arXiv:2602.10748v1 [cs.DB] 11 Feb 2026

用于知识图谱验证的大语言模型基准测试

Farzad Shami
farzad.shami@aalto.fi 阿尔托大学 埃斯波，芬兰

Stefano Marchesin
stefano.marchesin@unipd.it 帕多瓦大学 帕多瓦，意大利

Gianmaria Silvello
gianmaria.silvello@unipd.it 帕多瓦大学 帕多瓦，意大利

摘要

知识图谱（KG）通过关系连接实体来存储结构化事实知识，这对许多应用至关重要。这些应用依赖于知识图谱的事实准确性，因此验证事实十分重要，但也颇具挑战。由专家进行人工验证固然理想，但大规模开展并不现实。自动化方法展现出良好前景，但尚未准备好用于真实世界知识图谱。大语言模型（LLM）凭借其语义理解能力和知识获取潜力，为此带来了可能，但其是否适用于知识图谱事实验证以及效果如何，仍未得到充分探索。

本文介绍了 FactCheck，这是一个基准测试，旨在从三个关键维度评估大语言模型的知识图谱事实验证能力：(1) 大语言模型的内部知识；(2) 通过检索增强生成（RAG）获取外部证据；以及 (3) 运用多模型共识策略汇总知识。我们在三个多样化的真实世界知识图谱上评估了开源和商用大语言模型。FactCheck 还包含一个用于知识图谱事实验证的 RAG 数据集，其中有 2+百万篇文档。

实验分析表明，尽管大语言模型表现出良好潜力，但其稳定性和可靠性仍不足以应用于真实世界的知识图谱验证场景。通过 RAG 方法整合外部证据会导致性能波动，相较于更精简的方法，改进效果并不稳定，且计算成本更高。同样，基于多模型共识的策略并不总是优于单个模型，这凸显了不存在放之四海而皆准的 解决方案。这些发现进一步强调了 FactCheck 这类基准测试的必要性，以便系统地评估这项艰巨却至关重要的任务，并推动其不断进步。

关键词

知识图谱、大语言模型、事实验证

1 引言

知识图谱（KGs）是机器可解释的有向标记多重图，其中节点表示实体或概念，边表示类型化语义关系。它们为现实世界知识提供结构化表示，使信息源之间能够进行推理、整合和查询 [23, 26, 47]。知识图谱已广泛应用于各种场景 [23, 24], ，包括：(1) 网络搜索，用于查询和内容的语义理解 [16, 41]；(2) 电子商务，用于推荐 [56]和对话代理 [49]；(3) 社交网络，用于建模用户兴趣 [18, 41]；以及 (4) 金融 [3], 、交通运输 [21], 和能源 [72]等其他领域。然而，下游应用的有效性取决于知识图谱中事实的准确性。每一条独立的知识通常表示为一个 <S、P、O> 三元组（即，[redacted]必须在事实上正确。此外，整个知识图谱的可靠性不仅取决于这些原子事实的正确性，还取决于它们相互连接的方式 [8, 50]。[1]

知识图谱构建完成后，一个关键步骤是评估其中事实的真实性 [23, 51]。这需要确定数据对现实世界实体、关系和现象的反映有多准确。事实验证是一项重大挑战，且成本高昂 [36, 37]。最可靠的方式是由人类专家进行人工或计算机辅助标注 [42, 66]。然而，这一过程极其耗时 [12, 43]。由于专家通常需要依赖多个外部参考来源来审核事实，在大型知识图谱中（例如 DBpedia [33] 或 YAGO [22]），验证每个三元组都可能耗费数分钟，因此大规模人工检查和纠正错误并不可行。

因此，自动化事实核查方法 [11, 14, 58, 62,63],通常基于规则和可强制执行的约束 [10, 40],，作为更具可扩展性的替代方案应运而生，以解决人工方案在时间和成本方面的局限。虽然这些方法对定义明确且频繁出现的事实[29]）十分有效，但面对真实世界知识图谱中种类繁多的事实时，却难以实现泛化。另一方面，人工定义既困难又昂贵。因此，可以采用（半）自动化方法来提取规则和约束。尽管如此，这些方法主要针对用于识别高频正例的规则；对于涉及低频事实或需要应用负规则的情况，它们仍难以应对 [45]。

这些局限促使人们采用基于机器学习／深度学习解决方案的事实核查系统 [17]。在这一领域，一种可行的方法是利用大语言模型（LLM）进行事实核查，因为它们在各类任务中已展现出接近人类的表现 [46]。在此框架下，大语言模型具有多种优势：能够从文本中提取上下文信息、理解陈述的语义，并拥有广泛的内部知识库 [48, 65]。然而，当前的大语言模型会生成幻觉内容和不忠实的回答 [60]。此外，近期研究指出，大语言模型在事实验证任务中尤其存在问题，表现出可能影响其可靠性的系统性偏差和知识缺口 [59]。为克服大语言模型的知识截止和幻觉所带来的局限，当前基于大语言模型构建的系统通常采用检索增强生成（RAG）方法，即利用外部来源的数据补充大语言模型，从而改善其回答 [28]。然而，尽管大语言模型研究近期取得了诸多进展，且大语言模型具备处理广泛任务的能力，目前似乎仍没有专门衡量大语言模型在知识图谱事实验证方面表现的基准测试 [51]。

因此，我们提出了 FactCheck，这是一个通用基准测试，旨在从以下三个方面评估大语言模型验证知识图谱事实的能力：

[1]我们根据上下文交替使用事实、陈述和三元组这几个术语。

EDBT '26，坦佩雷（芬兰）© 2026 版权所有归所有者/作者所有。本文由 OpenProceedings.org 出版，国际标准书号 978-3-98318-104-9，系列国际标准连续出版物号 2367-2005。本文可依据知识共享许可协议知识共享署名-非商业性使用-禁止演绎 4.0 国际的条款进行分发。

主要维度：（1）大语言模型内部知识；（2）通过检索增强生成（RAG）获取的外部证据；以及（3）由多个模型综合而成的知识。

FactCheck 依托一套验证流程，将结构化三元组转换为自然语言陈述，以评估其事实准确性。验证程序首先处理知识图谱实体和关系，推导出结构化三元组，再依据可靠来源对其进行核查，并计算准确性评分。

FactCheck 由以下研究问题（RQ）驱动：**研究问题1：** 仅依靠内部知识时，大语言模型在知识图谱事实核查方面的效果如何?

研究问题2： 外部证据是否能提升大语言模型对知识图谱进行事实核查的能力?

研究问题3： 汇总多个大语言模型的预测，是否能提高知识图谱事实的验证可靠性?

RQ1 针对近期关于大语言模型能否充当知识库的讨论，旨在评估大语言模型内部知识的事实准确性和完整性，包括已见过和未见过的知识 [19, 20, 71]。我们不会提示大语言模型检索知识来评估其完整性和准确性，而是要求大语言模型判断外部提供的事实是否准确，这要求它完全依赖内部知识。我们的重点是这一研究方法，同时承认已有研究表明，与提示大语言模型生成或评估自身内容相比，询问其信息准确性验证结果会更理想 [27, 31]。

RQ2 关注通过为大语言模型补充外部证据来提升知识图谱事实核查的有效性，并为围绕检索增强生成及其在事实验证中作用的持续讨论作出贡献 [28, 53, 70]。经典的检索增强生成方法通常优于仅依赖内部知识的大语言模型，但近期研究发现，在复杂或多轮场景中，检索增强生成的效果可能会减弱，因为上下文管理和证据选择更容易出错 [32]。此外，整合外部证据可能引入上下文偏差，使模型过度信任检索内容 [34]。借助 FactCheck，我们希望推动相关研究，探讨外部证据是否有助于知识图谱事实验证、在何种程度上有帮助，以及在什么条件下发挥作用。

研究问题3聚焦于一项日益增多的研究，这些研究探讨汇总多个大语言模型的输出能否提高事实验证的准确性或可靠性 [7, 54]。不同大语言模型在事实准确性、推理模式以及易产生幻觉的程度上可能各不相同，但近期研究表明，通过投票、共识或仲裁机制结合多个模型，可以减轻个别模型的偏差并提高鲁棒性 [64, 67]。然而，这种方法也带来了自身的挑战，包括分歧解决、扩展成本，以及放大在重叠数据上训练的模型之间共同误解的风险。FactCheck 可用于探索多个大语言模型的集成式推理能否提高知识图谱事实核查的可靠性。

贡献. 我们提出 FactCheck，这是一个使用大语言模型进行知识图谱事实验证的基准测试，具有以下几项优势：

(1) FactCheck 整合了多种大语言模型，用于知识图谱事实验证。该基准测试利用这些模型的内部知识以及通过检索增强生成获取的外部证据进行评估。此外，它还探索了基于共识的验证方式，即采用多数投票策略。针对中等规模（7–9B 参数）和商用大语言模型开展的实验凸显了这项任务面临的挑战。

(2) FactCheck 基于三个真实世界知识图谱数据集构建：*Fact-Bench* [14], *YAGO* [43], 和*DBpedia* [38], ，涵盖了广泛的知识领域，从日常事实到复杂的特定领域信息，确保能够多样且有代表性地评估事实验证能力。

(3) FactCheck 包含一个大规模检索增强生成数据集，其中包含若干问题及其对应的 Google 搜索引擎结果页面（SERP）。该数据集包含 2M+ 份文档，涵盖广泛的事实信息，是规模最全面的公开知识图谱事实验证检索增强生成资源之一。FactCheck 还包含一个模拟 API，可模拟真实的搜索 API，让用户无需直接访问搜索引擎，即可复现数据检索、测试检索方法并扩展检索增强生成方法。

(4) 专用网络应用程序（https://factcheck .dei.unipd . it/），使用户能够直观地探索和分析验证过程的每个步骤；该应用还提供错误分析模块，对推理错误进行分类，从而能够系统地识别大语言模型在事实核查场景中的局限性。

(5) FactCheck 通过结合性能指标与资源使用分析，实现全面评估。将模型预测与金标准标签进行比较，以评估准确性和可靠性。该基准测试还会追踪计算成本（推理时间和令牌用量）。

采用不同的方法和数据集进行评估，凸显了知识图谱事实验证任务的难度和固有复杂性。我们的工作主要有三点发现：第一，尽管大语言模型在知识图谱事实验证方面展现出可喜的能力，但距离可靠地应用于现实世界的验证场景仍有很大差距。第二，通过检索增强生成整合外部知识会导致性能波动；与更精简的方法相比，它带来的改进并不稳定，且计算成本显著更高。最后，基于共识、使用多个模型的策略无法持续优于单个模型。总之，这些结果凸显了该任务的难度和复杂性，也表明需要建立专门的基准测试来推动这一领域的发展。

概述。 本文其余部分安排如下。第 2 节回顾自动化知识图谱事实核查和基准开发方面的相关工作。第 3 节介绍 FactCheck 基准测试。第 4 节详细介绍 FactCheck 的构建过程，涵盖数据集选择和检索增强生成语料库构建。第 5 节概述实验设置，第 6 节讨论实验结果。第 7 节对失败案例进行定性错误分析。最后，第 8 节总结全文。

2 相关工作

2.1 自动化知识图谱事实核查

事实核查方法可分为两类：一类直接利用知识图谱为给定陈述查找支持路径 [29, 57, 58, 61] ，另一类则依赖外部参考来源查找支持性或冲突性证据 [14, 62]。表 1 对这两种范式进行了比较分析。

(1) 基于内部知识图谱的事实核查. Knowledge Stream (KStream) 和 Relational KnowledgeLinker (KLinker) [58] 是无监督的基于网络流的方法，旨在评估以 <S,P,O> 三元组表示的事实陈述的真实性。KStream 将知识图谱建模为流网络，其中路径将流量从主语传递到宾语，以支持或反驳某个

表1：基于内部知识图谱与基于外部证据的事实核查机制比较分析。

特征	基于内部知识图谱的事实核查	基于外部证据的事实核查
原理	连贯性：与图模式一致。	对应性：与外部来源相符。
主要证据	图拓扑、路径和流网络	非结构化文本、网页和搜索摘要
假设	基于局部完备性，从缺失链接中提取负信号。	在不完备性条件下，使用外部数据验证缺失链接。
机制	路径挖掘、链接预测。	信息检索、自然语言处理、检索增强生成。
负例处理	通过采样策略合成（例如 [29]）。	检索失败或矛盾。
权衡	（+）快速、一致。（-）会漏掉图谱错误。	（+）有效性高。（-）速度慢，依赖来源。
示例	KStream [58],PredPath [57],COPPAL [61]。	DeFacto [14], KGValidator [5],FactCheck（本文方法）。

给定陈述。另一方面，KLinker 专注于发现彼此连接实体的关系路径。COPPAL [61]提出了一种用于查找支持陈述的路径的佐证元路径。这些方法只关注正向证据路径，并且受知识图谱不完整性的严重限制。PredPath 等方法 [57] 尝试同时利用负向和正向路径，以涵盖更广泛的事实陈述。PredPath 仅依据正确示例为具有判别力的谓词路径分配权重，而忽略反例，这可能导致规则的权重设置不当。此外，Kim 和 Choi [29] 提出了一种无监督的基于规则的方法，其表现显著优于该领域最先进的无监督方法。他们通过在知识图谱中查找正向和负向证据路径、为训练阶段生成示例、创建用于从正规则和负规则中学习的模型，并根据已确立的证据为三元组评分，从而计算给定陈述的真实性分数。

尽管这些方法有效，但它们完全依赖底层知识图谱，而底层知识图谱可能包含错误或不完整；因此，这些方法无法用于评估知识图谱本身的准确性。

（*2*）基于外部证据的事实核查。 DeFacto [14]是一种监督学习方法，通过使用从网络检索到的证据来验证知识图谱三元组。为计算证据分数，该方法将可信度指标与文本证据相结合。Syed 等人 [62] 提出了一种事实验证方法，将静态参考语料库中的文本证据用作外部知识。他们将三元组转化为自然语言，查询搜索引擎以检索语料库中相似的句子，然后从这些句子中提取证据和特征，并利用训练好的模型估算每个知识图谱三元组的置信度。最近，Boylan 等人 [5] 提出了 KGValidator，这是一个利用大语言模型自动评估知识图谱补全模型的框架。KGValidator 通过利用多种上下文来源来评估预测的三元组，包括大语言模型的内部知识、用户提供的文本文件以及网络资源。与这一方法论贡献不同，FactCheck 专注于提供必要的评估基础设施——即数据集、指标和精选证据语料库——以便系统地评估和比较此类验证方法。

与先前采用外部来源进行事实验证的研究一致， [5, 14, 62], FactCheck 允许大语言模型利用从网络搜索引擎结果页面检索到的外部证据。此外，FactCheck 提供了多种基于大语言模型的基线方法，可据此将大语言模型与由外部证据驱动的解决方案进行比较评估。此外，FactCheck 还针对这项任务，基于三个真实世界知识图谱数据集（13,530 条事实）评估大语言模型的性能，并以 2M+份检索到的文档作为外部证据。

2.2 基准测试和数据集

CRAG [69] 是一个**基准测试**，旨在评估检索增强生成系统的有效性，重点关注事实准确性。它包含 4,409 个问答对，涵盖五个领域和八种问题类别。为模拟真实使用场景，CRAG 提供了用于网络和知识图谱搜索的模拟 API。该基准测试专门针对诸如回答较少见或快速变化的事实等挑战，并评估大语言模型在不同实体流行度和时间相关性水平下的性能。尽管 CRAG 和 FactCheck 都采用检索增强生成，但两者解决的是本质上不同的问题，评估目标也各不相同。事实上，FactCheck 评估知识图谱事实验证，优先考虑准确性和一致性。CRAG 无法取代 FactCheck，因为高性能问答模型往往无法满足验证独立知识图谱三元组所要求的严格、细粒度逻辑。此外，FactCheck 还提供有关计算成本和资源效率的详细信息，而这两个方面并未在 CRAG 中得到广泛涉及。因此，尽管这些基准测试有所关联，但它们针对事实验证的不同方面。

除了 CRAG 之外，还有若干面向文本声明事实核查的流程和共享任务。RumourEval [15]通过立场检测和谣言真实性验证分析社交媒体帖子，评估分类系统，所用数据集包含来自 Twitter 和 Reddit 的数据。CLEF Check-That! [1] 提供针对新闻文章的句子级主观性检测。ClaimBuster [2] 提出了一种自动化端到端事实核查流程，整合了声明检测、匹配和验证。如前所述，这些基准测试主要针对非结构化文本声明，无法用于知识图谱事实验证。

目前提出的用于知识图谱验证的**数据集** 很少 [14, 38, 43] [14], [33] 。其中一个重要数据集是*FactBench*，它基于 DBpedia 和 Freebase [4] 知识图谱构建，用于评估验证系统识别系统性错误的能力。其他数据集包括*YAGO* [43] 和*DBpedia* [38],，它们由从各自知识图谱中抽取的样本组成，并由专家人工标注其正确性。尽管这些数据集已用于人工和自动化验证，但几乎没有被用于基于大语言模型的方法。因此，我们在 FactCheck 中采用 FactBench、YAGO 和 DBpedia，因为它们涵盖了事实验证挑战中相互补充的方面，从而能够对基于大语言模型的策略进行多角度评估。另一个相关数据集是 FactKG [30],，专为知识图谱上的事实验证而设计。不过，FactKG 使用知识图谱来验证文本声明，而我们的工作方向相反：利用外部证据帮助大语言模型验证知识图谱事实。

3 FactCheck

本节详细介绍 FactCheck 为解决研究问题所采用的策略。该基准测试包含多种策略，使用开源和商用大语言模型。在 §3.1 中，我们介绍两种仅依赖大语言模型内部知识来验证知识图谱事实的方法（研究问题1）。在 §3.2 中，我们介绍一种检索增强生成方法，通过外部证据增强大语言模型。最后，§3.3

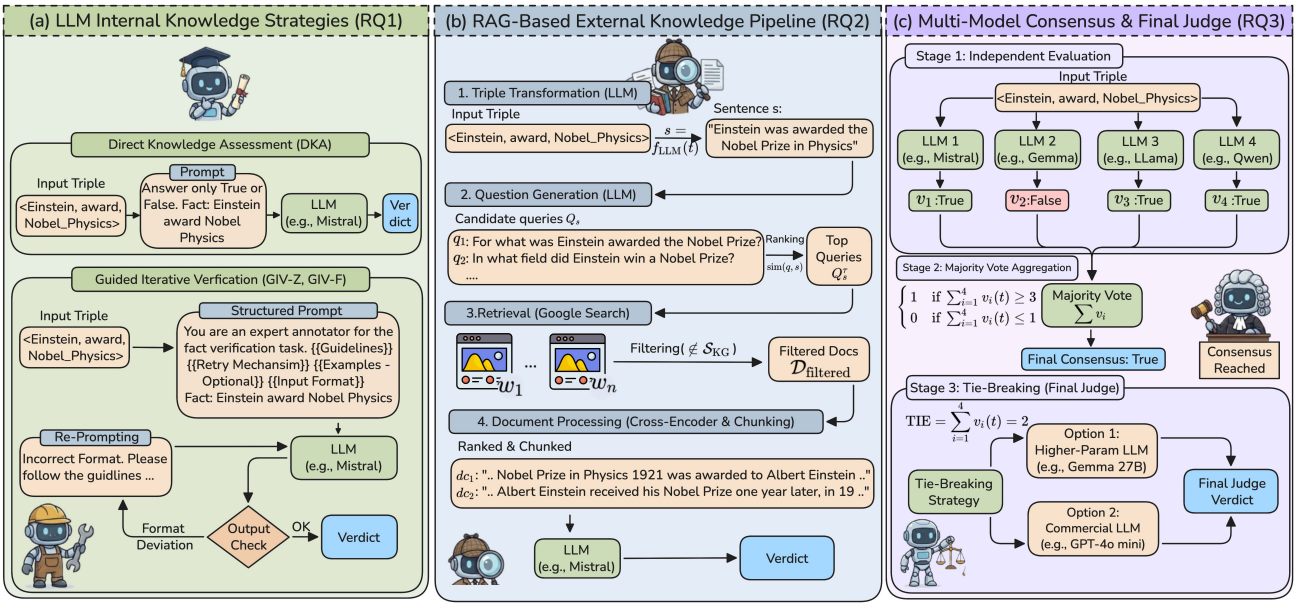


图 1：FactCheck 基准测试的总体概览。

介绍一种多模型共识策略，该策略汇总多个大语言模型的预测结果，以提高验证准确率（研究问题3）。

3.1 大语言模型内部知识

为了解决 RQ1，FactCheck 采用了两种不同的策略：

直接知识评估（DKA）是一种简单的策略，仅向大语言模型提供基本且直接的提示，不给予任何进一步指导。DKA 旨在评估大语言模型仅凭内部知识验证事实的能力。我们将 DKA 作为基线，用于比较不同的大语言模型和更高级的策略。图 1(a) 左上部分给出了一个示例。

引导式迭代验证（GIV）（见图 1(a) 左下部分）是一种迭代提示方法，它利用结构化提示模板来说明预期的输出格式，并可选择性地强制遵循数据集特定约束。如果模型的输出不符合要求，系统就会触发重新提示，明确指出其不合规之处。对于反复未能满足标准的响应，会将其标记为无效。我们同时考虑零样本和少样本设置。在少样本设置中，我们会加入一小组已正确评估的三元组作为示例，以帮助模型理解任务。这些示例在不同数据集之间共享，且在语义层面独立于知识图谱；同时，其编码会根据目标知识图谱进行调整，以符合谓词和模式约定。

3.2 外部知识

为了解决 RQ2，我们通过检索增强生成增强大语言模型。给定一个知识图谱三元组 t，我们检索一组文档 D，其中包含可能支持或反驳该三元组的证据。我们通过一个包含四个主要阶段的多阶段流水线实现这一过程：(1) 三元组转换，(2) 问题生成与排序，(3) 文档检索与筛选，以及 (4) 文档处理与分块。图 1(b) 展示了 FactCheck 中基于检索增强生成的验证引擎的核心组件。

在三**元组转换** 阶段 (1) 中，结构化的知识图谱三元组会被转换为可读性强的句子。我们使用大语言模型执行这一转换，以应对不同知识图谱在表示 S，P，O 数据时存在的显著差异。不同知识图谱采用不同的三元组编码约定，而这些特定于来源的格式往往会妨碍有效的信息检索。常见问题包括：(1) 知识图谱特定命名空间（例如，*dbpedia.org/resource/:term:*）；(2) 下划线或驼峰命名法等特殊表示法（例如，*isMarriedTo*、*Alexander_III_of_Russia*）；以及 (3) 缺乏充分语法或语义上下文的谓词。这类表示可能会使搜索结果局限于提取三元组的原始源页面，从而造成检索偏差，并限制评估期间的信息覆盖范围。相比之下，自然语言改写有助于发现更多相关网络来源。我们将此过程定义为转换函数 $s = f_{\text{LLM}}(t)$，它将一个三元组 t 映射为自然语言句子 s。

在**问题生成与排序** 阶段（2）中，对于给定的句子 s，我们提示大语言模型生成一组候选查询 $\mathrm{Q}_s = \{q_1, q_2, \ldots\ldots, q_{k_q}\}$。生成多个问题的目的是扩大原始三元组的语义覆盖范围，从而提高检索到相关证据的可能性，即使输入含糊、含噪或信息不足也是如此。生成多个问题还有助于缓解大语言模型将三元组转换为自然语言时可能引入的释义偏差。通过构造若干不同的问题，我们扩大了对给定三元组可能存在的解读范围，从而削弱其与某个单一方面的关联，而这种关联可能会因某个特定的大语言模型生成的释义而被强加。为找出信息量最大的查询，我们使用交叉编码器模型（*jina-reranker-v1-turbo-en*），其对应于交叉编码器最终表征与学习得到的相关性向量之间的归一化点积（即经 Sigmoid 缩放的点积得分）。该得分反映候选查询 $q \in \mathrm{Q}_s$ 与原始句子 s 之间的语义接近度。最终得到的集合为 $\mathrm{Q}_s^{\text{ranked}} = \{q_{(1)}, q_{(2)}, \ldots\ldots, q_{(k_q)}\}$，其中对于所有 $i \in \{1, 2, \ldots\ldots, k_q - 1\}$，$\mathrm{sim}(q_{(i)}, s) \geq \mathrm{sim}(q_{(i+1)}, s)$。我们保留

前-τ个查询，记为 Q_s^τ，并使用预定义阈值 $\tau \in [0,1]$，以确保仅使用最相关的查询。

在**文档检索与过滤**阶段（3）中，我们使用特定参数将 Q_s^τ 中的每个查询提交至 Google 搜索，以确保一致性。我们将 lr = “lang_en”和 hl = “en”设置为强制使用英语内容和界面设置，并将 gl = “us”设置为将地理位置标准化为美国，从而减轻本地个性化偏差。使用 num = “100”，我们收集排名前 $n_{\max} = 100$ 的检索网页，记为 $R(q) = \{w_1, w_2, \ldots, w_{n_{\max}}\}$。对于每个网页 $w_i \in R(q)$，我们提取其文本内容，记为 $\text{text}(w_i)$。针对给定查询 q 检索到的文档集合随后定义为 $D(q) = \{d_i = \text{text}(w_i) \mid w_i \in R(q)\}$。为获得与原始三元组 t关联的完整文档池，我们对 Q_s^τ中的所有查询取并集：$D = \bigcup_{q \in Q_s^\tau} D(q)$。为确保证据独立性并避免循环验证，我们将 S_{KG} 定义为原始知识图谱来源集合——例如，在验证来自 DBpedia 和 FactBench 数据集的事实时的维基百科条目。我们使用该集合过滤掉任何直接源自这些来源的检索文档。所得过滤后的文档集合定义为 $D_{\text{filtered}} = \{d \in D \mid \text{source}(d) \notin S_{KG}\}$。

最后，在**文档处理与分块**阶段（4）中，我们使用交叉编码器找出与句子 s最相关的 k_d 文档。对于每份文档$d \in D_{\text{filtered}}$，使用与上述相同的方法计算相似度分数 $\text{sim}_d(d, s)$。按相似度排序后，排名前 k_d 的文档构成最终集合 $D_{\text{final}} = \{d_1, d_2, \ldots\ldots, d_{k_d}\}$。$D_{\text{final}}$ 中的每份文档都使用滑动窗口分块策略划分为更小且相互重叠的段落。在事实验证阶段，这些分块随后作为上下文输入用于大语言模型提示。

3.3 多模型共识

由于大语言模型对同一事实核查任务可能给出不同答案，我们还探讨了一种模型共识策略（图 1(c)）。基于 §3.1 和 §3.2，设 $M = \{M_1, M_2, M_3, M_4\}$ 为大语言模型的集合。对于每个三元组 t，每个模型 $M_i \in M$ 都会给出一个二元判定 $v_i(t) \in \{0, 1\}$，其中 0 表示“假”，1 表示“真”。

我们采用简单多数投票策略来确定最终判定。给定三元组 t 的共识决策 V_{final}（t）为：

$$V_{\text{final}}(t) = \begin{cases} 1 & \text{if } \sum_{i=1}^{4} v_i(t) \geq 3 \\ tie & \text{if } \sum_{i=1}^{4} v_i(t) = 2 \\ 0 & \text{otherwise} \end{cases}$$

该策略旨在通过降低离群预测的影响来减少错误。若出现平局，我们将采用冲突解决策略。设 M_{judge} 为负责打破平局的最终裁决模块。我们探索了两种定义 M_{judge}的方法：

(1) 从模型集合 M 中选取一个模型的更高参数版本，其选择依据是一致性得分 CA_M。该得分表示模型输出在各数据集上与多数预测一致的实例比例，可作为其与正确结果对齐程度的替代指标。我们测试了一致性最高（CA_M 最高）和最低（CA_M 最低）的模型，并将它们升级为更高参数版本（例如，Gemma2:9B → 27B）。

（2）采用不同架构和训练流程的商业模型，例如*GPT-4o mini*，可为解决模糊案例提供独立视角。

4 基准构建

本节介绍构建 FactCheck 基准测试的完整流程。首先，在 §4.1 中，我们详细说明从现有知识图谱数据集中收集三元组的过程，以及专为检索增强生成方法创建新数据集的过程。接着，在 §4.2 和 §4.3 中，我们介绍 FactCheck 中使用的大语言模型、评估指标和自动化评估流程。

4.1 数据集

FactCheck 数据集由两个主要部分组成：(i) 来自三个真实世界知识图谱的三元组；(ii) 从 Google 搜索结果页检索的内容。本节将分别介绍这两个部分，并介绍模拟 API。该 API 模拟真实场景，并为数据集提供标准化访问方式，以支持可复现实验。

知识图谱数据集。我们收录了来自三个真实且广泛使用的知识图谱数据集的三元组——FactBench、YAGO 和 DBpedia。请注意，我们使用这些数据集时采用基于快照的语义：如果某个三元组得到用于构建该数据集的底层知识图谱快照的支持，则视为真，否则视为假。表 2 总结了这些数据集的关键统计信息。

表 2：FactBench、YAGO 和 DBpedia 数据集概要。

	FactBench	YAGO	DBpedia
事实数量	2,800	1,386	9,344
谓词数量	10	16	1,092
每个实体的平均事实数	2.42	1.69	3.18
金标准准确率 (μ)	0.54	0.99	0.85

FactBench 是 Gerber 等人开发的多语言基准，用于评估事实验证算法。它涵盖十种关系类型，并支持英语、德语和法语。在 FactCheck 中，我们仅关注英语子集。正（正确）事实来源于 DBpedia 和 Freebase，而负（错误）事实则通过系统性地篡改正确事实生成，确保符合定义域和值域约束。我们采用的配置中，正事实占比为 $\mu = 0.54$；该比例通过混合正确事实与采用各种负采样策略生成的错误事实实现 [37]。

YAGO 是从 YAGO知识图谱中抽样得到的评估数据集，最初由 Ojha 和 Talukdar 提出 [43]，并被广泛用于知识图谱准确率估计 [12, 36, 37]。该数据集包含 1，386 条事实，涵盖 16 种不同的谓词，每个实体平均对应 1.69 条事实。所有事实均由众包工作者标注，因此其金标准准确率为 $\mu = 0.99$。如此高的准确率给事实核查带来了独特挑战，因为大语言模型可能倾向于将所有事实都判定为正确，从而抬高性能指标。

DBpedia是一个从 DBpedia 知识图谱中抽取的评估数据集，最初由 Marchesin 等人 [38] 提出。该数据集采用抽样与主动学习技术相结合的方式构建，并由专家和非专业标注员共同参与，以确保较高的标注质量。这些三元组取自 DBpedia 的 2015-10 英语版本，其中主体实体必须属于包含 rdfs:label 和 rdfs:comment 谓词的三元组。为专注于事实断言，研究排除了 T-Box 三元组——即表示本体实体和

模式级关系的三元组——仅保留代表具体事实主张的 A-Box 断言。每个三元组至少由三名标注者进行标注，最终形成一个包含 9,934 个三元组的数据集，金标准准确率为 $\mu = 0.85$，涵盖 $1,092$ 个不同的谓词。

***RAG* 数据集。** 我们构建了一个 RAG 数据集，其中包含根据知识图谱事实生成的问题及其对应的搜索结果。创建该数据集旨在为涉及外部知识的事实验证任务提供支持，以便有效评估大语言模型性能。该数据集由两个主要部分组成：生成的问题，以及从 Google 搜索结果页获得的相关搜索结果。

对于**问题**，我们使用大语言模型为每个转换后的三元组生成 $k_q = 10$ 个不同的问题 s，以探究其底层事实的不同方面。在构建数据集时，我们纳入了从模型输出中成功提取的所有问题。每个问题都会与其对应的相似度分数一同发布，该分数是根据转换后的三元组计算得出的。FactCheck 共包含针对 13,530 条事实生成的 $Q = 130,820$ 个问题。每条事实关联的问题数量各不相同（q_t），范围从最少 $q_t = 2$ 到最多 $q_t = 10$，均值为 $\mu_{q_t} = 9.67$，中位数为 $\tilde{q}_t = 10.00$。

每个问题都被赋予一个相似度分数 $\delta \in [0, 1]$，用于量化其与转换后的三元组之间的语义接近度。所有问题的相似度分数均值为 $\mu_\delta = 0.63$，中位数为 $\tilde{\delta} = 0.66$。标准差为 $\sigma_\delta = 0.25$，表明多样性适中。第一四分位数为 $Q_1 = 0.44$，第三四分位数为 $Q_3 = 0.84$，因此四分位距（IQR）为 IQR $= Q_3 - Q_1 = 0.40$，这证实数据集中相似度分数存在显著差异。

为进一步分析这一分布，我们将问题分为三个相似度等级：高相似度（$\delta \geq 0.70$），占数据集的 45%；中等相似度（$0.40 \leq \delta < 0.70$），占 34%；低相似度（$\delta < 0.40$），占剩余的 21%。这一分布表明，数据集中 79% 的问题与转换后的三元组至少具有中等程度的相似性（$\delta \geq 0.40$），近半数问题则具有较高相似度。这些不同的相似度层级涵盖了与每个事实在语义上接近以及关联较为宽泛的解读。

关于**Google 搜索结果**，对于每个事实，我们将转换后的原始三元组以及生成的问题中相似度分数最高的三个问题提交给 Google 搜索。解析 HTML 响应后，我们使用 *GRequests* Python库检索每个 URL。随后，我们使用 *newspaper4k* [2] Python软件包提取生成的网页内容。

语料库包含 $D = 2,090,305$ 份文档，涵盖 13,530 个三元组。每个三元组 t 关联 d_t 份文档，其中最少为 $d_t = 0$，最多为 $d_t = 337$，均值为 $\mu_{d_t} = 154.51$，中位数为 $\tilde{d}_t = 160$。中位数略高，表明分布略呈负偏态，大多数三元组关联的文档数量接近或略高于均值。

我们将 $\mathrm{E}_{\text{text}} \subset D$ 定义为文本内容为空的文档子集。该子集包含 $|\mathrm{E}_{\text{text}}| = 263$，515 篇文档，占整个文档集合的 13%。因此，文本覆盖率——即包含文本内容的文档所占比例——为 $1 - |\mathrm{E}_{\text{text}}|/|D| = 0.87$ (87%)。如此高的覆盖率表明所构建的文档集合具有可靠性。

[2]https://newspaper4k.readthedocs.io

表 3：检索增强生成 RAG 数据集生成流水线各步骤的平均耗时和令牌用量汇总。

Task	平均耗时	平均令牌数
问题生成	9.60 秒	672.58
获取文档（Google 页面）	3.60 秒	–
获取每个三元组的文档	350 秒	–

表 3 报告了生成 RAG 数据集所耗费的时间和 token 开销。总体而言，每个事实的问题生成平均需要 9.60 秒，而完整的谷歌结果检索过程约需 364。4 秒。

为确保评估的公平性和可复现性，我们提前生成了所有问题，并收集了相应的 Google 搜索结果页结果。这为大语言模型提供了一致的证据基础，避免因实时搜索结果变化而产生差异。完整数据集已公开发布在我们的 HuggingFace 项目页面上，并可通过模拟 API 访问。[3]

模拟 ***API***。 在 FactCheck 中，我们集成了类似网络搜索的 API 来检索内容，以模拟真实的检索增强生成场景。该 API 通过提供对预先收集的搜索数据的标准化访问，支持可复现的基准测试，从而消除搜索结果随时间变化而产生的多样性。

对于所考察数据集中的每个事实，我们使用转换后的三元组和生成的前三个问题分别发起查询。我们保存了 Google 搜索结果页中每个查询的前 100 条结果，随后检索并留存每个链接网页的实际内容。如前所述，我们过滤掉了与原始事实直接相关的来源，以避免循环验证。

我们实现了标准化端点，模拟常规网络搜索 API，同时返回数据集中的一致结果。通过此模拟 API，研究人员可以在多次实验中执行相同的检索操作，确保对不同大语言模型配置、提示策略和验证方法进行公平比较。模拟 API 可通过 https://factcheck-api.dei.unipd.it/ 访问。完整文档可在 GitHub 上查阅。[4]

4.2 模型

我们将四种参数规模为 7–9B 的开源大语言模型整合为知识图谱事实验证流程的骨干模型：Gemma2、Qwen2.5、Mistral 和 Llama3.1。我们优先选择开源模型，原因有以下几点。首先，它们能够部署在各种环境中，包括对数据隐私要求严格或 API 访问受限的场景，因为它们可以托管在本地，无需依赖外部服务。其次，它们具有更高的可调节性，可以针对特定领域的数据进行微调，或适配专门的事实验证任务。第三，对于大规模应用而言，它们的成本效益显著更高，能够避免按 token 计费的 API 成本；在处理大型知识图谱时，这类费用可能高得令人难以承受。为了提供性能参考，并评估开源方案与商业方案之间的差距，我们还纳入了 OpenAI 的商业模型 GPT-4o mini。

Gemma2:9B由谷歌开发，是一款开源的 9B 参数模型，经过效率优化， [13],在自然语言理解和生成方面表现出色。

[3]https://huggingface.co/datasets/FactCheck-AI/FactCheck [4] https://github.com/FactCheck-AI/FactCheck-MockAPI

Qwen2.5:7B来自阿里云，是一款开源的 7B 参数模型，在指令遵循、推理和结构化数据处理方面表现突出 [52, 68]。

LLaMA3.1:8B由Meta推出，是一款开源的80亿参数模型，拥有超大的128k令牌上下文窗口，并增强了多语言支持，因此适用于长上下文和多种语言任务 [9]。

Mistral:7B由Mistral AI开发，是一款700亿参数模型，以兼顾性能与紧凑性而闻名，这一点已在多种基准测试中得到验证 [25]。

GPT-4o mini由OpenAI开发，是GPT-4o的较小版本，具备强大的推理能力，同时延迟和成本更低 [44], ，可作为先进知识检索与事实验证的商业基线。

4.3 性能指标与评估

为了评估所考虑的事实验证策略的有效性，我们重点关注两个关键指标：分类别 F1 分数和共识一致性。选择这些指标是为了兼顾类别不平衡、衡量分类别性能，并评估多模型共识方法的一致程度。我们还通过计算各策略给出验证答复所需的平均响应时间来评估效率。

分类别 F1 分数 ($F1(c)$) 分别针对“真”（T）和“假”（F）标签独立计算，以评估模型在各个类别上的表现，而非将其汇总。这样的细粒度视角能够凸显模型在这两个类别上的性能差异。给定类别 $c \in \{T, F\}$ 的 $F1$ 分数定义如下：

$$F1(c) = \frac{2 \cdot \text{Precision}(c) \cdot \text{Recall}(c)}{\text{Precision}(c) + \text{Recall}(c)},$$

其中，Precision(c) 和 Recall(c) 分别表示针对类别 c 计算的精确率和召回率。

共识一致性 (CA_M) 用于量化给定模型的预测与所有已评估事实的多数投票结果之间的一致程度。具体而言，对于模型 M，其定义如下：

$$\text{CA}_M = \frac{1}{|G|} \sum_{t \in G} \mathbb{I}(\text{response}(M, t) = \text{majorityVote}(t))$$

其中，$\mathbb{I}(\cdot)$ 表示指示函数，若条件成立则取值为 1，否则取值为 0。这里，response(M, t) 表示模型 M 对三元组 t的预测，而 majorityVote(t) 是由集成中多数模型赋予的标签。CA_M 分数的范围为 0 到 1。较高的 CA_M 值表示该模型是“最具代表性”的模型，可作为最能代表群体共识的单一代理；较低的 CA_M 值则表示该模型是“离群模型”。这表明该模型会系统性地偏离多数意见。

为评估**效率**，我们测量事实的平均响应时间（以秒为单位），记为 $\bar{\theta}$。为确保评估稳健且不受极端值影响，我们采用基于四分位距法的离群值剔除流程。给定一个模型-数据集对，设 $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ 为模型对该数据集的 n 个事实的响应时间集合。我们先计算第一四分位数 $Q_1 = P_{25}(\Theta)$ 和第三四分位数 $Q_3 = P_{75}(\Theta)$，然后得出四分位距 $= Q_3 - Q_1$。最后，将可接受值的下限和上限分别定义为 $L_{\text{lower}} = Q_1 - 1.5 \times \text{IQR}$ 和 $L_{\text{upper}} = Q_3 + 1.5 \times \text{IQR}$。我们剔除所有超出这些界限的响应时间，得到筛选后的集合 $\Theta' = \{\theta \in \Theta \mid L_{\text{lower}} \le \theta \le L_{\text{upper}}\}$。随后，通过计算筛选后集合的平均响应时间，得到每个事实的平均响应时间：$\bar{\theta} = \frac{1}{|\Theta'|} \sum_{\theta \in \Theta'} \theta$。 ___

表4：检索增强生成流水线中使用的配置参数。

RAG组件	参数
易于人类理解的文本	Gemma2:9b
问题生成	Gemma2:9B
问题相关性	Jina-reranker-v1-turbo-en
相关性阈值	0.5
选定的问题	3
选定的文档 (k_d)	10
文档选择	ms-marco-MiniLM-L-6-v2
嵌入模型	bge-small-en-v1.5
分块策略	滑动窗口（大小为 = 3）

5 实验设置

本节详细介绍实现 FactCheck 所采用的技术规格、计算基础设施和方法框架。我们将说明硬件环境、模型配置以及流程规程。

为检索 Google 搜索结果页，我们使用了一台配备 2 个 CPU 核心和 4 GB 内存的基于 Unix 的服务器。对于三元组转换和问题生成，我们使用了一台搭载 Apple M2 Max 芯片、配备 32 GB 内存的 MacBook Pro。所有其他涉及大语言模型的实验，包括提示和评估，均在一台 Mac Studio（型号：Mac14,14）上进行。该设备搭载 Apple M2 Ultra 芯片，拥有 24 个核心（16 个性能核心和 8 个能效核心）以及 192 GB 统一内存。

我们使用 *Ollama* 在本地运行开源大语言模型，[5] 这是一个开源框架，可简化在本地机器上部署和使用大语言模型的流程。为了监控模型行为（包括令牌用量和推理时间），我们集成了 OpenTelemetry，并借助 OpenLIT 项目的工具实现监控。[6] 这一设置可有效监控大语言模型、向量数据库和 GPU 的使用情况。

检索增强生成流水线的配置参数见表4。这些设置是通过一系列对比不同配置的实验确定的。消融研究的结果可在 GitHub 仓库中查看。[7]

对于多模型共识，我们设置了两种不同的实验场景：一种使用参数规模较大的开源模型，另一种使用商用大语言模型，如 §3.3 所述。在开源场景中，我们计算各数据集上的模型一致性后，选取一致性得分最高和最低的模型。随后，我们将基础版本替换为更大的对应版本：LLaMA3.1（8B → 70B）、Gemma2（9B → 27B）、Qwen2.5（7B → 14B）和 Mistral（7B → nemo:12B）。在商用基线场景中，我们使用 OpenAI GPT-4o mini，为与开源替代方案进行比较提供了有力的参照。

6 实验分析

在本节中，我们全面评估大语言模型在 FactCheck 基准测试上的性能，考察其进行知识图谱事实验证的能力。表 5 和表 7 分别报告了 $F1$ 分数

[5]https://ollama.com/ [6]https://openlit.io/ [7] https://github.com/FactCheck-AI/FactCheck/blob/main/extra-experiments/ 消融_研究_结果/README.md

表5：事实核查系统的性能评估。评估涵盖多种方法（DKA、GIV-Z、GIV-F、检索增强生成）。每列中表现最佳的方法以粗体突出显示，表现第二好的方法以下划线标出。

数据集	方法	Gemma2		Qwen2.5		Llama3.1		Mistral		GPT-4o mini	
		$F1(T)$	$F1(F)$	$F1(T)$	$F1(F)$	$F1(T)$	$F1(F)$	$F1(T)$	$F1(F)$	$F1(T)$	$F1(F)$
FactBench	DKA	0.75	0.74	0.55	0.71	0.73	0.74	0.68	0.73	0.52	0.72
	GIV-Z	0.73	0.73	0.51	0.70	0.52	0.70	0.77	0.72	0.48	0.71
	GIV-F	0.79	0.76	0.74	0.73	0.75	0.72	0.81	0.73	0.49	0.71
	RAG	**0.91**	**0.89**	**0.89**	**0.85**	**0.83**	**0.80**	**0.87**	**0.82**	**0.91**	**0.90**
Mean		0.80	0.78	0.67	0.75	0.71	0.74	0.78	0.75	0.60	0.76
YAGO	DKA	0.82	0.02	0.42	0.02	0.71	0.02	0.59	0.01	0.48	0.02
	GIV-Z	0.88	**0.03**	0.53	0.02	0.52	0.02	0.75	**0.02**	0.51	0.02
	GIV-F	**0.92**	0.02	0.72	**0.03**	0.83	0.02	0.90	0.01	0.53	0.02
	RAG	**0.92**	**0.03**	**0.92**	**0.03**	**0.91**	0.02	**0.96**	**0.02**	**0.89**	0.02
Mean		0.89	0.03	0.65	0.03	0.74	0.02	0.80	0.02	0.60	0.02
DBpedia	DKA	**0.85**	0.36	0.63	0.33	**0.81**	0.29	0.79	0.34	0.56	0.31
	GIV-Z	0.81	0.37	0.63	0.33	0.53	0.31	0.87	0.23	0.48	0.31
	GIV-F	**0.85**	0.35	0.78	0.36	0.69	0.32	**0.89**	0.20	0.36	0.30
	RAG	0.79	**0.38**	**0.82**	**0.39**	0.74	**0.33**	0.82	**0.38**	**0.75**	**0.37**
Mean		0.83	0.37	0.72	0.35	0.69	0.31	0.84	0.29	0.54	0.32

这些分数针对 Fact- Bench、YAGO 和 DBpedia 数据集中的每个模型，分别按真、假标签列出。本分析围绕前文提出的三个关键研究问题展开。

研究问题***1.*** 表5概述了大语言模型内部知识能力的评估结果。分析采用了三种验证范式：直接知识评估（DKA），以及零样本（GIV-Z）和少样本（GIV-F）情境下的引导式迭代验证。

我们观察到不同模型和数据集之间存在明显的性能多样性。在 FactBench 数据集中，Gemma2 在两类任务中均展现出稳健的能力，在 GIV-F 设置下，$F1(T)$ 达到 0.79，$F1(F)$ 达到 0.76。相比之下，GPT-4o mini 表现出明显的性能不对称性。尽管其检测错误事实的能力与其他模型相当 $F1(F) \approx 0.71$，但其验证真实事实的能力始终较弱 $F1(T)$ $[0.48, 0.52]$ 。

这一发现对商业模型或较大型模型优于较小型或开源模型的普遍观点提出了质疑。

在这些数据集中，FactBench 似乎最适合评估内部知识，因为大多数模型在 $F1(T)$ 与 $F1(F)$ 之间保持了合理的平衡。另一方面，由于正确事实数量庞大，YAGO 被证明最具挑战性。尽管模型取得了较高的 $F1(T)$ 分数（最高达 0.92），但 $F1(F)$ 分数却微乎其微（0.01 至 0.03）。这种巨大差异表明模型存在明显偏向正类分类的模型偏差，从而妨碍了在高度不平衡的情境中检测罕见错误事实。相比之下，DBpedia 的结果处于中间水平；大多数模型的 $F1(T)$ 分数尚可 $[0.53, 0.89]$，但它们难以可靠地识别错误信息，$F1(F)$ 值通常低于 0.40。

值得注意的是，少样本设置（GIV-F）的表现始终优于 DKA 和 GIV-Z 设置。例如，在 FactBench 上，Mistral 的表现从 0.68（DKA）提升至 0.81（GIV-Z），而其在虚假主张上的表现则稳定在 0.73 左右。这些提升在中等水平的模型上尤为明显，因为它们从结构化提示和基于示例的引导中获益更多。相比之下，Gemma2 等本已表现出色的模型，性能提升相对较小。

发现 1： Gemma2 或 Mistral 等开源模型在仅依赖内部知识时，表现优于 GPT-4o mini 等商业替代方案。此外，少样本提示始终能提升表现，但提升幅度会受到数据集特征的影响，例如类别平衡和标签分布。

研究问题 ***2.*** 我们评估了检索增强生成方法在所有模型和数据集上的表现，并将其与基于内部知识的方法进行比较，结果见表5。

总体而言，检索增强生成在几乎所有实验设置中都取得了最高性能。特别是在 FactBench 数据集上，检索增强生成带来了显著提升：例如，在 DKA 设置下，Qwen2.5 的 $F1(T)$ 达到 0.89，而 DKA 设置下的数值为 0.55。这一趋势在所有受评估模型中均成立，包括 GPT-4o mini；引入外部证据后，其性能显著提升，两个 $F1$ 分数均提高了 25% 以上。

然而，检索增强生成的影响因数据集而异。FactBench 和 YAGO 的绝对增益最大，这可能是因为它们涵盖的事实内容更加多样。相比之下，Dbpedia 的提升微乎其微，某些情况下性能甚至略有下降。这可能归因于模式多样性，它会使检索过程更加复杂，并降低提取的证据的相关性。

发现 2： 通过检索增强生成引入外部证据，是实现高准确率事实验证的一条有前景的途径。然而，其效果取决于数据集特征。

研究问题***3***。 我们研究多模型共识策略的有效性，在四个开源模型之间采用多数投票。在出现平局的情况下，我们引入平局判定机制，使用更高参数版本或商业模型（GPT-4o mini）进行判定。表7总结了结果。

在内部知识设置（DKA、GIV-Z 和 GIV-F）中，多模型共识的表现更可靠，尽管并非总能胜过所有单个模型。在许多情况下，它能在不同条件下稳定表现，而非取得最佳结果。有趣的是，平局判定模型的选择对最终表现影响甚微。无论我们使用一致性最高的模型（agg-cons-up），还是

表6：不同事实验证方法和数据集上的模型对齐分析。共识一致性（CA_M）衡量大语言模型预测与多数投票决策的一致率，其中最高和 标出各方法—数据集组合中表现最差的模型。平局百分比表示需要仲裁的分歧决策频率。

数据集	方法	Ties	Gemma2	Qwen2.5	Llama3.1	Mistral
FactBench	DKA	16%	0.919	0.861	0.906	0.938
	GIV-Z	21%	0.914	0.893	0.913	0.814
	GIV-F	14%	0.937	0.861	0.901	0.909
	RAG	6%	0.968	0.970	0.897	0.960
YAGO	DKA	19%	0.798	0.797	0.916	0.920
	GIV-Z	26%	0.790	0.872	0.859	0.886
	GIV-F	16%	0.934	0.771	0.901	0.944
	RAG	6%	0.968	0.969	0.916	0.974
DBpedia	DKA	17%	0.937	0.772	0.891	0.920
	GIV-Z	24%	0.948	0.875	0.765	0.758
	GIV-F	17%	0.960	0.879	0.779	0.876
	RAG	9%	0.953	0.961	0.848	0.945

表7：事实验证系统的性能评估。该评估涵盖多模型共识。在每一列中，性能最佳的方法以粗体突出显示，性能第二佳的方法加下划线。

数据集	方法	agg-cons 上升（参见表6）		agg-cons 下降（参见表6）		agg-GPT-4o mini	
		$F1(T)$	$F1(F)$	$F1(T)$	$F1(F)$	$F1(T)$	$F1(F)$
FactBench	DKA	0.68	0.75	0.69	0.75	0.69	0.75
	GIV-Z	0.74	0.76	0.64	0.74	0.63	0.74
	GIV-F	0.82	0.78	0.81	0.79	0.80	0.79
	RAG	**0.91**	**0.89**	**0.91**	**0.89**	**0.91**	**0.89**
Mean		0.79	0.80	0.76	0.79	0.76	0.79
YAGO	DKA	0.59	**0.02**	0.63	**0.02**	0.61	**0.02**
	GIV-Z	0.63	**0.02**	0.73	**0.02**	0.65	**0.02**
	GIV-F	0.84	**0.02**	0.84	**0.02**	0.84	**0.02**
	RAG	**0.93**	**0.02**	**0.94**	**0.02**	**0.93**	**0.02**
Mean		0.75	0.02	0.78	0.02	0.76	0.02
DBpedia	DKA	0.84	0.37	0.80	0.37	0.78	0.37
	GIV-Z	0.77	0.38	0.73	0.36	0.71	0.36
	GIV-F	**0.85**	**0.40**	**0.86**	**0.39**	**0.81**	0.38
	RAG	0.80	0.39	0.81	**0.39**	0.80	**0.39**
Mean		0.81	0.39	0.80	0.38	0.77	0.38

一致性最低的模型（agg-cons-down）或 GPT-4o mini，所得分数在所有数据集和方法中仍几乎相同。这表明，多数投票机制有效地捕捉了最可靠的信号，而仲裁者的具体选择不如采用一致的平局解决策略重要。

表6所示的一致性分析进一步表明，模型间的一致程度会随着方法复杂度的提高而增加。例如，与 GIV-Z 中 21% 到 26% 的平局率相比，检索增强生成的平局率较低，为 6% 到 9%。这进一步印证了外部证据不仅能提升单个模型的性能，还能增强跨模型对齐。不过，这种一致程度的提高也可能反映了共享上下文证据的影响力增强，从而可能减少对内部知识的依赖，并因此以牺牲模型个体性或特异性为代价，导致结果趋于一致。

表8：事实验证在不同方法（DKA、GIV-Z、GIV-F 和检索增强生成）下的执行时间（$\bar{\theta}$，单位为秒）。最快的配置以绿色突出显 慢的配置则以红色标记。

数据集	方法	Gemma2	Qwen2.5	Llama3.1	Mistral
FactBench	DKA	0.21	0.18	0.30	0.17
	GIV-Z	0.62	0.40	0.50	0.45
	GIV-F	0.78	0.51	0.67	0.65
	RAG	2.27	2.39	2.73	1.69
YAGO	DKA	0.22	0.19	0.31	0.19
	GIV-Z	0.62	0.41	0.45	0.47
	GIV-F	0.78	0.54	0.69	0.67
	RAG	2.10	2.39	2.68	1.63
DBpedia	DKA	0.35	0.25	0.37	0.24
	GIV-Z	0.70	0.43	0.58	0.53
	GIV-F	0.89	0.56	0.69	0.78
	RAG	2.55	2.55	2.87	1.77

发现 3： 多模型共识提供了一种简单而稳健的机制，可稳定事实验证性能。虽然它并非总能胜过单个模型，但可以减轻较弱模型带来的影响。仲裁者的具体选择影响有限。此外，外部证据有助于提升模型对齐，但应谨慎避免对上下文偏差过度拟合。

计算效率。除了准确率指标外，我们还评估了不同方法的计算效率。表8报告了四种开源大语言模型在三个参考数据集上进行事实验证的执行时间（$\bar{\theta}$，单位为秒）。在每个数据集中，DKA 的执行时间均最低：在 FactBench 上为 0.21 至 0.30 秒，在 YAGO 上为 0.19 至 0.31 秒，在 DBpedia 上为 0.24 至 0.37 秒。GIV-Z 的执行时间高于 DKA，在 FactBench 和 YAGO 上约为其两倍，例如，在 FactBench 上使用 Qwen2.5 时，执行时间从 0.18 秒增加到 0.40 秒。GIV-F 比 GIV-Z 耗时更长，最高达到 0.78 秒。检索增强生成在所有数据集和模型上的执行时间均最高，例如，Llama3.1 在 FactBench 上耗时 2.73 秒，DBpedia 上有多个模型的耗时超过 2.5 秒。

对各数据集进行比较后发现，正如预期，检索增强生成的计算成本最高，通常是 DKA 的六倍甚至更多。在所有配置中，执行时间均呈现出从 DKA 到 GIV-Z、再到 GIV-F，最后到检索增强生成逐渐增加的趋势。这一模式表明，验证策略的方法复杂度与其计算成本之间存在直接关系。

另一方面，多模型共识可以并行执行，这意味着推理延迟取决于速度最慢的模型，而非所有模型耗时之和。实践中，如果模型的响应时间各不相同（例如 0.3–0.5 秒），共识推理就需要等待最慢的响应，因此与仅选择最快模型相比，延迟会略高。平局判定还会增加推理开销，因为它需要额外查询一个模型。此外，协调多个模型并进行资源分配会带来轻微但不可忽视的计算开销。尽管如此，共识也有其优势：通过汇集不同的模型视角，预测的可信度得以提高。

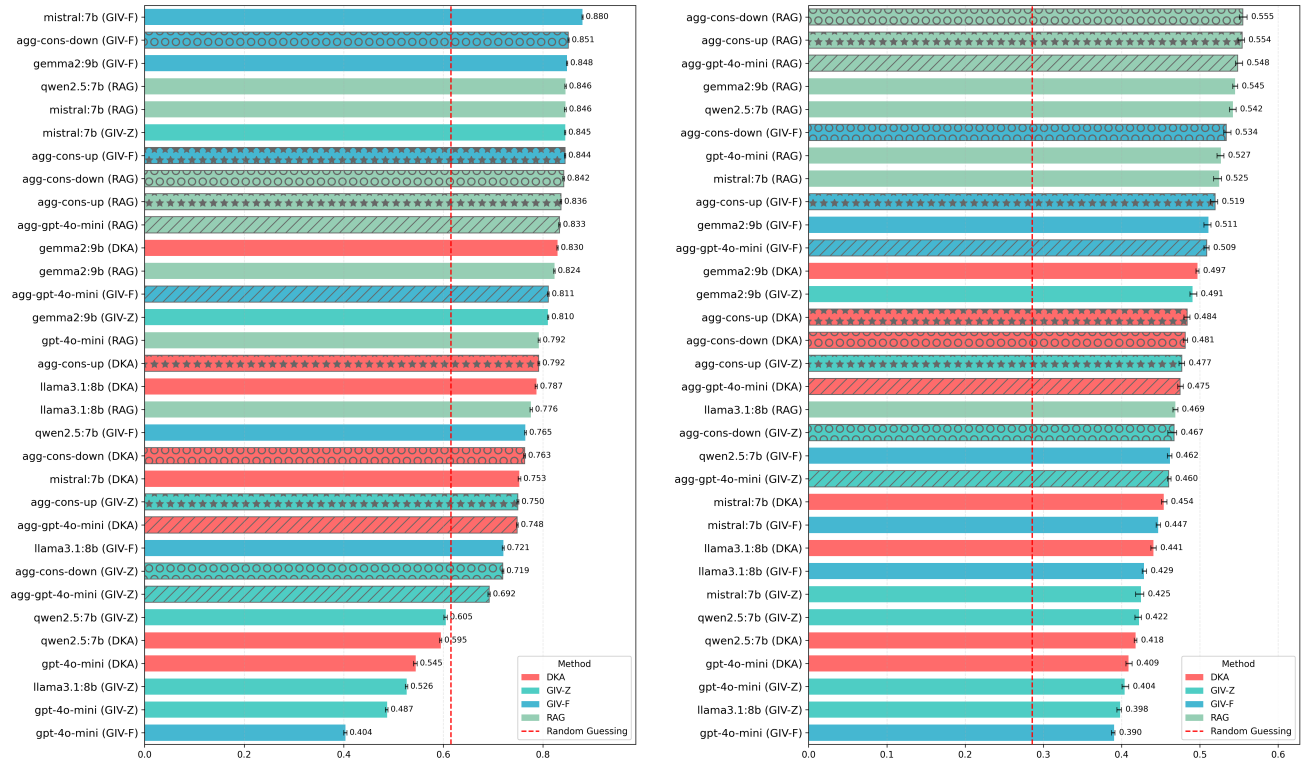


图2： $F1$ **得分多模型共识** **FactCheck 基准测试。左图显示** $F1(T)$ **得分，右图显示** $F1(F)$ **得分结果以阴影线表示，红色虚线表示猜测率。**

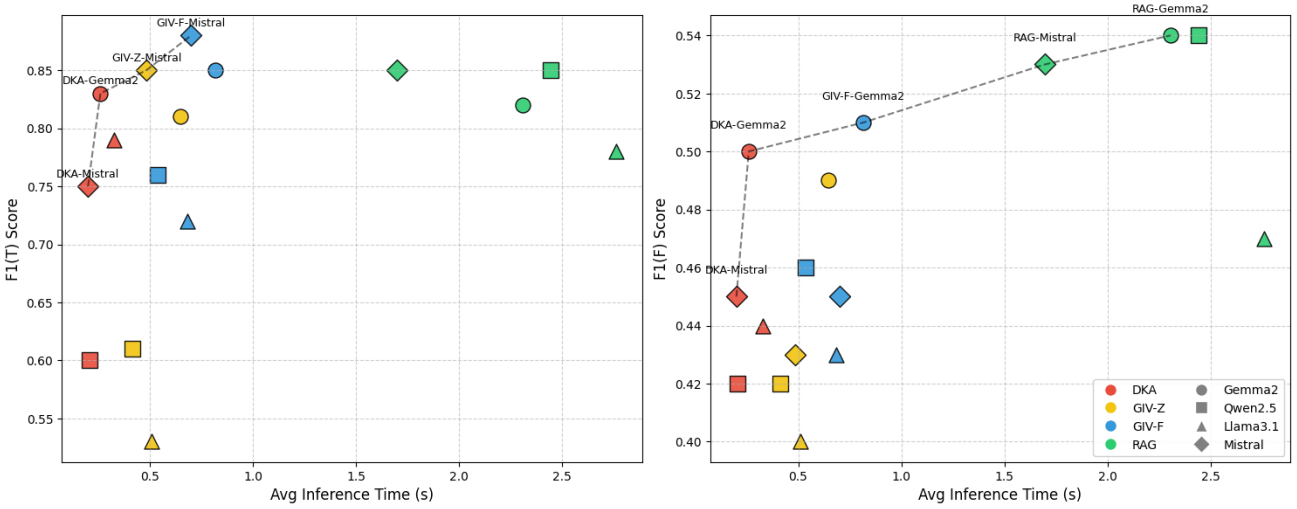


图3：计算成本（$\bar{\theta}$）与验证性能（$F1(F)$和 $F1(T)$）之间的权衡分析。虚线表示帕累托前沿，突出显示能够实现最佳效率的配置（在给定时间预算内达到最高准确率）。

为了刻画预测准确率与计算开销之间的平衡，我们考察了各种模型中各方法的帕累托效率（图3）。分析显示，各策略的效用存在明显差异：基于检索增强生成的技术通常聚集在右上象限，尤其是在 $F1(F)$ 指标上，这表明它们以更高的延迟（≈1.6s–2.9s）换取了对虚假主张更强的检测能力。相反，DKA 配置则占据主导地位

高速模式下，推理时间不到一秒（<0.3 s），适用于延迟敏感型应用场景，但敏感度较低。帕累托前沿表明，中等规模的方法（例如 GIV-F，尤其是与 Gemma2 和 Mistral 搭配使用时）实现了颇具吸引力的权衡：准确率具有竞争力，有时甚至在 $F1(T)$ 指标上超过检索增强生成，同时计算成本远低于完整的基于检索的系统。

发现 4： 不同方法的计算效率差异很大。一方面，检索增强生成所需的处理时间最多比内部知识方法高出 10× 。另一方面，共识策略可以并行化，从而确保相较于内部知识方法，延迟仅小幅增加。

跨数据集泛化与稳定性。 为评估基于大语言模型的事实验证能力的泛化能力和稳定性，我们分析了不同方法和聚合策略的表现，结果以柱状图呈现（图2）。图中按表现对真类（左图）和假类（右图）的 $F1$ 分数进行了排序。红色虚线表示随机猜测基线；对于 $F1$（T），其值约为0.62；对于 $F1$（F），其值约为0.29，这反映了数据集内在的类别分布挑战。

检索增强生成展现出最稳定的鲁棒性。在$F1$（F）图中，该图通常呈现更困难的任务，即识别错误事实；基于检索增强生成的方法及其聚合结果在排名中占据前列。另一方面，GIV-F（蓝色柱）表现出较高的方差。尽管Mistral（GIV-F）在 $F1$（T）图中达到绝对最高峰（0.88），但采用相同策略的其他模型（如gpt-4o-mini）表现则低得多，仅为0.40。这一结果显著低于随机猜测基线，表明GIV-F虽然能使特定模型对真实事实达到较高召回率，但其稳定性不及检索增强生成。DKA（红色柱）方法总体处于中下游，尤其是在 $F1$（F）分析中，这表明仅依赖内部参数化知识通常不足以区分虚假主张。最后，标记为“agg-cons-∗”的聚合方法始终位居两张图的前列。这证实了集成推理，尤其是多数投票策略，能够有效降低单个模型的波动，并平滑GIV-Z和GIV-F等策略中观察到的噪声。

发现 5： 检索增强生成在跨数据集泛化方面表现最强，在检测虚假主张时始终优于基于内部知识的方法。一些 GIV-F 模型在判断真实事实时表现最佳，但结果波动很大。值得注意的是，若干基于内部知识的方法表现低于随机猜测，这表明糟糕的方法可能会使推理能力降至抛硬币基线以下。因此，基于共识的聚合对于保持稳定性和减少模型特定偏差仍至关重要。

7 定性错误分析

在错误分析中，我们通过结合大语言模型生成的推理与上下文文档嵌入的半自动化流程，对开源模型产生的错误进行分类。我们收集错误预测的日志，并提示同一个大语言模型解释每个错误。随后，我们使用 cde-small-v1 模型 [39] 对这些解释进行编码，并先通过 UMAP 降低维度，再使用 HDBSCAN [6] 进行聚类，以找出密度各异的簇。最后，我们为每个簇赋予描述性标签。由此得到的错误类别如下：未标记（E1）：提供的上下文缺少所断言的细节或相关实体的提及。关系错误（E2）：模型提供了有关个人之间关系的错误信息，例如婚姻状况或宗教信仰。角色归属错误（E3）：模型错误地将人物与特定角色联系起来，地点或团队。地理/国籍错误（E4）：有关地点或国家隶属关系的信息与上下文不一致。类型/分类错误（E5）：模型错误地对与个人或工作室相关的电影、类型或创意作品进行分类。标识符/传记信息错误（E6）：标识符或传记事实（例如奖项名称）不准确。

表9：基于大语言模型生成的推理，按数据集划分的错误聚类。

数据集	模型	E1	E2	E3	E4	E5	E6	总计*
FactBench	Gemma2	4	36	45	176	13	1	275
	Qwen2.5	33	27	60	194	34	1	349
	Llama3.1	38	44	73	295	38	3	491
	Mistral	53	27	53	242	40	2	417
唯一性比率（%）		0.62	0.72	0.44	0.52	0.63	0.57	0.53
YAGO	Gemma2	6	134	0	14	51	2	207
	Qwen2.5	7	109	0	13	63	2	194
	Llama3.1	8	98	0	19	104	2	231
	Mistral	7	54	0	10	34	3	108
唯一性比率（%）		0.35	0.52	–	0.46	0.51	0.33	0.50
DBpedia	Gemma2	353	22	98	1729	459	299	2960
	Qwen2.5	339	19	91	1525	357	237	2568
	Llama3.1	382	28	109	2172	509	318	3518
	Mistral	325	20	94	1487	438	241	2605
唯一性比率（%）		0.41	0.43	0.44	0.42	0.42	0.40	0.41

表9显示了各个错误类型在所评估数据集中的数量。如表9所示，E4错误是FactCheck中最主要的挑战。此外，我们还使用 Marchesin 等人 [38] 提出的分层方法和主题建模，在 DBpedia 数据集上扩展了这项分析，以了解事实的流行度和领域所产生的影响。结果显示，在代表常识的分区中，错误率有所降低；“教育”和“新闻”等领域的错误率较低，而“建筑”和“交通运输”则仍然更具挑战性。本文介绍的整个验证过程和错误分析，可以通过我们位于 https://factcheck.dei.unipd.it/ [55]的网络平台进行交互式解读和可视化。

为了研究这些模型如何相互补充，我们使用 UpSet 图 [35] 考察了它们预测的重叠情况。如图4所示，最大的交集通常对应于四个模型均正确预测的事实，这表明开源大语言模型在内部知识及其错误特征上有很大共通之处。这种一致性在检索增强生成设置中最为明显，共同的外部证据会引导模型得出相同结论，从而降低方差。

然而，GIV-Z 偏离了这一模式：“全模型”交集相对于 DKA 显著缩小（从约 4,600 降至约 3,200），并被更强的成对重叠所取代（例如 Qwen2.5 与 Gemma2 之间）。这一模式表明，零样本提示会导致更加异质的推理轨迹，以及模型之间更大的分歧。相比之下，GIV-F 恢复了更强的共识，将全模型交集提升至 5,200 以上，表明少样本示例能够有效协调模型行为。总体而言，模型之间有限的真正互补性或许可以解释，为何共识方法能够稳定预测，却很少超越最佳的单一模型。

8 结语

在本研究中，我们推出了 FactCheck，这是一个用于系统评估大语言模型在知识图谱事实验证方面表现的基准测试。我们的评估

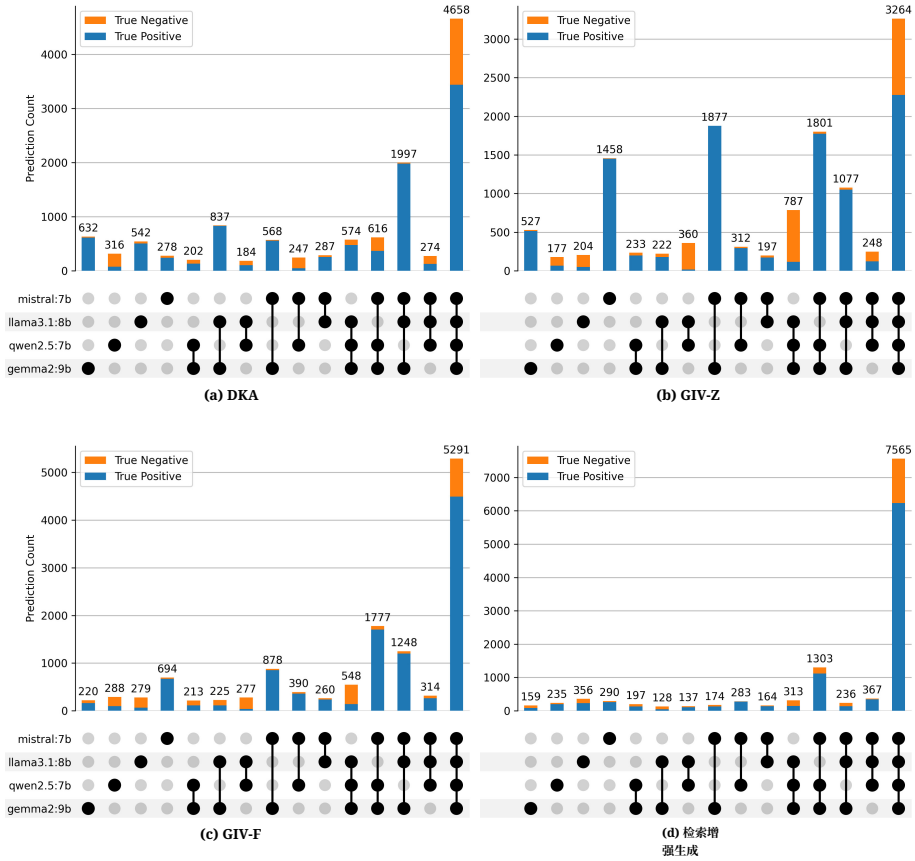


图4：各模型正确预测的交集。柱状图展示了由下方相连圆点所示的特定模型组合对应的正确样本数量。

基于 FactCheck 所含的三个真实世界数据集——FactBench、YAGO 和 DBpedia——开展评估，得出了几项重要发现。首先，Gemma2 等开源大语言模型展现出良好的验证性能，仅依靠内部知识时，$F1$ 分数最高可达 0.79 和 0.76；结合检索增强生成后则超过 0.89。其次，检索增强生成在大多数设置下都能提升性能，但计算成本显著增加——速度约比其他方法慢 10×。第三，与单个模型的预测相比，多模型共识能够减少错误并提供更可靠的回答，尤其是在依赖内部知识时。

与此同时，我们也发现了若干局限：（1）数据集特定挑战，例如 YAGO 中的类别不平衡以及 DBpedia 中的模式多样性；（2）基础设施限制，包括因网络问题和区域限制导致的 0.08% 检索失败率；以及（3）托管部署中的内容过滤，例如 Azure 的 GPT-4o-mini 会屏蔽敏感主题的事实性内容。

因此，FactCheck 利用知识图谱的结构化语义推进了对大语言模型事实推理的研究，这有别于以往聚焦于非结构化主张或通用领域问答的基准测试。它提供了一个受控环境，可对模型行为进行可复现、细粒度的分析，包括内部知识的使用、检索有效性和多模型交互。作为一个稳健的测试平台，FactCheck 支持开发用于事实验证的新提示策略、模型架构和检索技术。通过公开发布该平台，我们旨在促进透明度与协作，并加快构建可信、可扩展的知识图谱验证系统。

展望未来，我们的研究结果揭示了若干有前景的研究方向。首先，对用于知识图谱事实验证的大语言模型进行微调或预训练，有助于缓解数据集不平衡带来的局限。其次，将结构化的知识图谱遍历与非结构化网络数据相结合的混合检索策略，可能提升检索质量，尤其适用于 DBpedia 等数据集。最后，可以扩展该基准测试，以支持对同样利用知识图谱中逻辑规则的事实核验系统进行评估，例如利用知识图谱所基于的本体（如运用传递性、定义域/值域约束及其他属性来评估三元组的正确性和可靠性）。

致谢

本研究部分由 HEREDITARY 项目资助，该项目属于欧盟 Horizon Europe 研究与创新计划，资助协议编号为 GA 101137074。作者感谢 Andrea Segala 在其硕士论文期间为零样本和少样本提示实验所作的贡献。

工件

源代码和数据集已公开发布于 https://github.com/FactCheck-AI/ 和 https://huggingface.co/ FactCheck-AI。

参考文献

[1] Firoj Alam、Julia Maria Struß、Tanmoy Chakraborty、Stefan Dietze、Salim Hafid、Katerina Korre、Arianna Muti、Preslav Nakov、Federico Ruggeri、Sebastian Schellhammer 等。2025年。CLEF-2025 CheckThat! 实验室：主观性、事实核查、声明规范化与检索。载于欧洲信息检索会议。施普林格，467–478。[2] Fatma Arslan、Naeemul Hassan、Chengkai Li 和 Mark Tremayne。2020年。值得核查的事实声明基准数据集。载于*AAAI*国际网络与社交媒体会议论文集，第14卷。821–829。[3] Luigi Bellomarini、Daniele Fakhoury、Georg Gottlob 和 Emanuel Sallinger。2019年。知识图谱与企业人工智能：赋能技术的前景。载于*2019*年第*35*届*IEEE*国际数据工程会议（*ICDE*）。

26 37 10 1109 2019 00011 – 。doi: 。/ICDE. 。Kurt Bollacker、Colin [4] Evans、Praveen Paritosh、Tim Sturge 和 Jamie Taylor。2008年。Freebase：一个为组织人类知识而协作创建的图数据库。*SIGMOD*论文集，

1247 1250 10 1145 1376616 1376746 – 。doi: 。/ 。Jack [5] Boylan、Shashank Mangla、Dominic Thorn、Demian Gholipour Gha- landari、Parsa Ghaffari 和 Chris Hokamp。2024年。KGValidator：知识图谱构建自动验证框架。arXiv:

2404 15923 [ ] 2404 15923 . cs.AI https://arxiv.org/abs/ 。R. [6] J. G. B. Campello、D. Moulavi 和 J. Sander。2013。。基于层次密度估计的基于密度的聚类。载于知识发现与数据挖掘进展。施普林格 柏林 海德堡，柏林，海德堡，160–172。[7] Zhijun Chen、Jingzheng Li、Pengpeng Chen、Zhuoran Li、Kai Sun、Yuankai Luo、Qianren Mao、Dingqi Yang、Hailong Sun 和 Philip S. Yu。2025年。善用多个大语言模型：大语言模型集成综述。*CoRR* abs/

2502 18036 2025 10 48550 2502 18036 。（2月）。 [8] https://doi.org/ 。/arXiv. 。O. Deshpande、D. S. Lamba、M. Tourn、S. Das、S. Subramaniam、A. Rajaraman、V. Harinarayan 和 A. Doan。2013年。构建、维护和使用知识库：一线报告。载于*ACM SIGMOD*国际数据管理会议论文集，*SIGMOD 2013*，纽约，美国纽约州，*6*月*22*日至*27*日

，2013。美国计算机协会 ，1209–1220。doi:10.1145/2463676.2465297 [9] Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ah- mad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、An- gela Fan、Anirudh Goyal、Anthony Hartshorn、Aobo Yang、Archi Mitra、Archie Sravankumar、Artem Korenev、Arthur Hinsvark、Arun Rao、Aston Zhang、Au- rélien Rodriguez、Austen Gregerson、Ava Spataru、Baptiste Rozière、Bethany Biron、Binh Tang、Bobbie Chern、Charlotte Caucheteux、Chaya Nayak、Chloe Bi、Chris Marra、Chris McConnell、Christian Keller、Christophe Touret、Chun- yang Wu、Corinne Wong、Cristian Canton Ferrer、Cyrus Nikolaidis、Damien Allonsius、Daniel Song、Danielle Pintz、Danny Livshits、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、Frank Zhang、Gabriel Synnaeve、Gabrielle Lee、Georgia Lewis Anderson、Graeme Nail、Grégoire Mialon、Guan Pang、Guillem Cucurell、Hailey Nguyen、Hannah Korevaar、Hu Xu、Hugo Touvron、Iliyan Zarov、Imanol Arrieta Ibarra、Isabel M. Kloumann、Ishan Misra、Ivan Evtimov、Jade Copet、Jaewon Lee、Jan Geffert、Jana Vranes、Jason Park、Jay Ma- hadeokar、Jeet Shah、Jel mer 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、KalyanVasudenAlwala、KartikeyaUpasani、Kate Plawiak、Ke Li、Kenneth Heafield、Kevin Stone，以及其他作者。2024年。Llama 3 模型家族。*CoRR* abs/2407.21783（2024年）。 https://doi.org/10.48550/arXiv.2407.21783

[10] Basil Ell、Andreas Harth 和 Elena Simperl。2014年。用于解释语义搜索引擎查询的 SPARQL 查询自然语言表述。载于语义网：趋势与挑战，Valentina Presutti、Claudia d'Amato、Fabien Gandon、Mathieu d'Aquin、Steffen Staab 和 Anna Tordai（编）。施普林格国际出版公司，查姆，426–441。[11] Luis Antonio Galárraga、Christina Teflioudi、Katja Hose 和 Fabian Suchanek。2013年。AMIE：在本体知识库中基于不完整证据进行关联规则挖掘。载于第*22*届万维网国际会议论文集（巴西里约热内卢）（*WWW '13*）。美国计算机协会，纽约，纽约州，美国，

413 422 10 1145 2488388 2488425 – 。doi: 。/。J. [12] Gao、X. Li、Y. E. Xu、B. Sisman、X. L. Dong 和 J. Yang。2019。。高效的知识图谱准确性评估。《*VLDB*基金会会刊》12，11（2019年），1679–1691。 doi:10.14778/3342263.3342642

[13] Gemma 团队、Morgane Riviere、Shreya Pathak、Pier Giuseppe Sessa、Cas sidy Hardin、Surya Bhupatiraju、Léonard Hussenot、Thomas Mesnard、Bobak Shahriari、Alexandre Ramé、Johan Ferret、Peter Liu、Pouya Tafti、Abe Friesen、Michelle Casbon、Sabela Ramos、Ravin Kumar、Charline Le Lan、Sammy Jerome、Anton Tsitsulin、Nino Vieillard、Piotr Stanczyk、Sertan Girgin、Nikola Momchev、Matt Hoffman、Shantanu Thakoor、Jean-Bastien Grill、Behnam Neyshabur、Olivier Bachem、Alanna Walton、Aliaksei Severyn、Alicia Parrish、Aliya Ahmad、Allen Hutchison、Alvin Abdagic、Amanda Carl、Amy Shen、Andy Brock、Andy Coenen、Anthony Laforge、Antonia Paterson、Ben Bastian、Bilal Piot、Bo Wu、Brandon Royal、Charlie Chen、Chintu Kumar、Chris Perry、Chris Welty、Christopher A. Choquette-Choo、Danila Sinopalnikov、David Weinberger、Dimple Vijaykumar、Dominika Rogozińska、Dustin Herbison、Elisa Bandy、Emma Wang、Eric Noland、Erica Moreira、Evan Senter、Evgenii Eltyshev、Francesco Visin、Gabriel Rasskin、Gary Wei、Glenn Cameron、Gus Martins、Hadi Hashemi、Hanna Klimczak-Plucińska、Harleen Batra、Harsh Dhand、Ivan Nardini、Jacinda Mein、Jack Zhou、James Svensson、Jeff Stanway、Jetha Chan、Jin Peng Zhou、Joana Carrasqueira、Joana Iljazi、JocelynBecker、Joe Fernandez、Joost van Amersfoort、Josh Gordon、Josh Lipschultz、Josh Newlan、Ju yeong Ji、Kareem Mohamed、Kartikeya Badola、Kat Black、Katie Millican、Keelin McDonell、Kelvin Nguyen、Kiranbir Sodhia、Kish Greene、Lars Lowe Sjoesund、Lauren Usui、Laurent Sifre、Lena Heuermann、Leticia Lago、Lilly Mc- Nealus、Livio Baldini Soares、Logan Kilpatrick、Lucas Dixon、Luciano Martins、Machel Reid、Manvinder Singh、Mark Iverson、Martin Görner、Mat Velloso、Mateo Wirth、Matt Davidow、Matt Miller、Matthew Rahtz、Matthew Watson、Meg Risdal、Mehran Kazemi、Michael Moynihan、Ming Zhang、Minsuk Kahng、Minwoo Park、Mofi Rahman、Mohit Khatwani、Natalie Dao、Nenshad Bar- doliwalla、Nesh Devanathan、Neta Dumai、Nilay Chauhan、Oscar Wahltinez、Pankil Botarda、Parker Barnes、Paul Barham、Paul Michel、Pengchong Jin、Petko Georgiev、Phil Culliton、Pradeep Kuppala、Ramona Comanescu、Ramona Merhej、Reena Jana、Reza Ardeshir Rokni、Rishabh Agarwal、Ryan Mullins、Samaneh Saadat、Sara Mc Carthy、Sarah Cogan、Sarah Perrin、Sébastien M. R. Arnold、Sebastian Krause、Shengyang Dai、Shruti Garg、Shruti Sheth、Sue Ronstrom、Susan Chan、Timothy Jordan、Ting Yu、Tom Eccles、Tom Henni- gan、Tomas Kocisky、Tulsee Doshi、Vihan Jain、Vikas Yadav、Vilobh Meshram、Vishal Dharmadhikari、Warren Barkley、Wei Wei、Wenming Ye、Woohyun Han、Woosuk Kwon、Xiang Xu、Zhe Shen、Zhitao Gong、Zichuan Wei、Victor Cotruta、Phoebe Kirk、Anand Rao、Minh Giang、Ludovic Peran、Tris Warkentin、Eli Collins、Joelle Barral、Zoubin Ghahramani、Raia Hadsell、D. Sculley、Jeanine Banks、Anca Dragan、Slav Petrov、Oriol Vinyals、Jeff Dean、Demis Hassabis、Ko- ray Kavukcuoglu、Clement Farabet、Elena Buchatskaya、Sebastian Borgeaud、Noah Fiedel、Armand Joulin、Kathleen Kenealy、Robert Dadashi 和 Alek Andreev。2024年。Gemma 2：在实用规模下改进开放语言模型。arXiv:

2408 00118 [ ] 2408 00118 。cs.CL [14] https://arxiv.org/abs/ 。Daniel Gerber、Diego Esteves、Jens Lehmann、Lorenz Bühmann、Ricardo Us- beck、Axel-Cyrille Ngonga Ngomo 和 René Speck。2015年。DeFacto——时序与多语言深度事实验证。《网络语义学杂志》35（2015），85–101。 doi:10.1016/j.websem.2015.08.001 语义网机器学习与数据挖掘（MLDMSW）。[15] Genevieve Gorrell、Elena Kochkina、Maria Liakata、Ahmet Aker、Arkaitz Zu- biaga、Kalina Bontcheva 和 Leon Derczynski。2019年。SemEval-2019 任务7：RumourEval，判定谣言真实性并为谣言提供支持。载于第*13*届语义评估国际研讨会论文集，Jonathan May、Ekaterina Shutova、Aurelie Herbelot、Xiaodan Zhu、Marianna Apid- ianaki 和 Saif M. Mohammad（编）。计算语言学协会，美国明尼阿波利斯，

845 854 10 18653 1 19 2147 –。doi: 。/v / [16] S - R。Guha、Rob McCool 和 Eric Miller。2003。。语义搜索。载于第*12*届万维网国际会议论文集（匈牙利布达佩斯）（*WWW '03*）。美国计算机协会，美国纽约，NY，700–709。 doi:10.1145/775152.775250[17] Zhijiang Guo、Michael Schlichtkrull 和 Andreas Vlachos。2022年。自动化事实核查综述。计算语言学协会汇刊 10（2022），178–206。 [18] Qi He、Bee-Chung Chen 和 Deepak Agarwal。2016年。构建 *LinkedIn* 知识图谱。 LinkedIn 工程。https://engineering.linkedin.com/ blog/2016/10/building-the-linkedin-knowledge-graph 访问日期：2025-04-16。[19] Qiyuan He、Yizhong Wang 和 Wenya Wang。2024年。语言模型能否大规模充当知识库？*CoRR*（2024年）。[20] Qiyuan He、Yizhong Wang、Jianfei Yu 和 Wenya Wang。2025年。大规模知识库中的语言模型：关于新事实的容量、灵活性与推理。载于第*31*届国际计算语言学会议论文集，Owen Rambow、Leo Wanner、Marianna Apidi- anaki、Hend Al-Khalifa、Barbara Di Eugenio 和 Steven Schockaert（编）。计算语言学协会，阿联酋阿布扎比，1736–1753。 https://aclanthology.org/2025.coling-main.118/ [21] Cory Henson、Stefan Schmid、Anh Tuan Tran 和 Antonios Karatzoglou。2019年。利用场景知识图谱实现自动驾驶数据搜索。载于*ISWC*（卫星会议）。313–314。[22] Johannes Hoffart、Fabian M. Suchanek、Klaus Berberich、Edwin Lewis-Kelham、Gerard de Melo 和 Gerhard Weikum。2011年。YAGO2：探索并查询时空、情境及多语言世界知识。载于第*20*届万维网国际会议伴随会议论文集（印度海得拉巴）（*WWW '11*）。美国计算机协会，美国纽约，NY，229–232。doi:10.1145/1963192.1963296[23] Aidan Hogan、Eva Blomqvist、Michael Cochez、Claudia D'amato、Gerard De Melo、Claudio Gutierrez、Sabrina Kirrane、José Emilio Labra Gayo、Roberto Navigli、Sebastian Neumaier、Axel-Cyrille Ngonga Ngomo、Axel Polleres、Sabbir M. Rashid、Anisa Rula、Lukas Schmelzeisen、Juan Sequeda、Steffen Staab 和 Antoine Zimmermann。2021年。知识图谱。美国计算机协会计算调查

54, 4, 71 2021 37 10 1145 3447772 文章（7月），页码。 [24] doi: 。/ Shaoxiong Ji、Shirui Pan、Erik Cambria、Pekka Marttinen 和 Philip Yu。2021年。知识图谱综述：表示、获取与应用。*IEEE* 神经网络与学习系统汇刊 PP（2021年04月）。doi:10. 1109/TNNLS.2021.3070843[25] Albert Q. Jiang、Alexandre Sablayrolles、Arthur Mensch、Chris Bamford、Devendra Singh Chaplot、Diego de las Casas、Florian Bressand、Gianna Lengyel、Guillaume Lample、Lucile Saulnier、Lélio Renard Lavaud、Marie-Anne Lachaux、Pierre Stock、Teven Le Scao、Thibaut Lavril、Thomas Wang、Timothée Lacroix 和 William El Sayed。2023年。Mistral 7B。arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825[26] Xuhui Jiang、Chengjin Xu、Yinghan Shen、Xun Sun、Lumingyuan Tang、Saizhuo Wang、Zhongwu Chen、Yuanzhuo Wang 和 Jian Guo。2023年。论知识图谱的演变

：综述与展望。arXiv:2310.04835 [cs.AI] https://arxiv.org/abs/2310.04835 。Ryo Kamoi、Yusen Zhang、Nan Zhang、[27] Jiawei Han 和 Rui Zhang。2024年。大语言模型何时才能真正纠正自身错误？大语言模型自我纠正的批判性综述。《计算语言学协会汇刊》12 2024 1417 1440 10 1162 _ _00713 （），–。doi: 。/tacl a[28] Mohammed Abdul Khaliq、Paul Yu-Chun Chang、Mingyang Ma、Bernhard Pflugfelder 和 Filip Miletić。2024年。Ragar，你的谬误雷达：使用多模态大语言模型，通过检索增强推理进行政治事实核查。载于第七届事实提取与验证研讨会（FEVER）。280–296。[29] Jiseong Kim 和 Key-sun Choi。2020年。通过知识图谱中基于反例加权的正向与负向证据路径进行无监督事实核查。载于第28届国际计算语言学会议论文集，Donia Scott、Nuria Bel 和 Chengqing Zong（编）。国际计算语言学委员会，西班牙巴塞罗那（线上），1677–1686。doi:10.18653/v1/2020.coling-main.147[30] Jiho Kim、Sungjin Park、Yeonsu Kwon、Yohan Jo、James Thorne 和 Edward Choi。2023年。FactKG：通过知识图谱推理进行事实验证。载于计算语言学协会第61届年会论文集（第1卷：长篇论文），Anna Rogers、Jordan Boyd-Graber 和 Naoaki Okazaki（编）。计算语言学协会，加拿大多伦多，16190 16206 10 18653 1 2023 895 –。doi: 。/v / .acl-long。[31] Aviral Kumar、Vincent Zhuang、Rishabh Agarwal、Yi Su、John D Co-Reyes、Avi Singh、Kate Baumli、Shariq Iqbal、Colton Bishop、Rebecca Roelofs、Lei M Zhang、Kay McKinney、Disha Shrivastava、Cos min Paduraru、George Tucker、Doina Precup、Feryal Behbahani 和 Aleksandra Faust。2025年。通过强化学习训练语言模型进行自我纠正。载于第十三届国际学习表征会议。https://openreview.net/ forum?id=CjwERcAU7w [32] Philippe Laban、Hiroaki Hayashi、Yingbo Zhou 和 Jennifer Neville。2025年。大语言模型在多轮对话中迷失方向。arXiv:2505.06120 [cs.CL] https: // arxiv.org/abs/2505.06120[33] Jens Lehmann、Robert Isele、Max Jakob、Anja Jentzsch、Dimitris Kontokostas、Pablo Mendes、Sebastian Hellmann、Mohamed Morsey、Patrick Van Kleef、Sören Auer 和 Christian Bizer。2014年。DBpedia——从维基百科提取的大规模多语言知识库。《语义网杂志》6（2014年01月）。doi:10.3233/SW-140134[34] Quinn Leng、Jacob Portes、Sam Havens、Matei Zaharia 和 Michael Carbin。2024年。大语言模型长上下文检索增强生成的性能。载于《自适应基础模型：面向个性化与高效学习的 AI 演进》。https: // openreview.net/forum?id=Le9anH3kv1[35] Alexander Lex、Nils Gehlenborg、Hendrik Strobelt、Romain Vuillemot 和 Hanspeter Pfister。2014年。UpSet：相交集的可视化。IEEE 可视化与计算机图形学汇刊 20，12（2014），1983–1992。doi:10.1109/TVCG.2014.2346248[36] Stefano Marchesin 和 Gianmaria Silvello。2024年。高效可靠地估计知识图谱准确性。《VLDB基金会会刊》17，9（2024），2392–2404。doi:10.14778/3665844.3665865[37] Stefano Marchesin 和 Gianmaria Silvello。2025年。知识图谱准确性估计的可信区间。《美国计算机协会数据管理会刊》（SIGMOD）3，3，文章142 2025 26 10 1145 3725279 （），页码。doi: 。/ Stefano[38] Marchesin、Gianmaria Silvello 和 Omar Alonso。2024年。面向效用且标注有限的知识图谱准确性估计：以 DBpedia 为例。AAAI 人类计算与众包会议论文集 12，1（2024年10月），105–114。doi:10.1609/hcomp. v12i1.31605[39] John Xavier Morris 和 Alexander M Rush。2025年。上下文文档嵌入。载于第十三届国际学习表征会议。https://openreview.net/forum?id=Wqsk3FbD6D [40] Axel-Cyrille Ngonga Ngomo、Lorenz Bühmann、Christina Unger、Jens Lehmann 和 Daniel Gerber。2013年。抱歉，我不会说 SPARQL：将 SPARQL 查询转换为自然语言。载于第22届万维网国际会议论文集（巴西里约热内卢）（WWW '13）。美国计算机协会，美国纽约，NY，977–988。doi:10.1145/2488388.2488473[41] Natasha Noy、Yuqing Gao、Anshu Jain、Anant Narayanan、Alan Patterson 和 Jamie Taylor。2019年。产业级知识图谱：经验与挑战——五家各具特色的科技公司展示实践方法。《Queue》17，2（4月2019 48 75 10 1145 3329781 3332266 ），–。doi: 。/ 。Allard Oelen、[42] Markus Stocker 和 Sören Auer。2020年。从综述文章表格创建学术知识图谱。载于《社会巨变时期的数字图书馆》，Emi Ishita、Natalie Lee San Pang 和 Lihong Zhou（编）。施普林格国际出版公司，查姆，373–389。[43] Prakhar Ojha 和 Partha Talukdar。2017年。KGEval：自动构建知识图谱的准确性估计。载于2017年自然语言处理经验方法会议论文集，Martha Palmer、Rebecca Hwa 和 Sebastian Riedel（编）。计算语言学协会，丹麦哥本哈根，1741 1750 10 18653 1 17 1183 –。doi: 。/[44] v /D - OpenAI。2024 4。。GPT- o mini：推动经济高效的智能发展。https: // openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/。访问日期：2025-07-05。[45] Stefano Ortona、Venkata Vamsikrishna Meduri 和 Paolo Papotti。2018年。在知识库中稳健地发现正规则和负规则。载于2018 IEEE 第34届国际数据工程会议（ICDE）。1168–1179。doi:10. 1109/ICDE.2018.00108[46] Jeff Z. Pan、Simon Razniewski、Jan-Christoph Kalo、Sneha Singhania、Jiaoyan Chen、Stefan Dietze、Hajira Jabeen、Janna Omeliyanenko、Wen Zhang、Matteo Lissandrini、Russa Biswas、Gerard de Melo、Angela Bonifati、Edlira Vakaj、Mauro Dragoni 和 Damien Graux。2023。大语言模型与知识图谱：机遇与挑战。《图数据与知识汇刊》1, 1 2023 2 1 2 38 10 4230 1 1 2 ( ), ： –: 。doi:。/TGDK。。Ciyuan Peng、Feng Xia、Mehdi Naseriparsa 和 Francesco[47] Osborne。2023。知识图谱：机遇与挑战。《人工智能评论》56，11（2023），13071–13102。[48] Fabio Petroni、Tim Rocktäschel、Sebastian Riedel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu 和 Alexander Miller。2019。语言模型是知识库吗？载于《2019年自然语言处理经验方法会议暨第9届国际自然语言处理联合会议论文集》（EMNLP-IJCNLP）。2463–2473。[49] R. J. Pittman。2017。《破解对话式商务的密码》。eBay公司。https://www.ebayinc.com/stories/news/cracking-the-code-on-conversational-commerce/ 访问日期：2025-04-16。[50] J. Pujara、E. Augustine 和 L. Getoor。2017。稀疏性与噪声：知识图谱嵌入的不足之处。载于《2017年自然语言处理经验方法会议论文集》（EMNLP 2017），丹麦哥本哈根，9月*9 11, 2017* 1751 1756 10 18653 1 17 1184 -。ACL，–。[51] doi:。/v /d - Umair Qudus、Michael Röder、Muhammad Saleem 和 Axel-Cyrille Ngonga Ngomo。2025。知识图谱事实核查——综述。《美国计算机协会计算调查》（2025年7月）。doi:10.1145/3749838。已接受。[52] Qwen团队。2024。Qwen2.5：基础模型家族。https://qwenlm. github.io/blog/qwen2.5/ [53] Daniel Russo、Stefano Menini、Jacopo Staiano 和 Marco Guerini。2024。直面事实！在真实场景中评估基于RAG的事实核查流水线。CoRR abs/2412 15189 2024 10 48550 2412 15189 。（）。[54] https://doi.org/。/arXiv.。Philipp Schoenegger、Indre Tuminauskaite、Peter S. Park、Rafael Valdece Sousa Bastos 和 Philip E. Tetlock。2024。硅基群体的智慧：大语言模型集成的预测能力可媲美人类群体的准确性。《Science Advances》10，45（2024 1528 10 1126 1528 ），eadp。doi:。/sciadv.adp Farzad[55] Shami、Stefano Marchesin 和 Gianmaria Silvello。2025。使用大语言模型验证知识图谱中的事实。载于《第48届国际ACM信息检索研究与发展会议论文集》（意大利帕多瓦）（SIGIR '25）。美国计算机协会，纽约州纽约，美国，3985 3989 10 1145 3726302 3730142 –。doi:。/。Chetan Sharma 和 Ja[56] n Overgoor。。在Airbnb扩展知识获取与检索。Airbnb技术博客。2018 https://medium.com/airbnb-engineering/scaling-knowledge-access-and-retrieval-at-airbnb-665b6ba21e95 访问日期：2025-04-16。[57] Baoxu Shi 和 Tim Weninger。2016。用于知识图谱事实核查的判别式谓词路径挖掘。《基于知识的系统》104（2016），123–133。doi:10.1016/j.knosys.2016.04.015[58] Prashant Shiralkar、Alessandro Flammini、Filippo Menczer 和 Giovanni Luca Ciampaglia。2017。挖掘知识图谱中的流以支持事实核查。载于《2017年IEEE国际数据挖掘会议》（ICDM）。859–864。doi:10.1109/ICDM.2017.105[59] Kai Sun、Yifan Xu、Hanwen Zha、Yue Liu 和 Xin Luna Dong。2024。从头到尾：大语言模型（LLM）知识有多丰富？又称：LLM会取代知识图谱吗？载于《2024年北美计算语言学协会分会：人类语言技术会议论文集》（第1卷：长篇论文），Kevin Duh、Helena Gomez 和 Steven Bethard（编）。计算语言学协会，墨西哥城，墨西哥，311 325 10 18653 1 2024 18 –。doi:。/v / .[60] naacl-long。Kai Sun、Yifan Ethan Xu、Hanwen Zha、Yue Liu 和 Xin Luna Dong。2024。从头到尾：大语言模型（LLM）知识有多丰富？又称：LLM会取代知识图谱吗？载于《2024年北美计算语言学协会分会：人类语言技术会议论文集》（第1卷：长篇论文），NAACL 2024，墨西哥城，墨西哥，2024年6月16–21日，Kevin Duh、Helena Gómez-Adorno 和 Steven Bethard（编）。计算语言学协会，311–325。doi:10.18653/V1/2024.NAACL-LONG.18[61] Zafar Habeeb Syed、Michael Röder 和 Axel-Cyrille Ngonga Ngomo。2019。用于事实验证的佐证路径无监督发现。载于《语义网——ISWC 2019：第18届国际语义网会议论文集，2019年10月26–30日，新西兰奥克兰，第一部分》（新西兰奥克兰）。施普林格出版社，柏林、海德堡，630–646。doi:10.1007/978-3- 030-30793-6 36[62] Zafar Habeeb Syed、Michael Röder 和 Axel-Cyrille Ngonga Ngomo。2018。FactCheck：使用文本证据验证RDF三元组。载于《第27届美国计算机协会信息与知识管理国际会议论文集》（意大利都灵）（CIKM '18）。美国计算机协会，纽约州纽约，美国，1599 1602 10 1145 3269206 3269308 –。doi:。/。Zafar Habeeb[63] Syed、Nikit Srivastava、Michael Röder 和 Axel-Cyrille Ngonga Ngomo。2019。COPAAL——一个使用佐证路径解释事实的界面。载于《ISWC 2019卫星会议论文集（海报与演示、产业及突破性创意）》，Mari Carmen Suárez-Figueroa、Gong Cheng、Anna Lisa Gentile、Christophe Guéret、Maria Keet 和 Abraham Bernstein（编），第2456卷。施普林格国际出版公司，201–204。https: // papers.dice-research.org/2019/ISWC2019_COPAAL_Demo/public.pdf

[64] Guangya Wan、Yuqi Wu、Jie Chen 和 Sheng Li。2025。推理感知自一致性：利用推理路径实现高效的大语言模型采样。载于*2025*年美洲国家计算语言学协会分会：人类语言技术会议论文集（第*1*卷：长篇论文），Luis Chiruzzo、Alan Ritter 和 Lu Wang（编）。计算语言学协会，阿尔伯克基，新墨西哥州，3613–3635。doi: 10 18653 1 2025 184 。/ v / .naacl-long。[65] Chenguang Wang、Xiao Liu 和 Dawn Song。2020。语言模型是开放知识图谱。arXiv: 2010 11967 [ ] 2010 11967 。cs.CL https://arxiv.org/abs/。[66] Magdalena Wysocka、Oskar Wysocki、Maxime Delmas、Vincent Mutel 和 André Freitas。2024。大语言模型、科学知识与事实准确性：简化人类专家评估的框架。《生物医学信息学杂志》158（2024），104724。doi:10.1016/j.jbi.2024.104724[67] Mingfeng Xue、Dayiheng Liu、Wenqiang Lei、Xingzhang Ren、Baosong Yang、Jun Xie、Yidan Zhang、Dezhong Peng 和 Jiancheng Lv。2023。用于大语言模型高效推理的动态投票。载于计算语言学协会论文集：*EMNLP 2023*成果，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，新加坡，3085–3104。doi:10.18653/v1/2023.findings-emnlp.203[68] 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、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、Yang Fan、Yang Yao、Yichang Zhang、Yu Wan、Yunfei Chu、Yuqiong Liu、Zeyu Cui、Zhenru Zhang 和 Zhihao Fan。2024。《Qwen2 技术报告》。*arXiv* 预印本 *arXiv:2407.10671*（2024）。[69] Xiao Yang、Kai Sun、Hao Xin、Yushi Sun、Nikita Bhalla、Xiangsen Chen、Sajal Choudhary、Rongze Daniel Gui、Ziran Will Jiang、Ziyu Jiang、Lingkun Kong、Brian Moran、Jiaqi Wang、Yifan Ethan Xu、An Yan、Chenyu Yang、Eting Yuan、Hanwen Zha、Nan Tang、Lei Chen、Nicolas Scheffer、Yue Liu、Nirav Shah、Rakesh Wanga、Anuj Kumar、Wen-tau Yih 和 Xin Luna Dong。2024。《CRAG——综合 RAG 基准》。载于神经信息处理系统进展，A. Globerson、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. Tomczak 和 C. Zhang（编），第 37 卷。Curran Associates, Inc.，10470–10490。https://proceedings.neurips.cc/ paper_ files/paper/2024/file/1435d2d0fca85a84d83ddcb754f58c29-Paper- 数据集_和_基准测试_Track.pdf[70] Zhenrui Yue、Huimin Zeng、Lanyu Shang、Yifan Liu、Yang Zhang 和 Dong Wang。2024。《通过合成对比论据进行检索增强事实验证》。载于计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文），Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，10331 10343 10 18653 1 2024 556 – 。doi: 。/v / 。[71] 2024 acl-long。Danna Zheng、Mirella Lapata 和 Jeff Z. Pan。。大语言模型作为知识库有多可靠？重新思考事实性与一致性。arXiv:2407.13578 [cs.CL] https://arxiv.org/abs/2407.13578[72] Zhengzuo Zhengzuo、Zhengzuo Liu、Lanyu Li、Ling Fu、Jing Li、Tianrui Sun 和 Xiaonan Wang。2023。《面向低碳电力与能源系统的知识图谱》。doi:10.46855/energy-proceedings-10361