多个语言中验证声明，尤其是在资源匮乏的环境中，此类环境中存在有限的事实核查支持 (Hrckova et al., 2024) 。为了解决这个问题，开发多语言事实核查方法至关重要，这样可以帮助事实核查员有效识别和验证虚假信息。

事实核查中的一个关键任务是声明检索，也被称为先前核查声明检索 (Pikuliak et al., 2023) ，其目标是从数据库中识别出与给定输入最相似的事实核查记录。这个任务至关重要，因为许多声明并非完全新颖，而是对先前被揭穿的错误信息的改述或重复。高效的检索可以使事实核查员迅速检测到重复的声明，减少重复劳动，并优先处理新出现或复杂的声明 (Hrckova et al., 2024) 。然而，检索结果可能包括仅松散相关或无关的事实核查，增加了工作量。为减轻这一问题，可以利用 LLM 来评估检索到的事实核查的相关性，从而简化审查流程 (Vykopal et al., 2025) 。

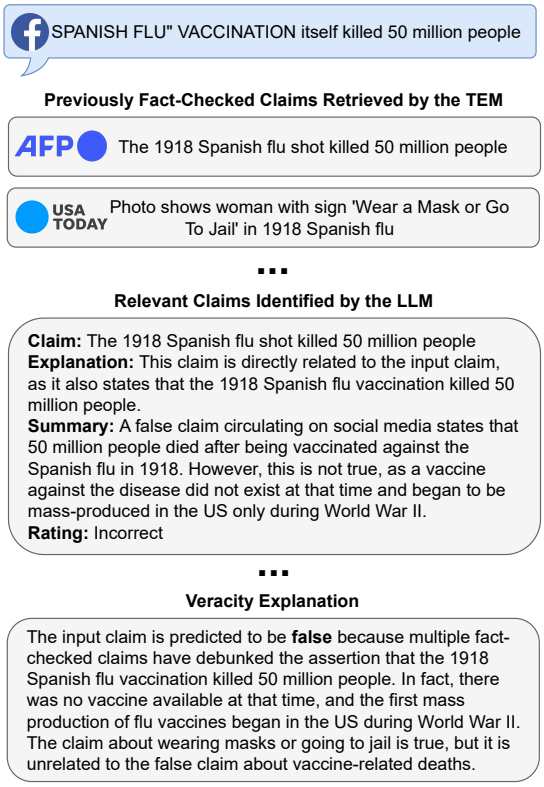

Figure 1: 这是一个帖子实例，其中包含两个由嵌入模型检索的事实核查声明。LLM 选择相关的声明，解释其选择，总结事实核查文章，并预测帖子的真实性。

在本文中，我们提出了一种新的流程，用于检索之前已核实的事实主张，并协助事实核查员评估其与给定查询的相关性。我们的实验涵盖了来自不同语言家族和文字系统的 10 多种语言，包括资源丰富和资源匮乏的语言。我们分析了语言模型在结合总结和解释的同时检索相关事实核查的能力。图 1 中显示了一个示例。此外，我们评估了大型语言模型（LLMs）在基于检索到的事实核查和支持信息来确定真

实性方面的性能。[1]

我们的贡献如下：

- 我们提供了一个新的 AFP-Sum 数据集，包括大约 19K 篇事实核查文章及其摘要，涵盖 23 种语言。此外，我们创建了一个包含 2300 篇事实核查的小型数据集，这些事实核查用 23 种语言书写，并包含原始语言的摘要和翻译成英文的摘要。

- 我们评估了多种文本嵌入模型（TEMs）在 20 种语言中检索先前事实核查过的声明的效果，以及 TEMs 根据自然语言指令过滤事实核查的能力。

- 我们提出了一种新的流程，将 LLMs 融入到验证过程中，通过利用 LLMs 识别相关的先前事实核查的声明，提供事实核查摘要，并基于先前检索到的事实核查预测给定声明的真实性。

2 相关工作

先前核实的声明检索。 之前的事实核查声明检索，也被称为已验证声明检索（Barrón-Cedeño et al., 2020）或声明匹配（Kazemi et al., 2021），旨在通过检索相似的、已验证的声明来减少事实核查员的工作量。尽管大多数研究集中在单语环境中（Shaar et al., 2020, 2022; Hardalov et al., 2022），但多语言检索仍然未得到充分探索（Vykopal et al., 2024）。最近的工作，例如 Pikuliak et al. (2023)，引入了用于多语言声明检索的 MultiClaim 数据集，在单语和跨语言背景下评估各种 TEMs 的已核查声明排序。

近年来，LLMs 的进步为增强声明检索提供了新的机会。现有方法主要依赖于两种策略。第一种涉及文本蕴含，其中模型将输入声明与事实核对之间的蕴含分类为三种类别（Choi and Ferrara, 2024a,b）。相反，第二种策略采用生成式重排序，根据 LLMs 生成的条件概率对之前已核对的声明进行排序，从而优先考虑更相关的声明（Shliselberg and Dori-Hacohen, 2022; Neumann et al., 2023）。

事实核查流程 & 工具。 随着在线事实核查的重要性日益增长，已经开发了许多管道来对抗错误信息。其中许多系统依赖于根据给定的声明检索证据，并利用 LLMs 来评估真实性并提供理由。然而，大多数研究主要集中在英语（Hassan et al., 2017; Shu et al., 2019; Li et al., 2024）或阿拉伯语（Jaradat et al., 2018; Althabiti et al., 2024; Sheikh Ali et al., 2023），而较少有针对其他语言的研究。

已经开发了多种在线工具来处理虚假信息。WeVerify [2] 提供了一套识别虚假信息的工具，包括用于检测被操纵内容的图像分析。此外，BRENDA（Botnevik et al., 2020）评估声明的可信度，帮助用户评估在线信息。此外，由 Pikuliak et al. (2023) 训练的检索先前事实核查过的声明的模型已被整合到 Fact-Check Finder [3] 中，这是一种旨在协助事实核查员识别跨多种语言的相关事实核查声明的工具。

多语言摘要 多语种摘要研究得到了广泛数据集的发展和 LLM 的应用推动（Scialom et al., 2020; Hasan et al., 2021; Bhattacharjee et al., 2023）。这些资源已经使得多语种模型如 mT5（Xue et al., 2021）的微调成为可能，这些模型在多语种和低资源摘要任务中表现出竞争力。此外，研究探讨了 LLM 如 GPT-3.5 和 GPT-4 在跨语言摘要中的零样本和少样本能力，突出了它们在无需广泛微调的情况下处理多种语言对的潜力（Wang et al., 2023）。为了增强多语种摘要中的事实一致性，也进行了努力，例证为使用多语种模型来提高机器生成摘要在不同语言中的可靠性（Aharoni et al., 2023）。

3 方法论

我们的实验旨在评估 TEM 和 LLM 在通过提供附加信息来协助事实核查员的能力。这包括检索之前被核查过的最相似的声明，总结事实核查文章及其评级，并可能基于检索到的信息预测给定输入的真实性。许多此类过程可以使用 LLM 自动化，从而减少事实核查员在识别相关事实核查时所需的努力。

我们提出的流程，如图 2 所示，包括四个关键步骤：检索（第 4 节）、过滤（第 5 节）、总结（第 6 节）和真实性预测（第 7 节）。在检索步骤中，TEM 基于给定的输入检索出最相似的前 K 个事实核查。在过滤步骤中，通过使用 LLM 来筛选出仅与输入直接相关的事实核查，同时提供选择的解释并排除无关的声明。在总结步骤中，LLM 生成相关事实核查文章的简洁摘要。最后，在真实性预测步骤中，LLM 利用检索到的事实核查、它们的评级以及生成的摘要，根据可用的信息来评估和预测给定输入的真实性。

[1] 数据可在 Zenodo 上按请求获取，仅用于研究目的：https://zenodo.org/records/15267292 。源代码可在以下地址获取：https://github.com/kinit-sk/claim-retrieval 。

[2] https://weverify.eu/

[3] https://fact-check-finder.kinit.sk/

此外，我们提供了实验中使用的数据集（见第 3.1 节）和模型（见第 3.2 节）的概述。我们还在第 3.3 节详细介绍了管道每个步骤的评估。

3.1 数据集

多重索赔 我们选择了 MultiClaim 数据集 (Pikuliak et al., 2023) 来评估嵌入模型和大型语言模型在检索已核实的声明和评估声明真实性方面的效率。MultiClaim 包括 206K 篇事实核查文章，涵盖 39 种语言，以及 27 种语言中的 28K 条社交媒体帖子。此外，该数据集还包含 31K 对社交媒体帖子和事实核查文章的配对，这些配对将帖子与相应的事实核查文章连接起来。而且，每篇事实核查文章都被分配了真实性评级，并包含一个 URL，可以检索到完整的文章内容。这些补充信息通过使我们能够对检测已核实声明进行更结构化和全面的评估，增强了我们的工作流程。

AFP-总和。 为了评估大型语言模型汇总事实核查文章的能力，我们创建了 AFP-Sum 数据集，该数据集由来自 AFP（法新社）的事实核查文章及其摘要组成 [4] 。我们收集了 23 种语言的事实核查文章，总计约 19K 篇由事实核查人员撰写摘要的文章。在我们的实验中，我们为每种语言选择了 100 篇事实核查文章，评估大型语言模型生成的英文摘要。为了便于评估，我们使用谷歌翻译将所有参考摘要翻译成英文。附录 B.2 中的表格 8 展示了数据集的统计信息。

3.2 语言模型

我们使用了两类模型，特别是文本嵌入模型和大型语言模型。文本嵌入模型用于检索阶段，以识别给定输入的最相关事实核查。虽然存在众多的文本嵌入模型，我们选择了英文和多语言模型，并使用 BM25 作为基准进行比较。我们研究中使用的文本嵌入模型列在表格 2 中。

除了 TEMs，我们还评估了一组不同的 LLMs，包括闭源和开源的，依据它们在 NLP 任务中的最新性能进行选择。在摘要方面，我们也尝试了参数少于 30 亿的小型 LLMs。我们实验中使用的 LLMs 的完整列表如表 1 所示。

3.3 评价

我们采用了针对我们提出的流程不同阶段量身定制的各种评估指标。

在检索步骤中，我们使用成功率 @K（S@K）作为评估 TEM 性能的主要评价指标。S@K 衡量在前 K 个检索结果中出现正确事实核查的情况百分比。此外，我们应用该指标评估 LLM 从 TEM 检索集合中识别最相关的事实核查的能力。

在摘要实验中，我们使用了两个标准指标：BERTScore 和 ROUGE-L。BERTScore 通过基于 BERT 模型的上下文词嵌入计算 F1 分数来评估语义相似性。而 ROUGE 则测量生成摘要与参考摘要之间的 n-gram 重叠。具体来说，我们采用了 ROUGE-L，该指标侧重于最长的公共单词序列。ROUGE-L 还帮助检测生成的摘要中出现非英语语言的情况，这比使用 BERTScore 来识别更具挑战性。

最后，对于真实性预测实验，我们采用了不平衡数据的标准分类指标：Macro F1 得分、精确度和召回率。

Model	# Params [B]	# Langs	Organization	Citation
`GPT-4o (2024-08-06)`	N/A	N/A	OpenAI	
`Claude 3.5 Sonnet`	N/A	N/A	Anthropic	
`Mistral Large`	123	11	Mistral AI	Mistral AI Team (2024)
`C4AI Command R+`	104	23	Cohere For AI	Cohere For AI (2024)
`Qwen2 Instruct`	72	29	Alibaba	Yang et al. (2024)
`Qwen2.5 Instruct`	0.5, 1.5, 3, 72	29	Alibaba	Yang et al. (2024)
`Llama3.1 Instruct`	70	8	Meta	Grattafiori et al. (2024)
`Llama3.2 Instruct`	1, 3	8	Meta	Grattafiori et al. (2024)
`Llama3.3 Instruct`	70	8	Meta	Grattafiori et al. (2024)
`Gemma3`	27	140	Google	Team et al. (2025)

Table 1: 我们实验中使用的大型语言模型，包括闭源模型和开源模型。

4 检索实验

在本节中，我们描述了在两种设置下对各种 TEMs 进行的实验。首先是直接检索（第 4.1 节），在这种设置中，我们评估现有 TEMs 在基于帖子检索最相似的已事实核查声明时的性能。其次是基于标准的检索（第 4.2 节），在这种设置中，我们评估 TEMs 在根据英语自然语言中指定的标准（例如特定命名实体的存在、发布日期等）过滤结果时的表现。

4.1 直接检索

我们评估了各种 TEM，以及它们在基于给定输入对之前已事实核查的声明进行排名的性能。我们将任务公式化为一个排名问题，目的是基于余弦相似性对数据库中给定输入的所有事实核查进行排名。我们选择了至少有 100 个帖子语言的 20 种语言，设置类似于 Pikuliak et al. (2023) 所提议的。除了在 (Pikuliak et al., 2023) 中评估的 TEM，我们还包括了更新的多语言 TEM，特别是各种大小的多语言 E5 模型。我们仅在单语言环境中评估 TEM 的性能，其中事实核查的声明和帖子为同一种语言。

结果。 在单语环境中的 TEMs 结果如表 2 所示。这些结果表明，一些多语言的 TEMs 可以

[4] https://www.afp.com

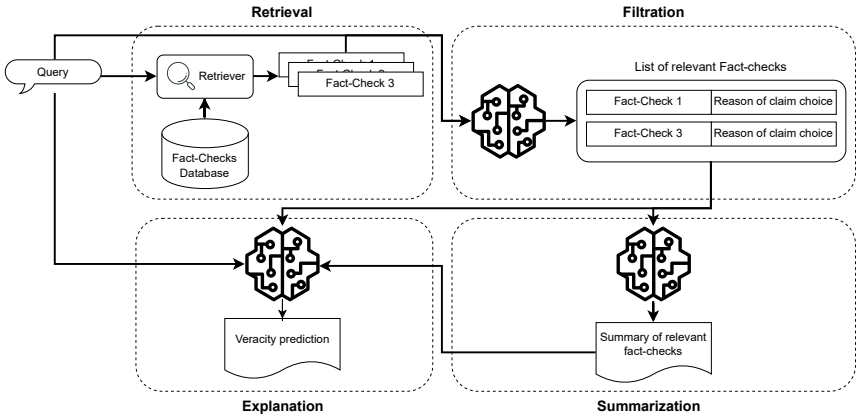

Figure 2: 我们提出的流程包括：(1) 检索最相似的前 N 个事实核查项，(2) 确定相关的已核查的声明，(3) 摘要相关的事实核查文章，以及 (4) 预测查询的真实性并提供解释。

Model	Size [M]	Ver.	Avg. S@10
BM25		Og	0.62
English TEMs			
DistilRoBERTa	82	En	0.75
MiniLM-L6	22	En	0.79
MiniLM-L12	33	En	0.79
MPNet-Base	109	En	0.77
GTE-Large-En	434	En	0.80
GTR-T5-Large	737	En	0.83
Multilingual TEMs			
BGE-M3	568	Og	0.82
DistilUSE-Base-Multilingual	134	Og	0.66
LaBSE	470	Og	0.69
Multilingual E5 Small	117	Og	0.78
Multilingual E5 Base	278	Og	0.78
Multilingual E5 Large	559	Og	0.84
MiniLM-L12-Multilingual	117	Og	0.63
MPNet-Base-Multilingual	278	Og	0.69

Table 2: 使用 S@10 评估指标在单语环境中对英语和多语种 TEMs 的平均表现。Ver. 表示数据集的原始版本（Og）或英文版本（En）。两种版本的最佳结果用粗体显示。

在性能上超过英文翻译加英文 TEMs 的组合，但差异并不具有统计学意义。Multilingual E5 Large 在 S@10 上取得了最高的成绩（具有统计学意义；$p < 0.05$），而 GTR-T5-Large 在英文翻译中取得了最佳结果（$p < 0.05$）。其他多语言 TEMs 在表现上落后于英文 TEMs。

表 9 显示了所有研究的 TEMs 在 20 种语言上的结果。基于英语 TEMs 的结果，GTR-T5-Large 在大多数语言上取得了卓越的表现（$p < 0.05$）。然而，对于德语，结果低于 0.70。另一方面，Multilingual E5 Large 在所有语言上都证明是有效的（$p < 0.05$），除了在泰语中，较小的 Multilingual E5 超过了更大的版本。

4.2 基于标准的检索

除了基于输入声明或帖子的检索之外，我们还尝试了基于标准的检索，其中我们使用 TEMs 根据给定的标准过滤结果，例如，特定命名实体的存在。目的是评估 TEMs 是否可以通过事实核查者提供的自然语言指令来过滤结果。我们为实验定义了四种设置：基于语言、日期范围、事实核查领域或命名实体的过滤。我们选择了性能最好的 TEM – Multilingual E5 Large 进行实验。我们提出了如图 5 所示的模板，包括经过事实核查的声明和元数据，如语言、事实核查组织和发布日期。

作为真实数据，我们使用了通过 Multilingual E5 Large 获得的结果对已经基于特定条件过滤的数据子集进行排名，使用的是手动设计的过滤器（例如，只对西班牙语事实核查进行排名）。我们的基于标准的检索管道包括两个步骤：(1) 基于标准进行检索（例如，特定语言），我们仅选择相似度得分超过 0.8 的事实核查；(2) 基于帖子进行排名，在其中我们使用帖子内容对之前检索到的结果进行排名，与直接检索相似。

在评估中，我们采用了斯皮尔曼等级相关系数和肯德尔 Tau 来评估 TEMs 使用自然语言查询对结果进行排序的能力。我们计算了 Multilingual E5 Large 使用我们的两步方法产生的排序与在已根据人工设计的过滤器过滤后的结果上使用 Multilingual E5 Large 获得的排序之间的相关性。

结果。 表格 3 展示了在四种设置下进行过滤检索的结果：命名实体、语言、事实核查域和日期范围。我们计算了斯皮尔曼等级相关系数的平均值、肯德尔的陶相关系数以及预测与参考核查列表之间常见事实核查的比例。正相关表示预测排序与参考排序一致，而负相关则表

Settings	Avg. Spearman	Avg. Kendall's Tau	Avg. Common FCs
Named Entities	−0.31	−0.20	0.32
Languages	−0.58	−0.43	0.17
Domains	−0.66	−0.51	0.12
Dates	−0.82	−0.64	0.02

Table 3: 平均斯皮尔曼相关系数、肯德尔 Tau 以及真实排序列表与预测排序列表之间的公共事实核查（FCs）比例的得分。我们报告所有设置的平均得分，每个类别至少包含 100 个事实核查。

Model	S@10 ↑	MRR ↑	Macro F1 ↑	TNR ↑	FNR ↓
`Multilingual E5 Large`	0.76	0.58	54.75	86.27	25.59
`Mistral Large 123B`	0.70	0.40	59.82	90.23	15.38
`C4AI Command R+`	0.66	0.35	55.50	85.83	15.38
`Qwen2.5 72B`	0.57	0.32	58.37	90.81	30.65
`Llama3.3 70B`	0.67	0.38	59.96	90.82	19.61
`Llama3.1 70B`	0.63	0.37	59.62	91.08	24.25
`Gemma3 27B`	0.65	0.35	57.77	89.14	21.78
`Llama3.1 8B`	0.60	0.24	52.38	82.30	21.16
`Qwen2.5 7B`	0.47	0.35	59.25	93.20	43.86

Table 4: 在 10 种语言的 100 篇帖子上的检索和过滤性能结果。`Multilingual E5 Large` 作为基准。最好的结果以粗体显示，次好的结果是 underlined 。

明它们之间存在反向关系。

我们的研究结果显示，基于命名实体的筛选产生了预测与真实事实核查列表之间的最高重叠（$p < 0.05$），这表明当事实核查是基于命名实体检索时，TEMs 表现最好。尽管如此，−0.31 的 Spearman 相关性表明，尽管 TEMs 可能识别出相关的事实核查，其排序与真实排序并不完全匹配。

按语言、领域和日期范围过滤会导致性能下降，其中日期范围的过滤效果最差。这表明，虽然 TEMs 可以根据自然语言指令检索相关的事实核查，但它们的过滤改变了候选集，从而限制并降低了整体排名性能。此外，我们指定了一个日期范围，而事实核查的嵌入仅包含每个事实核查的确切日期。这种差异使得 TEMs 更难根据提示中未明确提及的日期来检索事实核查。

5 过滤实验

为了筛除无关的已验证事实的声明，我们在 MultiClaim 数据集的一个子集上进行了几种 LLM 的实验。我们选择了 10 种语言，具体为捷克语、英语、法语、德语、印地语、匈牙利语、波兰语、葡萄牙语、西班牙语和斯洛伐克语，每种语言 100 个帖子。这些帖子是根据其真实性标签选择的。然而，由于 MultiClaim 数据集主要包含虚假帖子，实现平衡分布是不现实的。最终的数据集由 55 个真实帖子、65 个不可验证帖子和 880 个虚假帖子组成，导致明显的不平衡。我们使用这些数据来评估 LLM 在过滤掉给定输入的无关事实核查中的效率。

我们的方法包括两个步骤。首先，我们使用 `Multilingual E5 Large` 检索出 50 个最相似的事实核查声明。然后，我们指示 LLM 筛选这一集合（参见图 7），仅选择那些与输入帖子直接相关的声明，同时删除不相关的事实核查。

为了评估 LLMs 在此任务中的性能和效率，我们计算了用于检索的 S@10 和 MRR（平均倒数排名）分数。另外，我们还计算了宏 F1、真负率（TNR）和假负率（FNR）来识别 LLMs 的能力。为了计算分类指标，我们创建了由 `Multilingual E5 Large` 模型识别的帖子和事实核查对，这里的相关性标签是从 MultiClaim 数据集中的标记对中获得的。TNR 表示正确过滤掉多少无关事实核查的比例，而 FNR 表示错误过滤掉多少相关事实核查的比例。在这种情况下，我们希望最大化 TNR 并最小化 FNR。

5.1 结果

表格 4 总结了我们在过滤无关核实信息方面的结果。以 `Multilingual E5 Large` 为基准，我们在前 10 个结果中正确检索了 76 % 的相关核实信息。为了进一步评估性能，我们将排序任务框架化为二元分类，使用 Youden's Index 选择最佳阈值。宏观 F1 显示基准优于 `Llama3.1 8B` 。

在检索之后，我们应用了一个大型语言模型来筛选出前 50 个检索到的事实核查。虽然与基线相比，这降低了 S@10 和 MRR 得分，但目的是减少呈现给事实核查员的无关事实核查数量。我们测量了移除的相关和不相关事实核查的比例。`Mistral Large` 在 TNR 和 FNR 之间实现了最佳平衡（$p < 0.05$），同时在 S@10 上也优于其他大型语言模型。

我们的研究结果表明，尽管 LLMs 能够有效地去除不相关的事实核查，但它们可能会排除一些相关的。`Multilingual E5 Large` 和 LLMs 之间的性能差距表明，相关的事实核查有时会被误分类为不相关的，尽管 LLMs 也可能将排名较低的事实核查提升到前 10 名。

6 摘要评估

我们在 23 种语言的 AFP-Sum 数据集子集中评估了 LLM 在摘要事实核查文章方面的表现。实验在两种设置下进行：（1）文章优先--在指令之前提供文章；（2）文章最后--在指令之后提供文章（见图 6）。我们检查了提示顺序和量化如何影响摘要质量。文章以其原始语言提供，并指示生成英文摘要。生成的摘要与使用谷歌翻译的参考摘要的英文翻译进行了比较。

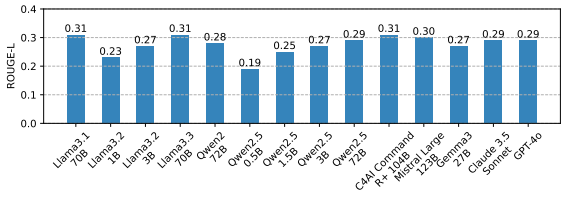

Figure 3: LLMs 在事实核查摘要中的整体表现。我们报道了每个 LLM 使用 “Article first” 设置的平均 ROUGE-L 得分，该设置中文章是在指令之前提供的。

6.1 结果

图 3 显示了在文章优先设置下使用 ROUGE-L 指标的整体结果。结果显示出不同的 LLM 性能差异。较小的 Llama 模型在摘要生成方面表现不佳，常常生成文章原语言的摘要而非英文，从而导致较低的 ROUGE-L。相比之下，其他 LLM 更好地遵循了生成英文摘要的指令。此外，在文章前提供指令加剧了这一问题，并导致 ROUGE-L 非常低（见表 13 ）。

表 12 比较了不同文章顺序设置和三个量化水平下的 Llama3.2 模型（1B 和 3B 参数）。采用 4 位、8 位和全精度模型生成摘要。结果显示，在指令之前提供文章显著提高了性能（$p < 0.05$ ），在指令之后提供文章时产生了更好的结果。对于 Llama3.2 1B ，8 位模型通常表现最佳（没有统计显著性），紧随其后的是全精度。在 Llama3.2 1B 中，全精度和 4 位之间的性能差距约为 0.3 BERTScore 分，而在 Llama3.2 3B 中仅为 0.1，表明量化对较大 LLMs 的影响较小。

总体而言，Llama3.3 70B 和 Mistral Large 在不同语言中的表现最佳（见表 11 ），而其他 LLMs 相对滞后（无统计显著性）。结果表明，覆盖较少语言的 LLMs（例如，Llama ）可以比覆盖更广泛语言的多语言 LLMs 表现得更好，例如，C4AI Command R+ 或 Qwen2.5 。

7 LLM 真实性预测的评估

为了评估大型语言模型（LLMs）在使用检索到的先前经过事实核查的声明和事实核查摘要时预测声明真实性的效果（见图 7 ），我们使用了与第 5 节中相同的数据。最终数据集包含三类：真实、虚假或无法验证，这些类别是不平衡的。因此，我们利用宏观 F1、宏观精确度和宏观召回率来评估 LLMs 的性能。在这种情况下，补充信息是英文的，尤其是摘要、评级和经过事实核查的声明。这对于人类事实核查员理解 LLMs 提供的结果也有益。作为基线，我们选择了 Mistral Large 和 Llama3.3 ，这些模型仅通过帖子和任务描述进行指导，没有额外的信息。

Model	Macro F1	Macro Precision	Macro Recall
Baseline (without retrieved fact-checks)			
Mistral Large 123B	26.53	39.57	33.25
Llama3.3 70B	30.29	34.22	33.30
Mistral Large 123B	63.05	64.88	61.62
C4AI Command R+	54.92	55.50	54.38
Qwen2.5 72B	57.28	57.28	57.33
Llama3.3 70B	52.62	52.18	53.09
Llama3.1 70B	51.68	50.67	53.25
Gemma3 27B	52.39	52.62	52.36
Llama3.1 8B	49.15	46.66	53.65
Qwen2.5 7B	51.99	56.47	49.23

Table 5: 跨各种 LLMs 的真实性预测结果。结果显示了基线（不使用检索的事实核查）和使用检索的事实核查的 LLMs。最佳结果以粗体显示。

7.1 结果

我们的结果如表 5 所示，我们使用了八个不同模型大小的 LLM。带有检索信息的 Mistral Large 表现优于所有其他 LLM，以及基线（统计上不显著）。它取得了最高的性能，使其成为实验的 LLM 中预测真实度最可靠的模型。

Qwen2.5 72B 的性能显著下降，表明仅靠模型大小并不能决定有效性。Llama 型号的表现相似，显示出在区分真实性类别时基于检索信息的能力有限。较小的模型表现最差，并且在泛化方面存在困难。

总体而言，虽然更大的 LLM 往往表现更好，但上下文信息起着重要作用。Mistral Large 的强劲表现突显了其在改进事实核查应用方面的潜力。

8 人工评估

为了评估我们所提出的多语言声明检索流程的有效性，我们开发了一个面向事实核查人员的基于网络的工具。除了对各个组成部分进行自动评估外，我们还利用开发的工具专注于对整个流程进行人工评估。我们将该工具提供给学生和学者，他们对其性能和可用性进行了评估。通过评估研讨会和结构化问卷收集的反馈，为系统的适用性提供了见解。

8.1 开发工具

我们的基于网络的应用程序 [5] 集成了 3 节中描述的流程，采用了表现最佳的 TEM 模型 – Multilingual E5 Large 。后端采用了 Llama3.3 70B ，因其强大的摘要能力和过滤

[5] https://fact-exu-miner.kinit.sk/

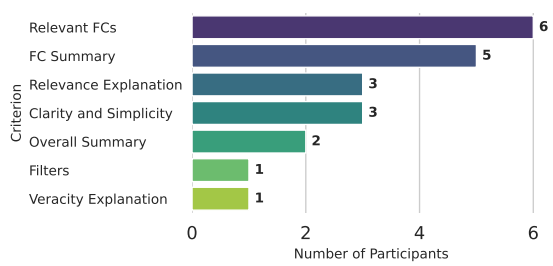

Figure 4: 每个评价标准被认为有益的参与者数量（$N = 6$）。

无关事实核查的能力而被选中。我们在 Milvus [6] 向量数据库中存储来自超过 80 种语言的事实核查声明，以及元数据和 `Multilingual E5 Large` 嵌入。系统的结果提供了一个由 LLM 识别的相关事实核查的排名列表，连同其摘要和解释。此外，我们还提供给用户一个真实性标签分布图和一个判决解释，以辅助决策。

8.2 评估 & 结果

为了评估该工具，我们进行了一个用户研究，研究对象包括六名参与者（五名新闻专业学生和一位学者）。每位参与者都对该工具进行了互动，并完成了一份问卷，该问卷旨在评估系统的可用性、输出质量以及在支持事实核查方面的整体效果。

参与者对该工具的各个方面进行了满意度评分，包括摘要、相关事实核查的解释、真实性图以及整体可用性。评分范围从 1（非常不满意）到 5（非常满意）。大多数特性得到了 3.5 到 4 之间的平均满意度评分，其中相关事实核查的解释和真实性解释获得了最高的平均评分。用户普遍认为该工具在识别相关事实核查时很有帮助，并赞赏摘要和解释的清晰度。

我们要求参与者评估该工具的主要好处（见图 4）。参与者重点指出了相关事实核查（FCs）的检索、简洁的总结和解释，以及界面的清晰度作为主要好处。这些结果表明该工具对于事实核查员来说是一个有前途的辅助工具。

9 讨论

多语言 TEMs 的表现优于英语 TEMs。 `Multilingual E5 Large` 在大多数语言中实现了最佳检索性能。然而，基于标准的检索实验显示 TEMs 在处理自然语言指令时存在困难，尤其是在按日期范围过滤时。

使用大型语言模型进行过滤可以提高精准度，但存在权衡。 `Mistral Large` 提供了在保留相关事实核查和过滤掉不相关的事实核查之间的最佳平衡，显示了对协助事实核查员的潜力。然而，精确度与召回率之间的权衡仍然具有挑战性，因为某些有用的事实核查可能会被排除在外。

较大的大规模语言模型在摘要和真实性预测方面表现出色。 较小的 LLM 经常未能遵循指令，生成的摘要是原语言的而非英语。较大的 LLM 表现更好，尤其是在文章先于指令时，并且在预测声明的真实性方面更有效。然而，由于准确评估声明真实性的固有难度，总体表现仍然是中等的。

10 结论

本文提出了一种用于多语言检索先前经过事实核查的声明的管道，集成了 LLMs 以增强事实核查过程。除了检索之外，我们的方法还通过过滤无关的事实核查、总结事实核查文章和预测真实性标签及解释来支持事实核查员。我们还开发了一个基于网络的应用程序，并评估了其在事实核查过程中的有效性。我们的研究结果表明，LLMs 有潜力提高事实核查工作流程，使其在不同语言中更高效且易于访问。

11

局限性

使用的模型。 我们的实验依赖于挑选出的最先进的 LLM 和 TEM，包括闭源模型（例如，`GPT-4o` 或 `Claude 3.5 Sonnet`）和开源模型（例如，`Mistral Large`，`Llama3`）。然而，模型性能高度依赖于训练数据和微调策略。因此，我们的发现可能无法推广到所有的 LLM 和架构，并且随着更新模型的问世，未来可能会有改进。

语言支持。 尽管我们的方法在超过 10 种语言上进行了评估，并结合了来自 20 种语言的事实核查数据，但我们的系统在处理资源稀缺的语言时可能仍会面临挑战。TEMs 和 LLMs 的性能可能会因语言而异，特别是那些缺乏预训练资源的语言。

此外，所选的 LLM 表现出不同程度的多语言能力。尽管 Hugging Face [7] 上的模型卡显示了预期的语言支持，由于训练数据的多样性和数据污染的可能性，这些模型可能在额外的语言上表现出能力。

人工评估。 我们的用户研究包括来自学术环境的六名参与者——五名新闻学学生和一名学者。虽然专业的事实核查员会是我们工具更合

[6] https://github.com/milvus-io/milvus

[7] https://huggingface.co/

适的评估者，但由于时间限制和资源有限，他们的参与不可行。然而，新闻学学生作为潜在最终用户具有专业性和相关性，因此是一个合理的替代。我们承认这一限制，并认为今后与专业事实核查员的评估是一个重要的发展方向。

自动化真实性预测。 我们的流程包括一个基于 LLM 的真实性预测，它根据检索到的事实核查来建议一个声明的真实性。然而，自动评估仍然受到事实核查数据的可用性和准确性的限制。在没有相关事实核查的情况下，系统可能难以提供可靠的预测。

12

伦理考虑

偏差。 由于我们对大型语言模型进行了实验，我们的系统可能会继承嵌入模型和大型语言模型训练数据中的偏见。这些偏见可能影响到索引声明、相关事实检查的选择和真实性评估，可能导致偏颇或误导性的输出，特别是对政治敏感或有争议的话题。

由于事实核查员决定他们将核查的内容，因此产生了额外的偏差。

开发工具。 开发工具的最终版本采用了 `Llama3.3 70B` [8] 。与较大的模型相比，选择此模型是因为其在摘要和高效推理方面的高级功能。该工具包括从使用的 LLM 中继承的偏差。

为了增强透明度并帮助用户评估输出，该工具还提供与给定声明相关的支持和反驳事实核查的数量。此信息包含在工具中的真实性分布图上。用户可以使用此信息对给定声明的真实性进行最终决策，并与大型语言模型的真实性预测进行比较。

管道的分类准确性和效率取决于最终模型——在我们的案例中为 `Llama3.3 70B` 。对应模型及其在真实性预测方面的有效性在第 7 节中进行了评估。众所周知，大型语言模型会产生幻觉（Rawte et al., 2023），因此它们可能会创建虚假的、非真实的甚至有害的信息。

此外，该工具还整合了事实核查文章和相应的声明，其中许多是在网上传播的虚假或误导性言论。因此，应用程序的用户可能会接触到虚假、误导性甚至有害的声明。为了解决这一问题，该工具包含了使用条款，明确其预期目的，识别目标用户，并指定不适合使用该工具的用户群体。

[8] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

个人信息 原始的 MultiClaim（Pikuliak et al., 2023）数据集可能包含个人信息和社交媒体帖子中的数据（例如，用户的姓名）。然而，我们在实验或开发的工具中并未使用任何个人信息。

平台和数据集的使用条款。 在我们的研究中，我们使用了 MultiClaim 数据集（Pikuliak et al., 2023），该数据集在特定条件下可访问--仅限于学术和研究用途。

此外，我们引入了法新社（AFP）[9] 的事实核查文章，这些文章可用于个人、私人和非商业用途。超出这些允许用途的任何复制或再分发都是禁止的。

我们确保我们对 MultiClaim 和 AFP 内容以及 AFP-Sum 的使用符合各自的条款和条件。

预期用途。 本研究所提供的标注数据仅供研究用途。它们来源于现有的 MultiClaim 数据集（Pikuliak et al., 2023），该数据集同样仅供研究用途。在我们的工作中，我们选择了一个子集，并为真实性预测任务标注了特定部分。

此外，我们引入了 AFP-Sum 数据集，该数据集包含来自 AFP 组织的事实核查文章及其摘要。由于 AFP 数据的版权限制，其使用严格限于研究目的。因此，我们仅将 AFP-Sum 数据集及其衍生资源发布给研究人员用于非商业研究用途。

为了促进实验结果的可重复性，我们也发布了用于获取结果的代码。这些数据集和代码仅供研究使用，要复制我们的发现需要访问原始的 MultiClaim 数据集，该数据集根据其各自的条款和条件提供。

人工智能助理的使用。 我们已使用 AI 助手进行语法检查和句子结构改进。除了方法部分（Sec. 3）详细描述的实验外，我们在研究过程中没有使用 AI 助手。

13

致谢

该项目由欧洲媒体与信息基金资助（资助编号 291191）。由欧洲媒体与信息基金支持的任何内容的唯一责任在于作者，它不一定反映 EMIF 及基金合作伙伴、卡卢斯特·古本基安基金会和欧洲大学研究所的立场。

本研究得到了捷克教育、青年和体育部通过 e-INFRA CZ（ID:90254）的支持。

[9] https://factcheck.afp.com/

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Saud Althabiti, Mohammad Ammar Alsalka, and Eric Atwell. 2024. Ta'keed: The first generative fact-checking system for arabic claims. *Preprint*, arXiv:2401.14067.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 215–236, Cham. Springer International Publishing.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2117–2120, New York, NY, USA. Association for Computing Machinery.

Eun Cheol Choi and Emilio Ferrara. 2024a. Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1441–1449, New York, NY, USA. Association for Computing Machinery.

Eun Cheol Choi and Emilio Ferrara. 2024b. Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 883–886, New York, NY, USA. Association for Computing Machinery.

Cohere For AI. 2024. c4ai-command-r-plus-08-2024 (revision dfda5ab).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Al-

biero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. CrowdChecked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 266–285, Online only. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*

2021, pages 4693–4703, Online. Association for Computational Linguistics.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10(12):1945–1948.

Andrea Hrckova, Robert Moro, Ivan Srba, Jakub Simko, and Maria Bielikova. 2024. Autonomation, not automation: Activities and needs of fact-checkers as a basis for designing human-centered ai systems. *Preprint*, arXiv:2211.12143.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A. Hale. 2021. Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

Mistral AI Team. 2024. Large enough.

Anna Neumann, Dorothea Kolossa, and Robert M Nickel. 2023. Deep learning-based claim matching with multiple negatives training. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 134–139, Online. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, Singapore. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Shaden Shaar, Nikola Georgiev, Firoj Alam, Giovanni Da San Martino, Aisha Mohamed, and Preslav Nakov. 2022. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zien Sheikh Ali, Watheq Mansour, Fatima Haouari, Maram Hasanain, Tamer Elsayed, and Abdulaziz Al-Ali. 2023. Tahaqqaq: A real-time system for assisting twitter users in arabic claim verification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3170–3174, New York, NY, USA. Association for Computing Machinery.

Michael Shliselberg and Shiri Dori-Hacohen. 2022. Riet lab at checkthat!-2022: Improving decoder based re-ranking for claim matching. In *CLEF (Working Notes)*, pages 671–678.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian

Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluciska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ivan Vykopal, Matú Pikuliak, Simon Ostermann, Tatiana Anikina, Michal Gregor, and Marián imko. 2025. Large language models for multilingual previously fact-checked claim detection. *Preprint*, arXiv:2503.02737.

Ivan Vykopal, Matú Pikuliak, Simon Ostermann, and Marián imko. 2024. Generative large language models in automated fact-checking: A survey. *Preprint*, arXiv:2407.02351.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).

A 计算资源

在我们的实验中，我们利用了由 A40 PCIe 40GB、A100 80GB 和 H100 NVL 94GB NVIDIA GPU 组成的计算基础设施。此外，我们使用了 Anthropic 的 API 来运行 Claude 3.5 Sonnet 的实验，并使用 Azure 来部署 GPT-4o。

使用 TEM 的实验大约花费了 30 个 GPU 小时。我们进行的总结和各种量化变体的比较实验大约需要 600 个 GPU 小时。最后，对具有真实性预测的整体流程的实验大约花费了 400 个 GPU 小时。

B 数据集统计

B.1 MultiClaim 数据集

在我们的实验中，我们选择了 MultiClaim (Pikuliak et al., 2023)，这是此前事实核查过

Language	Lang. Code	Average WC	# False	# True	# Unverifiable
Czech	cs	168.60 ± 242.44	100	0	0
German	de	86.08 ± 84.90	94	1	5
English	en	111.11 ± 142.39	92	5	3
French	fr	109.14 ± 129.62	95	4	1
Hindi	hi	63.36 ± 108.82	95	4	1
Hungarian	hu	123.73 ± 178.21	97	0	3
Polish	pl	102.00 ± 130.70	96	2	2
Portuguese	pt	92.25 ± 176.08	76	6	18
Slovak	sk	126.59 ± 214.57	100	0	0
Spanish	es	95.73 ± 130.48	35	33	32
Total			880	55	65

Table 6: 用于实验过滤和真实性预测的 MultiClaim 数据集子集的统计数据。我们提供了平均字数（WC）及其标准差，以及每种语言中虚假、真实和无法验证的断言数量。

的声明检索中最全面的多语言数据集。我们使用了完整的数据集来进行检索实验，如在第 4 节所述。对于其他组件，我们使用了 MultiClaim 的一个子集，选择了一组包括高资源和低资源语言的 10 种语言。从每种语言中，我们选取了 100 篇社交媒体帖子，同时力求平衡真实性标签的分布。然而，原始 MultiClaim 数据集严重失衡，主要是错误的社交媒体帖子。因此，我们的子集包含了相当大比例的错误声明。表格 6 提供了我们实验所用子集的详细统计数据。

最终的真实性评级来源于与特定帖子相关联的事实核查文章。我们手动评估了这些链接，以确保它们是从相应事实核查的元数据中正确提取的。

B.2 AFP-Sum 数据集

为了评估大型语言模型总结事实核查文章的能力，我们从 AFP 组织收集了数据。具体来说，我们提取了以 23 种语言撰写的事实核查文章，这些语言列在表格 7 中，其中也包括每种语言的文章数量。我们的数据集包括截至 2023 年 9 月发布的事实核查文章。

在最终评估中，我们仅使用了部分数据，特别是从 AFP-Sum 数据集中随机抽取每种语言 100 篇事实核查文章使用。抽取数据集的统计信息，包括 23 种语言的 2300 篇事实核查文章，如表 8 所示。除了事实核查文章的数量，我们还提供了文章及其摘要的平均字数及标准差。

由于提取的摘要是原语言的，我们使用谷歌翻译 API 将摘要翻译成英语，然后我们用它进行最终评估，并计算 BERTScore 和 ROUGE-L。

C 检索实验

表 9 提供了在 20 种语言中进行简单检索实验的结果，我们旨在评估 TEMs 根据社交媒体帖子内容检索相关事实核查的准确性。我们报告 S@10 作为评估的主要指标。

Language	Lang. Code	Domain	# Articles
English	en	https://factcheck.afp.com	6358
Spanish	es	https://factual.afp.com	3999
French	fr	https://factuel.afp.com	2883
Portuguese	pt	https://checamos.afp.com	1320
German	de	https://faktencheck.afp.com	564
Indonesian	id	https://periksafakta.afp.com	506
Polish	pl	https://sprawdzam.afp.com	386
Korean	ko	https://factcheckkorea.afp.com	359
Thai	th	https://factcheckthailand.afp.com	349
Serbian	sr	https://cinjenice.afp.com	306
Finnish	fi	https://faktantarkistus.afp.com	289
Malay	ms	https://semakanfakta.afp.com	233
Slovak	sk	https://fakty.afp.com	226
Czech	cs	https://napravoumiru.afp.com	216
Dutch	nl	https://factchecknederland.afp.com	192
Bulgarian	bg	https://proveri.afp.com	139
Bengali	bn	https://factcheckbangla.afp.com	136
Romanian	ro	https://verificat.afp.com	135
Burmese	my	https://factcheckmyanmar.afp.com	128
Hindi	hi	https://factcheckhindi.afp.com	125
Greek	el	https://factcheckgreek.afp.com	121
Hungarian	hu	https://tenykerdes.afp.com	112
Catalan	ca	https://comprovem.afp.com	110

Table 7: AFP-Sum 数据集的统计，包括语言、语言代码、领域和每种语言的文章数量。

Template

Fact-checked claim: `{claim}`
Language: `{language}` (`{language_code}`)
Published date: `{yyyy/mm/dd}`
Fact-checking organization: `{organization_name}`

An Example

Fact-checked claim: Vaccines cause autism
Language: English (en)
Published date: 2019/03/26
Fact-checking organization: healthfeedback.org

Figure 5: 用于构建过滤检索事实核查的模板，以及一个示例，展示其格式，包括被核查的声明、语言、发布日期和事实核查组织。

C.1 基于标准的检索

图 5 展示了用于过滤检索实验的模板。每个事实核查都是按照这个模板结构化的，其中包括已核查的声明、核查文章的语言、发布日期以及核查机构。然后，使用选定的 TEM 嵌入这个结构表示。

为了根据自然语言指令检索相关的事实核查，我们测试了不同的检索条件，例如按语言或按特定命名实体进行过滤。一旦我们获得了相似度评分高于 0.8 的事实核查清单，我们会基于社交媒体帖子的内容进行第二步检索。在这一步中，每个事实核查仅由经过核查的声明表示，不包含任何元数据，并使用特定的 TEM 进行嵌入以便于检索。

Lang. Code	Language	# Articles	Average WC Article	Average WC Summary
bg	Bulgarian	100	965.66 ± 533.28	81.57 ± 20.09
bn*	Bengali	100	308.93 ± 114.53	55.07 ± 17.23
ca*	Catalan	100	822.30 ± 454.67	82.69 ± 18.19
cs	Czech	100	691.35 ± 353.20	62.31 ± 14.28
de	German	100	869.32 ± 510.19	62.19 ± 15.57
el	Greek	100	1116.24 ± 500.74	86.51 ± 17.89
en	English	100	463.63 ± 197.19	58.18 ± 13.51
es	Spanish	100	713.13 ± 477.01	75.87 ± 18.69
fi*	Finnish	100	754.15 ± 369.82	57.50 ± 17.54
fr	French	100	659.96 ± 568.90	61.38 ± 23.46
hi	Hindi	100	507.20 ± 142.50	78.07 ± 17.16
hu	Hungarian	100	884.79 ± 570.36	78.02 ± 17.50
id*	Indonesian	100	458.79 ± 173.84	56.58 ± 12.63
ko	Korean	100	309.15 ± 131.40	46.99 ± 11.12
ms	Malay	100	521.20 ± 163.88	59.05 ± 13.16
my	Burmese	100	233.89 ± 77.18	31.18 ± 10.57
nl	Dutch	100	998.47 ± 515.52	73.51 ± 19.30
pl	Polish	100	836.52 ± 474.79	59.31 ± 17.34
pt	Portuguese	100	715.00 ± 343.31	80.21 ± 15.72
ro	Romanian	100	1156.78 ± 566.54	88.75 ± 19.20
sk	Slovak	100	850.55 ± 552.95	62.53 ± 22.07
sr	Serbian	100	954.83 ± 497.00	71.55 ± 19.63
th	Thai	100	121.34 ± 42.42	10.71 ± 4.68

Table 8: 用于摘要实验的数据集统计包括来自 23 种语言的 100 篇事实核查文章。标有 * 号的语言未包含在除摘要以外的其他实验中。阿拉伯语缺失，但在其他实验中使用。

Article Last

Create a 3-5 sentence summary of the following article, focusing on the main idea. Provide only the summary in English without any additional text.
Article: `{text}`
Summary:

Article First

Article: `{text}`

Create a 3-5 sentence summary of the article, focusing on the main idea. Provide only the summary in English without any additional text.
Summary:

Figure 6: 用于摘要实验的提示。

D 摘要实验

图 6 展示了我们在摘要实验中使用的最终提示格式。我们提供了“文章最后”和“文章最前”两种变体。

表 10 展示了使用各种开源和闭源大型语言模型（LLMs）在 23 种语言上的摘要实验结果，并使用 BERTScore 指标进行评估。对于每个 LLM，我们报告了两种设置下的表现：当文章在指令之前提供（文章优先）和当文章在指令之后提供（文章置后）。

类似地，表 11 汇总了基于 ROUGE-L 指标的结果。

除了评估这两种设置外，我们还研究了不同量化变体对 LLM 性能的影响。具体而言，我们比较了非量化模型及量化到 4 位和 8 位精度的版本。在这些实验中，我们选择了 `Llama` 模型，重点关注 1B 和 3B 变体中的 `Llama 3.1 70B` 和 `Llama3.2` 。23 种语言的 BERTscore 结果展示在表 12 中，而表 13 则报告了 ROUGE-L 得分。

E 真实性解释

图 7 提供了我们流程中每个步骤的提示模板。这些提示在流程中被用来获得最终的真实性预测。

E.1 错误分析

在本节中，我们研究真实性预测中的错误和不正确的解释。我们进行了错误预测子集的人工检查和自动分析来评估这些错误。

人工分析。 在我们的人工调查中，我们随机选择了每个 LLM [10] 的 20 个预测标签错误的样本，总共 140 个样本。作者之一分析了检索到的相关事实核查和 LLM 生成的解释，并将它们分类为几种类型。

最普遍的错误类别是主张中缺少上下文，在我们手动检查的样本中，这种情况占了 27 %。缺少的上下文使得很难识别相关的事实核查并正确预测真实性。值得注意的是，在这些情况下，63 % 的 LLMs 提供了正确的解释，承认了上下文的缺失。

第二常见的错误（占 16 % 的情况）源于前面步骤的失败，其中相关的事实核查未被识别。在这些情况下，LLMs 正确地解释了所检索的事实核查信息中没有一项与给定的声明直接相关，但仍然产生了错误的真实性评估。

错误的 17 % 是由于对声明的误解和相关事实核查的提供。在这些情况下，LLMs 集中在所提供的事实核查的不同方面，或者未能抓住声明的重点。我们观察到一些实例，其中 LLMs 错误地依赖于社交媒体帖子本身的信息来解释其真实性，尤其是在较长的帖子中。

另一种错误模式（案例的 12 %）涉及 LLMs 根据其生成摘要中提到的评级来预测真实性，而实际的事实核查评级不同。例如，一个摘要可能将某个说法描述为骗局，而从事实核查中得出的评级是“无法验证”。

在 4 % 的情况下，LLMs 仅依赖于第一次事实核查的评级，尽管在提示上下文中存在正确评级的后续事实核查。这表明了一种错误的假设，即在真实性预测中应该使用第一次事实核查。

[10] 由于是在手动调查已经完成之后的最后阶段加入研究的，`Gemma3` 模型未包含在误差分析中。

Model	Ver.	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
BM25	Og	0.75	0.71	0.70	0.63	0.61	0.63	0.74	0.46	0.61	0.49	0.58	0.75	0.31	0.56	0.56	0.77	0.70	0.78	0.73	0.31	0.62
English TEMs																						
DistilRoBERTa	En	0.79	0.86	0.88	0.58	0.73	0.64	0.79	0.65	0.65	0.82	0.82	0.75	0.77	0.72	0.65	0.64	0.86	0.85	0.72	0.89	0.75
MiniLM-L6	En	0.84	0.89	0.85	0.64	0.80	0.69	0.82	0.70	0.75	0.87	0.84	0.78	0.79	0.76	0.70	0.70	0.86	0.84	0.77	0.90	0.79
MiniLM-L12	En	0.84	0.90	0.86	0.64	0.80	0.70	0.82	0.72	0.77	0.86	0.83	0.78	0.80	0.73	0.72	0.71	0.86	0.86	0.78	0.89	0.79
MPNet-Base	En	0.80	0.87	0.89	0.57	0.77	0.68	0.81	0.70	0.72	0.87	0.80	0.79	0.80	0.74	0.67	0.67	0.86	0.85	0.75	0.88	0.77
GTE-Large-En	En	0.82	0.88	0.88	0.65	0.82	0.73	0.84	0.72	0.74	0.85	0.84	0.76	0.81	0.78	0.71	0.69	0.87	0.86	0.79	0.89	0.80
GTR-T5-Large	En	0.86	0.86	0.88	0.69	0.83	0.77	0.86	0.74	0.79	0.89	0.86	0.82	0.88	0.78	0.74	0.80	0.88	0.87	0.84	0.90	0.83
Multilingual TEMs																						
BGE-M3	Og	0.84	0.87	0.90	0.74	0.80	0.69	0.87	0.67	0.82	0.89	0.90	0.86	0.86	0.74	0.72	0.79	0.88	0.89	0.84	0.93	0.82
DistilUSE-Base-Multilingual	Og	0.74	0.81	0.71	0.50	0.60	0.56	0.69	0.57	0.53	0.78	0.74	0.60	0.62	0.61	0.60	0.58	0.80	0.77	0.64	0.72	0.66
LaBSE	Og	0.77	0.84	0.81	0.48	0.70	0.44	0.72	0.57	0.56	0.82	0.77	0.67	0.77	0.61	0.57	0.66	0.78	0.74	0.64	0.79	0.69
Multilingual E5 Small	Og	0.81	0.89	0.82	0.71	0.80	0.61	0.80	0.63	0.72	0.87	0.85	0.77	0.69	0.72	0.71	0.76	0.89	0.83	0.81	0.89	0.78
Multilingual E5 Base	Og	0.81	0.87	0.85	0.70	0.77	0.64	0.83	0.60	0.67	0.88	0.86	0.80	0.74	0.73	0.66	0.77	0.88	0.84	0.81	0.89	0.78
Multilingual E5 Large	Og	0.84	0.90	0.92	0.78	0.82	0.75	0.86	0.74	0.81	0.90	0.91	0.88	0.81	0.83	0.77	0.82	0.90	0.89	0.87	0.85	0.84
MiniLM-L12-Multilingual	Og	0.49	0.83	0.75	0.48	0.58	0.58	0.66	0.55	0.49	0.79	0.61	0.54	0.58	0.64	0.61	0.51	0.79	0.77	0.57	0.81	0.63
MPNet-Base-Multilingual	Og	0.70	0.81	0.78	0.53	0.63	0.61	0.73	0.56	0.63	0.83	0.71	0.62	0.75	0.66	0.60	0.57	0.84	0.80	0.64	0.86	0.69

Table 9: 使用 S@10 指标检索 20 种语言中已被事实核查的声明的 TEM 结果。每个配置的最佳分数——英文翻译 (En) 或原始语言 (Og) ——已用加粗显示。GTR-T-Large 在英文翻译中表现最佳，而 Multilingual E5 Large 在多语言数据中表现优异，超过了英文 TEM。

Model	Version	Quant.	bg	bn	ca	cs	de	el	en	es	fi	fr	hi	hu	id	ko	ms	my	nl	pl	pt	ro	sk	sr	th	Avg.
Open-Source LLMs																										
C4AI Command R+	Article first	4bit	0.75	0.76	0.74	0.74	0.75	0.76	0.76	0.73	0.75	0.74	0.76	0.75	0.77	0.77	0.76	0.74	0.75	0.75	0.74	0.76	0.75	0.75	0.75	0.75
	Article last	4bit	0.70	0.71	0.72	0.69	0.71	0.67	0.76	0.70	0.71	0.69	0.71	0.68	0.72	0.75	0.74	0.66	0.73	0.70	0.72	0.69	0.73	0.73	0.64	0.71
Llama3.1 70B Instruct	Article first	4bit	0.75	0.77	0.74	0.74	0.75	0.76	0.76	0.73	0.75	0.74	0.76	0.75	0.77	0.77	0.76	0.72	0.75	0.75	0.74	0.76	0.75	0.75	0.75	0.75
	Article last	4bit	0.75	0.77	0.74	0.74	0.75	0.76	0.76	0.74	0.74	0.74	0.77	0.69	0.77	0.77	0.76	0.72	0.75	0.75	0.74	0.75	0.75	0.75	0.75	0.75
Llama3.3 70B Instruct	Article first	4bit	0.74	0.76	0.73	0.74	0.74	0.75	0.75	0.73	0.75	0.73	0.76	0.75	0.76	0.76	0.75	0.70	0.75	0.74	0.74	0.76	0.74	0.74	0.75	0.74
	Article last	4bit	0.75	0.76	0.73	0.74	0.74	0.75	0.75	0.73	0.75	0.73	0.76	0.74	0.76	0.76	0.75	0.70	0.75	0.74	0.73	0.75	0.74	0.74	0.75	0.74
Mistral Large	Article first	4bit	0.75	0.77	0.73	0.74	0.75	0.76	0.76	0.73	0.75	0.74	0.75	0.75	0.77	0.77	0.76	0.74	0.75	0.75	0.74	0.75	0.74	0.74	0.75	0.75
	Article last	4bit	0.75	0.76	0.74	0.74	0.75	0.75	0.76	0.73	0.75	0.74	0.75	0.74	0.77	0.77	0.76	0.74	0.75	0.74	0.74	0.75	0.74	0.74	0.75	0.75
Qwen2 72B Instruct	Article first	4bit	0.74	0.75	0.73	0.73	0.74	0.75	0.75	0.72	0.74	0.73	0.75	0.73	0.76	0.76	0.74	0.74	0.74	0.74	0.73	0.74	0.74	0.74	0.74	0.74
	Article last	4bit	0.74	0.76	0.73	0.73	0.74	0.74	0.75	0.72	0.74	0.73	0.75	0.74	0.76	0.76	0.74	0.73	0.74	0.74	0.73	0.75	0.74	0.73	0.74	0.74
Qwen2.5 0.5B Instruct	Article first	-	0.70	0.68	0.71	0.70	0.72	0.69	0.74	0.71	0.68	0.71	0.69	0.69	0.73	0.72	0.71	0.68	0.71	0.71	0.71	0.71	0.70	0.69	0.72	0.70
	Article last	-	0.70	0.68	0.70	0.70	0.71	0.68	0.73	0.70	0.68	0.70	0.68	0.69	0.73	0.72	0.71	0.67	0.70	0.70	0.70	0.71	0.70	0.69	0.72	0.70
Qwen2.5 1.5B Instruct	Article first	-	0.73	0.73	0.73	0.72	0.73	0.72	0.75	0.73	0.71	0.73	0.73	0.72	0.76	0.74	0.74	0.70	0.74	0.73	0.73	0.74	0.73	0.72	0.74	0.73
	Article last	-	0.73	0.73	0.72	0.72	0.73	0.72	0.75	0.72	0.72	0.72	0.73	0.72	0.75	0.75	0.74	0.69	0.73	0.73	0.72	0.73	0.72	0.72	0.74	0.73
Qwen2.5 3B Instruct	Article first	-	0.74	0.75	0.73	0.73	0.74	0.74	0.76	0.73	0.73	0.73	0.75	0.73	0.76	0.75	0.75	0.71	0.74	0.74	0.74	0.75	0.74	0.74	0.74	0.74
	Article last	-	0.71	0.71	0.71	0.69	0.71	0.71	0.74	0.71	0.71	0.72	0.72	0.71	0.72	0.74	0.72	0.68	0.71	0.69	0.71	0.72	0.69	0.68	0.73	0.71
Qwen2.5 7B Instruct	Article first	-	0.73	0.75	0.73	0.73	0.74	0.74	0.75	0.72	0.73	0.73	0.74	0.73	0.76	0.75	0.75	0.70	0.73	0.73	0.73	0.74	0.73	0.73	0.74	0.74
	Article last	-	0.74	0.75	0.73	0.73	0.74	0.74	0.76	0.73	0.74	0.73	0.74	0.73	0.76	0.75	0.75	0.72	0.74	0.74	0.74	0.75	0.74	0.74	0.75	0.74
Qwen2.5 72B Instruct	Article first	4bit	0.75	0.77	0.74	0.74	0.74	0.75	0.76	0.73	0.75	0.74	0.76	0.74	0.77	0.76	0.76	0.74	0.75	0.75	0.74	0.75	0.75	0.75	0.75	0.75
	Article last	4bit	0.75	0.76	0.74	0.74	0.74	0.75	0.75	0.73	0.75	0.74	0.76	0.75	0.77	0.76	0.76	0.74	0.75	0.74	0.74	0.76	0.74	0.75	0.75	0.75
Gemma3 27B	Article first	4bit	0.72	0.75	0.72	0.72	0.73	0.74	0.74	0.72	0.73	0.73	0.74	0.73	0.75	0.75	0.74	0.73	0.73	0.72	0.72	0.73	0.73	0.72	0.73	0.73
	Article last	4bit	0.72	0.74	0.72	0.73	0.73	0.74	0.74	0.72	0.73	0.72	0.74	0.73	0.75	0.74	0.74	0.72	0.73	0.73	0.72	0.73	0.73	0.73	0.73	0.73
Closed-Source LLMs																										
Claude 3.5 Sonnet	Article first	-	0.74	0.76	0.73	0.74	0.73	0.74	0.75	0.72	0.74	0.73	0.75	0.74	0.76	0.75	0.75	0.73	0.74	0.74	0.73	0.74	0.74	0.73	0.75	0.74
	Article last	-	0.74	0.76	0.73	0.74	0.74	0.75	0.76	0.73	0.74	0.74	0.75	0.74	0.76	0.76	0.75	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.75	0.74
GPT-4o	Article first	-	0.74	0.76	0.73	0.74	0.74	0.75	0.75	0.72	0.74	0.73	0.75	0.74	0.76	0.76	0.75	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.75	0.74
	Article last	-	0.74	0.76	0.73	0.73	0.74	0.75	0.75	0.72	0.74	0.73	0.75	0.74	0.76	0.76	0.75	0.74	0.74	0.74	0.73	0.74	0.74	0.74	0.75	0.74

Table 10: 在两种设置下（文章优先和文章最后），对各种大型语言模型（LLM）在 23 种语言上的摘要性能进行 BERTScore 评估。每种语言的最佳结果用粗体显示。

Model	Version	Quant.	bg	bn	ca	cs	de	el	en	es	fi	fr	hi	hu	id	ko	ms	my	nl	pl	pt	ro	sk	sr	th	Avg.
Open-Source LLMs																										
C4AI Command R+	Article first	4bit	0.32	0.34	0.28	0.28	0.30	0.32	0.33	0.27	0.30	0.28	0.34	0.30	0.37	0.34	0.35	0.28	0.31	0.30	0.30	0.32	0.29	0.30	0.32	0.31
	Article last	4bit	0.12	0.19	0.18	0.12	0.12	0.08	0.32	0.08	0.19	0.05	0.16	0.10	0.11	0.28	0.24	0.15	0.23	0.14	0.11	0.06	0.25	0.26	0.09	0.16
Llama3.1 70B Instruct	Article first	4bit	0.31	0.34	0.29	0.29	0.31	0.32	0.33	0.29	0.31	0.28	0.34	0.31	0.37	0.34	0.34	0.24	0.32	0.30	0.31	0.33	0.28	0.30	0.33	0.31
	Article last	4bit	0.30	0.32	0.28	0.28	0.31	0.30	0.33	0.28	0.28	0.27	0.33	0.13	0.34	0.34	0.32	0.24	0.31	0.30	0.28	0.27	0.28	0.29	0.32	0.29
Llama3.3 70B Instruct	Article first	4bit	0.30	0.35	0.29	0.29	0.31	0.32	0.32	0.28	0.31	0.27	0.35	0.30	0.37	0.34	0.34	0.22	0.31	0.30	0.30	0.33	0.29	0.29	0.34	0.31
	Article last	4bit	0.31	0.34	0.29	0.28	0.31	0.31	0.32	0.29	0.31	0.27	0.35	0.26	0.35	0.35	0.34	0.22	0.31	0.29	0.30	0.31	0.29	0.29	0.34	0.31
Mistral Large	Article first	4bit	0.31	0.33	0.27	0.28	0.30	0.32	0.33	0.27	0.30	0.27	0.32	0.30	0.35	0.34	0.33	0.27	0.31	0.29	0.30	0.30	0.28	0.29	0.32	0.30
	Article last	4bit	0.30	0.33	0.28	0.29	0.31	0.32	0.31	0.27	0.31	0.27	0.33	0.29	0.36	0.34	0.33	0.28	0.31	0.29	0.29	0.31	0.30	0.29	0.33	0.31
Qwen2 72B Instruct	Article first	4bit	0.28	0.31	0.25	0.26	0.27	0.29	0.29	0.25	0.28	0.24	0.30	0.27	0.32	0.31	0.30	0.26	0.28	0.27	0.26	0.28	0.27	0.27	0.29	0.28
	Article last	4bit	0.28	0.32	0.25	0.26	0.28	0.29	0.30	0.24	0.28	0.24	0.29	0.28	0.32	0.31	0.29	0.26	0.28	0.27	0.27	0.28	0.26	0.27	0.30	0.28
Qwen2.5 0.5B Instruct	Article first	-	0.19	0.16	0.19	0.18	0.24	0.16	0.25	0.21	0.15	0.19	0.17	0.16	0.25	0.23	0.21	0.15	0.21	0.19	0.20	0.19	0.17	0.16	0.24	0.19
	Article last	-	0.20	0.16	0.18	0.19	0.21	0.16	0.25	0.18	0.15	0.19	0.18	0.16	0.24	0.24	0.22	0.15	0.18	0.19	0.17	0.19	0.17	0.17	0.24	0.19
Qwen2.5 1.5B Instruct	Article first	-	0.25	0.24	0.24	0.22	0.25	0.23	0.26	0.25	0.21	0.24	0.26	0.21	0.31	0.26	0.28	0.17	0.27	0.25	0.25	0.27	0.23	0.23	0.29	0.25
	Article last	-	0.26	0.24	0.22	0.21	0.24	0.23	0.26	0.25	0.23	0.23	0.25	0.21	0.29	0.28	0.27	0.17	0.25	0.24	0.23	0.26	0.24	0.23	0.29	0.24
Qwen2.5 3B Instruct	Article first	-	0.27	0.28	0.25	0.24	0.27	0.27	0.30	0.25	0.26	0.24	0.28	0.25	0.31	0.30	0.28	0.21	0.27	0.26	0.26	0.28	0.25	0.26	0.30	0.27
	Article last	-	0.24	0.25	0.23	0.19	0.25	0.23	0.28	0.23	0.21	0.23	0.24	0.22	0.27	0.29	0.26	0.16	0.25	0.20	0.23	0.25	0.18	0.17	0.29	0.23
Qwen2.5 7B Instruct	Article first	-	0.25	0.28	0.23	0.24	0.26	0.26	0.26	0.23	0.24	0.23	0.27	0.25	0.30	0.26	0.27	0.17	0.25	0.25	0.24	0.26	0.25	0.24	0.27	0.25
	Article last	-	0.28	0.30	0.25	0.26	0.28	0.28	0.29	0.25	0.27	0.24	0.29	0.25	0.32	0.31	0.30	0.23	0.28	0.26	0.27	0.29	0.27	0.27	0.31	0.28
Qwen2.5 72B Instruct	Article first	4bit	0.29	0.32	0.27	0.28	0.29	0.29	0.30	0.26	0.30	0.25	0.31	0.29	0.34	0.32	0.32	0.26	0.29	0.29	0.28	0.29	0.28	0.28	0.31	0.29
	Article last	4bit	0.29	0.32	0.26	0.27	0.28	0.30	0.30	0.25	0.30	0.25	0.31	0.29	0.34	0.32	0.31	0.25	0.29	0.26	0.28	0.29	0.28	0.29	0.31	0.29
Gemma3 27B	Article first	4bit	0.26	0.30	0.24	0.25	0.26	0.26	0.29	0.24	0.26	0.25	0.30	0.26	0.32	0.29	0.31	0.26	0.27	0.25	0.25	0.26	0.25	0.25	0.30	0.27
	Article last	4bit	0.26	0.30	0.24	0.25	0.27	0.27	0.28	0.24	0.26	0.24	0.30	0.27	0.31	0.29	0.30	0.25	0.26	0.25	0.25	0.27	0.25	0.26	0.30	0.27
Closed-Source LLMs																										
Claude 3.5 Sonnet	Article first	-	0.30	0.34	0.26	0.28	0.29	0.29	0.30	0.26	0.29	0.27	0.33	0.30	0.33	0.32	0.31	0.27	0.29	0.28	0.28	0.30	0.28	0.27	0.33	0.29
	Article last	-	0.29	0.33	0.27	0.29	0.30	0.30	0.32	0.27	0.29	0.27	0.33	0.28	0.34	0.34	0.31	0.28	0.29	0.29	0.29	0.29	0.28	0.28	0.33	0.30
GPT 4o	Article first	-	0.29	0.32	0.27	0.28	0.29	0.30	0.30	0.26	0.29	0.27	0.33	0.28	0.33	0.33	0.31	0.27	0.28	0.28	0.28	0.29	0.28	0.27	0.31	0.29
	Article last	-	0.29	0.33	0.26	0.27	0.29	0.29	0.30	0.26	0.29	0.25	0.32	0.29	0.32	0.33	0.31	0.26	0.28	0.27	0.27	0.29	0.27	0.28	0.31	0.29

Table 11: ROUGE-L 对 23 种语言的总结性能进行了评估，评估对象是不同的大语言模型，分为两种设置：文章在前和文章在后。每种语言的最佳结果用粗体显示。

Model	Version	Quant.	bg	bn	ca	cs	de	el	en	es	fi	fr	hi	hu	id	ko	ms	my	nl	pl	pt	ro	sk	sr	th	Avg.
Llama3.2 1B Instruct	Article first	4bit	0.70	0.67	0.71	0.69	0.71	0.70	0.74	0.70	0.66	0.69	0.64	0.64	0.72	0.72	0.70	0.65	0.72	0.68	0.70	0.70	0.69	0.68	0.60	0.69
		8bit	0.72	0.72	0.71	0.71	0.73	0.73	0.74	0.71	0.70	0.72	0.72	0.67	0.74	0.73	0.73	0.66	0.73	0.71	0.72	0.72	0.72	0.71	0.72	0.72
		-	0.72	0.72	0.72	0.71	0.73	0.73	0.74	0.71	0.70	0.71	0.72	0.66	0.74	0.73	0.73	0.67	0.72	0.71	0.71	0.73	0.71	0.71	0.72	0.72
	Article last	4bit	0.61	0.60	0.66	0.64	0.67	0.62	0.74	0.68	0.61	0.67	0.63	0.63	0.69	0.63	0.67	0.52	0.67	0.65	0.67	0.65	0.61	0.62	0.59	0.64
		8bit	0.64	0.63	0.68	0.67	0.70	0.64	0.75	0.69	0.64	0.70	0.64	0.65	0.71	0.69	0.69	0.53	0.69	0.67	0.69	0.67	0.65	0.65	0.61	0.66
		-	0.64	0.63	0.69	0.67	0.70	0.64	0.75	0.70	0.64	0.70	0.64	0.64	0.71	0.69	0.69	0.53	0.69	0.67	0.69	0.67	0.65	0.66	0.61	0.66
Llama3.2 3B Instruct	Article first	4bit	0.74	0.75	0.73	0.73	0.74	0.74	0.76	0.73	0.72	0.73	0.75	0.69	0.76	0.75	0.74	0.69	0.74	0.74	0.72	0.75	0.73	0.73	0.74	0.73
		8bit	0.74	0.75	0.73	0.73	0.74	0.74	0.75	0.72	0.74	0.73	0.75	0.70	0.75	0.75	0.74	0.70	0.74	0.73	0.72	0.75	0.73	0.73	0.74	0.74
		-	0.73	0.75	0.73	0.73	0.74	0.75	0.76	0.72	0.74	0.73	0.75	0.69	0.75	0.76	0.74	0.70	0.74	0.74	0.72	0.75	0.73	0.73	0.74	0.74
	Article last	4bit	0.73	0.75	0.72	0.72	0.73	0.73	0.76	0.72	0.70	0.73	0.75	0.66	0.76	0.76	0.75	0.68	0.73	0.72	0.71	0.73	0.72	0.72	0.74	0.73
		8bit	0.70	0.76	0.71	0.69	0.72	0.66	0.75	0.72	0.67	0.73	0.73	0.66	0.74	0.76	0.73	0.69	0.72	0.71	0.71	0.69	0.68	0.69	0.73	0.71
		-	0.69	0.76	0.70	0.69	0.72	0.67	0.75	0.72	0.66	0.73	0.73	0.66	0.74	0.76	0.73	0.70	0.72	0.71	0.71	0.69	0.68	0.69	0.74	0.71
Llama3.1 70B Instruct	Article first	4bit	0.75	0.77	0.74	0.74	0.75	0.76	0.76	0.73	0.75	0.74	0.76	0.75	0.77	0.77	0.76	0.72	0.75	0.75	0.74	0.76	0.75	0.75	0.75	0.75
		-	0.75	0.77	0.74	0.74	0.74	0.75	0.76	0.73	0.75	0.74	0.76	0.75	0.77	0.77	0.76	0.75	0.75	0.75	0.74	0.75	0.74	0.75	0.75	0.75

Table 12: BERTScore 评估大型语言模型在 23 种语言中的摘要表现，比较未量化模型与 4 比特和 8 比特量化版本。每种语言的最佳结果以粗体显示。

Model	Verstion	Quant	bg	bn	ca	cs	de	el	en	es	fi	fr	hi	hu	id	ko	ms	my	nl	pl	pt	ro	sk	sr	th	All
Llama3.2 1B Instruct	Article first	4bit	0.21	0.16	0.20	0.17	0.22	0.19	0.28	0.20	0.11	0.17	0.06	0.06	0.21	0.24	0.15	0.14	0.22	0.14	0.15	0.17	0.18	0.14	0.04	0.17
		8bit	0.24	0.26	0.22	0.23	0.26	0.25	0.29	0.24	0.19	0.23	0.26	0.10	0.29	0.27	0.27	0.17	0.26	0.22	0.24	0.25	0.22	0.21	0.27	0.24
		-	0.24	0.27	0.22	0.22	0.26	0.26	0.29	0.24	0.19	0.22	0.26	0.09	0.29	0.27	0.27	0.17	0.26	0.22	0.24	0.25	0.22	0.20	0.28	0.23
	Article last	4bit	0.01	0.00	0.04	0.05	0.04	0.02	0.27	0.06	0.02	0.05	0.01	0.03	0.06	0.05	0.04	0.02	0.07	0.02	0.05	0.04	0.03	0.01	0.01	0.04
		8bit	0.03	0.02	0.08	0.08	0.11	0.03	0.26	0.11	0.03	0.12	0.03	0.04	0.14	0.16	0.09	0.05	0.09	0.07	0.09	0.05	0.06	0.04	0.04	0.08
		-	0.02	0.01	0.08	0.08	0.11	0.03	0.26	0.11	0.03	0.13	0.03	0.04	0.14	0.16	0.09	0.04	0.09	0.07	0.09	0.05	0.06	0.05	0.04	0.08
Llama3.2 3B Instruct	Article first	4bit	0.27	0.31	0.25	0.27	0.28	0.27	0.31	0.25	0.25	0.24	0.32	0.14	0.32	0.30	0.30	0.19	0.28	0.28	0.22	0.30	0.26	0.26	0.30	0.27
		8bit	0.29	0.31	0.26	0.27	0.28	0.29	0.31	0.26	0.28	0.25	0.32	0.17	0.32	0.31	0.30	0.21	0.28	0.28	0.20	0.30	0.26	0.27	0.31	0.27
		-	0.29	0.32	0.27	0.27	0.28	0.29	0.31	0.26	0.28	0.25	0.31	0.14	0.32	0.32	0.30	0.22	0.29	0.27	0.17	0.30	0.26	0.26	0.31	0.27
	Article last	4bit	0.26	0.30	0.19	0.22	0.24	0.24	0.29	0.23	0.18	0.24	0.29	0.06	0.30	0.30	0.28	0.19	0.25	0.24	0.12	0.21	0.22	0.21	0.29	0.23
		8bit	0.13	0.31	0.09	0.13	0.20	0.06	0.30	0.18	0.08	0.22	0.23	0.04	0.22	0.30	0.19	0.19	0.16	0.17	0.10	0.07	0.09	0.12	0.29	0.17
		-	0.11	0.31	0.09	0.12	0.21	0.06	0.29	0.19	0.06	0.22	0.24	0.04	0.19	0.31	0.17	0.21	0.17	0.16	0.11	0.07	0.10	0.10	0.29	0.17
Llama3.1 70B Instruct	Article first	4bit	0.31	0.34	0.29	0.29	0.31	0.32	0.33	0.29	0.31	0.28	0.34	0.31	0.37	0.34	0.34	0.24	0.32	0.30	0.31	0.33	0.28	0.30	0.33	0.31
		-	0.32	0.35	0.29	0.30	0.31	0.32	0.33	0.29	0.32	0.28	0.36	0.31	0.37	0.35	0.34	0.30	0.32	0.30	0.31	0.33	0.29	0.30	0.34	0.32

Table 13: 在 23 种语言中对大型语言模型的摘要进行 ROUGE-L 评估，比较非量化模型与 4 位和 8 位量化模型的变体。每种语言的最佳结果用粗体表示。

Article Summary

Article: `{document}`

Create a 3-5 sentence summary of the article, focusing on the main idea. Provide only the summary in English without any additional text.
Summary:

Overall Summary

Retrieved claim: `{claim1}`
Summary: `{article_summary1}`

Retrieved claim: `{claim2}`
Summary: `{article_summary2}`

...

Generate a brief, one-paragraph summary that captures the key information from all the relevant claims and fact-checks. Ensure the summary covers the main points of each claim and addresses all the topics presented, while remaining concise and comprehensive.

Filtration

Input claim: `{post_text}`

Claim ID: 1
Fact-checked claim: `{claim1}`

...

Claim ID: 50
Fact-checked claim: `{claim50}`

Identify only fact-checked claims that are implied by the input claim. For each claim, provide the claim ID, the fact-checked claim, and an explanation of fact-checked claim's implication to the input claim.

Output Format (JSON):

```
{
    "fact_checked_claims": [
    {
        "claim_id": "<ClaimID1>",
        "fact_checked_claim": "<Claim1>",
        "explanation": "<Explanation of Claim1>"
    },
    {
        "claim_id": "<ClaimID2>",
        "fact_checked_claim": "<Claim2>",
        "explanation": "<Explanation of Claim2>"
    }
    ]
}
```

Veracity prediction

Input Claim: `{post_text}`

Fact-checked claim: `{claim1}`
Summary of the fact-check article: `{article_summary1}`

Fact-checked claim: `{claim2}`
Summary of the fact-check article: `{article_summary2}`

...

Based only on the provided fact-checked information that is directly relevant to the input claim, determine the veracity of the claim.
Ignore fact-checks that do not apply. The veracity should be classified as one of the following:
- "True" if the claim is accurate based on the relevant fact-checked information.
- "False" if the claim is inaccurate based on the relevant fact-checked information.
- "Unverifiable" if there is insufficient or no relevant fact-checked information to assess the claim.

Provide a concise explanation that justifies your prediction.

Output Format (JSON):

```
{
    "veracity": "<True/False/Unverifiable>",
    "explanation": "<Explanation for the prediction>"
}
```

Figure 7: 用于真实性预测的管道中的提示模板。

Model	Missing FC [%]
Mistral Large	25.5
C4AI Command R+	25.3
Qwen2.5 72B	39.1
Llama3.1 70B	33.5
Llama3.3 70B	29.2
Llama3.1 8B	29.9
Qwen2.5 7B	48.6

Table 14: 对于每个 LLM，没有在检索上下文中出现的与事实核对相关的基准真实信息的帖子百分比。

最后，15 % 的错误可以归因于事实真相问题，主要是在事实核查将声称归类为 “无证据” 的情况下——而我们的标准化过程将其转换为 “无法验证”。然而，在所有这些情况下，LLMs 对这些声称及其真实情况的解释都是正确的，并且得到了事实核查总结信息的支持。

自动分析。 由于观测到的一个错误来源于前几步未能识别相关的事实核查，我们进行了一次自动分析，重点关注缺少相关事实核查的案例比例。表 14 显示了每个模型中没有将任何真实相关的事实核查包括在相关事实核查列表中的帖子百分比。如果无法获取相关的事实核查，模型可能无论其推理能力如何，都难以准确预测真实性。分析显示了大型语言模型之间的差异，其中较小的模型通常表现出更高的缺失事实核查率。值得注意的是，Qwen2.5 7B 显示出最高比例（48.6 %）的帖子缺少相关的事实核查，而 C4AI Command R+ 和 Mistral Large 表现最佳，约有 25 % 的帖子缺少相关的事实核查。这些发现表明，检索质量仍然是事实核查流程中的瓶颈，尤其是对于较小的模型。

F 开发的应用程序

基于网络的应用程序集成了在第 3 节介绍的实施流程。对于检索，我们使用表现最好的 TEM 模型 Multilingual E5 。后端运行 Llama3.3 70B ，因其强大的摘要能力和有效过滤不相关的事实核查而被选中。

我们的事实核查数据库汇总了来自超过 80 种语言的多个事实核查组织的核查声明。我们在 Milvus [11] 向量数据库中存储了核查声明、元数据（例如，语言、事实核查文章、评级）以及核查声明的计算 Multilingual E5 嵌入。

用户提交查询，系统返回由 LLM 识别的相关事实核查的排名列表，以及它们的摘要和解释。此外，系统提供一个总体摘要、真实性标签分布图和判决解释。这些信息支持用户做出最终决策。

F.1 接口

开发的应用程序由四个主要组件组成。(1) 文本输入（见图 8），用户在此提供需要工具返回相关事实核查文章的声明。(2) 相关事实核查列表（见图 9），我们在此提供由 LLM 识别的所有相关事实核查。(3) 非相关事实核查列表（见图 10），我们在此列出检索步骤中检索到但未被 LLM 分类为相关的事实核查。由于 LLM 在识别相关事实核查方面的准确率不是 100 %，我们还提供其他事实核查，以使应用程序更具鲁棒性，并提供我们的管道中获得的所有信息，便于事实核查者做出明智决定。(4) 系统响应（见图 11），其中包括输入声明和所有相关事实核查的总体总结、基于相关事实核查文章评分的真实性分布图以及预测的真实性标签的解释。

[11] https://github.com/milvus-io/milvus

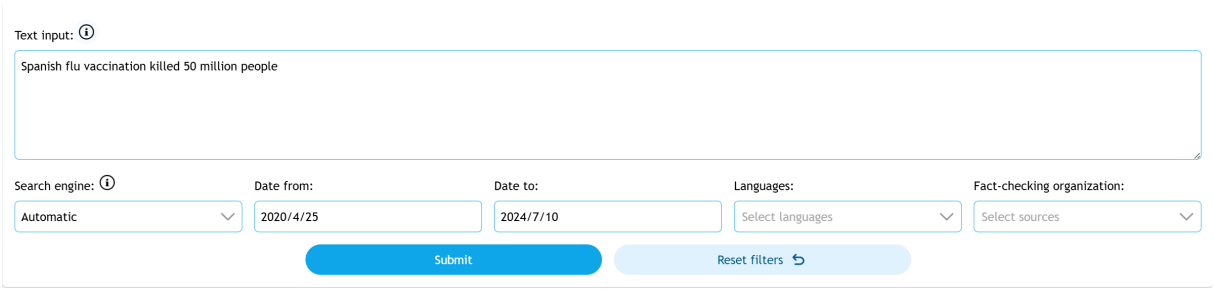

Figure 8: 用于文本输入的用户界面组件。

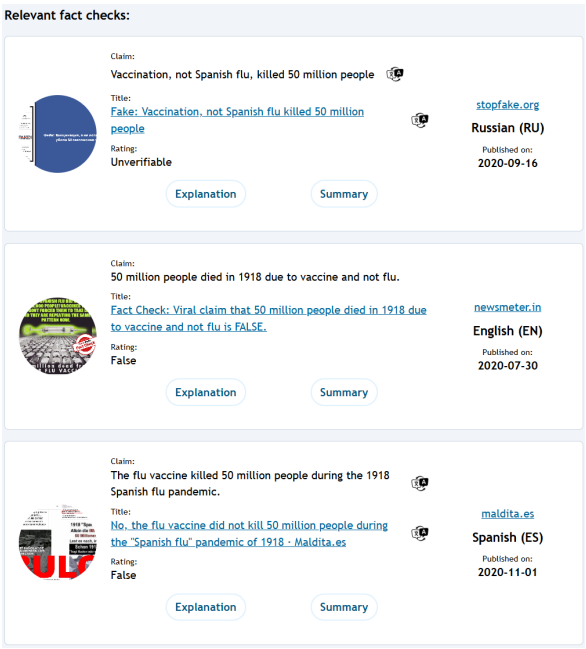

Figure 9: 用户界面组件用于显示由我们流程中的 LLM 识别的相关事实核查列表。对于每个相关的事实核查，我们提供事实核查文章的摘要以及为什么该事实核查被归类为相关的解释。

Figure 10: 用于显示非相关事实核查的用户界面组件。

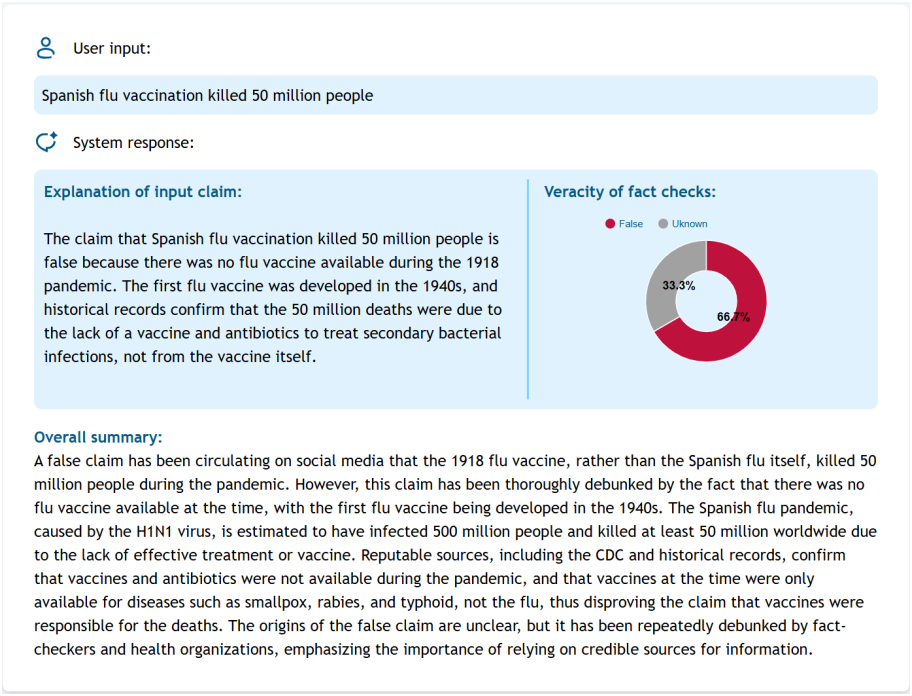

Figure 11: 系统响应的用户界面组件，其中我们提供声明的总体概述和相关的事实核查、真实性分布图以及预测真实性的解释。