从全文科学文章中进行假设与统计证据提取的上下文选择

Sai Koneru
saidileep543@gmail.com 宾夕法尼亚州立大学 大学公园市，宾夕法尼亚州，美国

Jian Wu
j1wu@odu.edu 欧道明大学 诺福克，弗吉尼亚州，美国

Sarah Rajtmajer
smr48@psu.edu 宾夕法尼亚州立大学 大学公园市，宾夕法尼亚州，美国

arXiv:2603.21193v1 [cs.CL] 22 Mar 2026

摘要

从全文科学文章中提取假设及其支持性统计证据，对于综合实证研究发现至关重要，但由于文档篇幅较长，且科学论证分布在论文各个部分，因此仍然颇具挑战。本文研究一种顺序式全文提取任务：将文章摘要中的主要发现陈述与（i）论文正文中相应的假设陈述，以及（ii）支持或反驳该假设的统计证据关联起来。这种任务设定带来了颇具挑战的文档内检索问题：许多候选段落在主题上与研究发现相关，但修辞作用各不相同，从而成为检索和抽取的困难负例。我们采用先检索、后抽取的两阶段框架，对检索设计选项开展受控研究，考察上下文数量、上下文质量（标准检索增强生成、重排序，以及与重排序结合的微调检索器），并设置预言机段落，以便在四种大语言模型抽取器中区分检索失败与抽取上限。研究发现，与全文提示相比，有针对性的上下文选择始终能提升假设抽取效果；这一提升主要出现在能够优化检索质量和上下文洁净度的配置中。相比之下，统计证据抽取仍然困难得多。即使提供预言机段落，性能仍处于中等水平，这表明抽取器在处理数值与文本混合陈述方面仍存在局限，而问题并非仅仅源于检索失败。

CCS 概念

• **计算方法** → **信息抽取**.

关键词

科学信息抽取、检索增强生成、大语言模型、证据抽取、假设抽取

ACM 参考格式：

Sai Koneru、Jian Wu 和 Sarah Rajtmajer。2026年。面向全文科学文章的假设与统计证据提取的上下文选择。载于.ACM，纽约，美国，14 页。https://doi.org/10.1145/ nnnnnnn.nnnnnnn

1 引言

假设是科学方法的核心组成单位。它们表达了关于构念与变量之间关系的可检验预期 [44], ，并指导研究的设计、分析和解释。在科学文章中，假设通常以概念性方式表述，并可能在多个章节中以不同措辞重复出现 [17],；而其所获实证支持则以统计证据的形式报告，这些证据是数值与文本混合陈述，包含效应方向、系数、检验统计量、置信区间和 p值 [22]。因此，提取这些假设及用于评估它们的证据，是实现实证研究发现大规模综合的关键一步。然而，每年出版物数量的显著增长，使得大规模综合这些主张变得日益困难 [6], ，从而推动了自动化方法的发展。

此前为实现这一目标所做的工作，主要集中于主张抽取，以支持各种下游应用，包括自动化文献综述 [30, 31], 假设生成 [34], 和事实核查 [40]。尽管大语言模型（LLM）近期的进展在文本摘要和问答方面展现出潜力 [7], ，现有的大部分科学陈述提取工作主要侧重于识别摘要中的主张类句子 [1, 42]。这种聚焦于摘要的做法并未解决假设—证据抽取问题，因为假设及其支撑性统计证据往往出现在论文全文的不同部分。

全文抽取之所以具有挑战性，不仅是因为文档篇幅较长，还因为科学推理分散在各个章节中，并嵌于分层论证和多样的修辞结构之中 [24]。尽管大多数最先进的大语言模型都拥有足够大的上下文窗口，能够在单次处理中将一篇科学论文的全文作为输入 [5, 46], ，现有评估表明，端到端性能远低于人类水平 [11, 35, 48]。在全文中，相关假设或证据往往稀疏、经过改述，且被主题相关但并非目标内容的文本所包围，因此，准确抽取有赖于谨慎的上下文选择。

本研究的关注点遵循 Alipourfard等人 [2], 提出的数据结构，该结构将高层次的摘要发现与论文正文中的（i）相应核心假设以及（ii）支持该假设的统计证据相联系。具体而言，我们将主张溯源提取划分为两个连续的子任务，它们各自具有不同的特征。假设抽取是一项语义对齐任务；假设可能经过改述、以不同具体程度表述，并分布在多个章节中。相比之下，证据抽取侧重于识别包含稠密统计信息（例如数值、检验结果）的文本片段，这些信息通常出现在结果部分（图 1）。

我们提出一种一致的两阶段检索与抽取流水线（图 2）。在阶段1，我们将摘要中预先抽取的发现陈述作为查询，检索候选段落，并由大语言模型从中抽取一个与该发现陈述对应的核心假设或其释义变体。在阶段2，我们将摘要发现

Conference'17, 美国华盛顿特区 2026年。AC
M ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

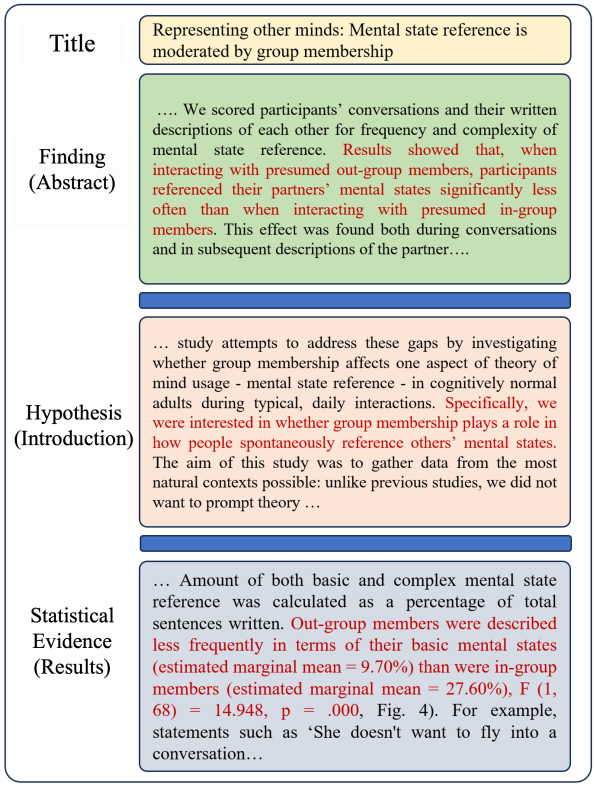


图 1：一个论断追踪链示例，贯穿论文的摘要、引言和结果部分。我们的两阶段流水线利用摘要级发现来抽取假设陈述（本例中来自引言），随后定位统计证据（结果部分）。

与抽取出的假设拼接，构造复合查询，以检索用于统计证据提取的候选段落。我们将这一两阶段流水线的每个阶段都置于检索增强生成（RAG）范式中，并将其与全文提示基线方法进行比较，以分离上下文选择的影响。

我们通过在这一范式下进行广泛的受控消融实验，得出了关键结果，实验考察了：(i)检索数量，检索排名前-k个段落（$k \in \{5, 10, 20\}$）；(ii) 检索质量，比较标准稠密检索、重排序检索，以及与同一重排序器配对的微调检索器；(iii)抽取器局限，通过特定阶段的预言机上下文提供阶段1中的标准假设段落和阶段2中的标准证据段落，从而界定性能上限，并区分检索失败与抽取失败。我们发现：

• **任务相关瓶颈。** 假设抽取受到检索质量和上下文洁净度的显著限制，而即使在预言机设置下，证据抽取的准确度仍只是中等水平（不同模型的预言机证据 F1介于0.47至0.55之间）。

• **在假设抽取方面，基于上下文的检索优于全文提示。** 将检索到的上下文从 k=5 → 10 →增加到20，通常有利于假设抽取，或至少不会产生负面影响，而全文提示的表现往往不如基于上下文的检索。

• **检索改进带来性能提升，但仍有很大的提升空间。** 重排序和微调检索提升了两个阶段的性能，但即使提供金标准段落，证据抽取的效果仍远未达到预言机水平。这表明，仅改进检索器还不够，抽取器本身也是一个重要瓶颈。

总之，这些实验在文档内场景中对假设和证据抽取的上下文选择进行了受控比较，其中包括预言机上下文，以区分段落选择效应与抽取器的局限。[1]

2 相关工作

2.1 从科学文本中抽取陈述

以往从学术文本中抽取科学陈述（如论断和假设）的研究，通常将该问题表述为句子／片段检测或序列标注。早期研究侧重于从科学摘要中识别论断句，利用话语结构和迁移学习，将类似论断的句子与周围上下文区分开来 [1, 19]。近期研究则通过监督式对比学习等目标函数，增强论断句分类器，使其能够将论断句与主题相似的非论断句区分开来 [42]。另一些相关研究则超越句子级检测，转而对科学陈述进行结构化表示，抽取细粒度的论断图，以编码实体、关系和修饰语，为后续的科学知识图谱构建提供支持 [28]。

修辞分区和句子序列分类提供了区分陈述类型的机制（例如背景、目标、方法、结果、结论），有助于将假设/主张类内容与其他话语角色区分开来 [12, 37]。聚焦引文的资源进一步捕捉科学论断如何被置于特定语境中并获得支持，例如通过引文意图分类以及相关的引文与话语建模 [9]。

该领域文献反复强调，从摘要扩展到全文论文会同时带来规模和歧义方面的问题，因为目标内容可能较为稀疏、以释义形式分布于不同章节之间，并与修辞上相邻但并非目标的陈述交织在一起 [43]。文档级科学信息抽取基准测试凸显了这样一种难题：当证据跨越多个句子，或必须在整篇文档范围内进行汇总时，如何抽取结构化信息 [18, 27]。由于科学主张和假设嵌入结构化话语之中，信息抽取通常可从话语线索和论证结构中获益 [14, 23, 24]。最后，既有研究还强调了定义歧义和标注挑战，而这些问题会因任务和研究群体的不同而有所变化 [39, 44]。

我们的工作超越了句子层面的主张和假设识别，进一步从全文论文中抽取研究发现、假设与支持性证据之间的结构化链接。我们通过系统地改变上下文的构建方式来处理全文歧义，并将段落选择错误与抽取器局限区分开来。

[1]代码、提示词和实验配置可在 https://github.com/SaiDileepKoneru/ScientificClaimTraceExtraction.git 获取

2.2 用于主张验证的证据检索

另一类相关研究关注科学主张验证：系统需要检索相关文档，并识别支持真实性判定的证据句。通用领域事实验证和理由评估的研究采用先检索后推理范式，并明确进行证据选择 [13, 38]。在科学领域，这包括检索包含证据的摘要，并选择支持或反驳某项主张的理由句 [40]。后续研究通过将任务分解为摘要检索、句子选择和标签预测 [32], ，并纳入全文上下文和弱监督，以提升鲁棒性和领域迁移能力，从而强调了检索和上下文范围的重要性 [41]。我们采用类似的先检索后提取范式，但将其调整为抽取结构化的主张轨迹，而非预测真实性标签。此外，验证流程通常会跨多个文档检索证据，而我们的任务则聚焦于在单篇文档中抽取与假设相关的统计证据，以重构论文的实证论证。

2.3 检索配置的影响

近期针对科学长上下文理解的基准测试表明，仅向模型提供长输入，并不能保证模型在各项任务中实现可靠的科学理解与推理 [11]。这促使人们谨慎配置检索，因为无关但语义相关的段落可能会积极误导下游大语言模型的行为，并降低检索增强设置下的正确性 [3, 10, 47]。对照研究表明，大语言模型易受干扰。即使提示中包含有用信息，加入无关上下文也可能使回答产生偏差；当无关上下文与目标高度相似时，这种影响最为显著 [33, 45], ——这种情况在文档内检索中十分常见。除了相关性，所提供上下文的长度和位置也会影响性能；先前的长上下文评测记录了中间信息遗失现象，即模型对埋在上下文中间的相关信息利用不足 [26]。此外，研究表明，增加检索到的上下文数量会带来递减收益：性能起初有所提升，但输入较长时则可能受到损害 [20]。检索召回率与精确率之间的相互作用也很复杂；后续的过滤或选择步骤可能只有在初始检索达到较高召回率时才有帮助 [25]。

这些研究发现共同表明，两阶段检索设计值得采用：在第一阶段检索时优先保证高召回率，然后通过对检索到的段落进行重排序并将其纳入候选集，以提高精确率 [8, 16]。我们采用了这一方法。首先检索高召回率候选集，然后使用重排序来提高精确率，从而在存在高度相似的文档内干扰项时，能够对上下文数量和质量进行受控比较。

3 任务描述

我们的目标是从给定科学论文的全文中提取一条结构化的实证论证脉络。我们的一个关键假设是，论断本质上是实证性的，因此我们只关注提取支持既定假设的统计证据，不考虑其他形式的学术推理，例如理论论证。

形式上，设一篇科学文章 D可建模为由段落组成的序列（p_1、 p_2、……、 p_n），其中每个段落 p_i都是由句子构成的序列 p_i =（$s_{i,1}$、 $s_{i,2}$、……、s_{i,m_i}）。在预处理中，我们仅将文章正文表示为纯文本段落，排除表格、图（及图注）和方程等非正文内容，因此每个 p_i 都对应于正文文本中的一个连续文本块。设 F_A 表示来自 D摘要的一个摘要发现句子。我们的目标是从 D正文中提取两个文本片段：

- 核心假设是一种与摘要发现相关的解释性或假设层面的陈述。H 它可以是一个句子，也可以跨越多个连续句子。 H
- 统计证据 E（H）是包含统计结果（例如 p 值、置信区间、回归系数）并用于支持假设的文本片段 H。

这项任务面临若干独特挑战，使其有别于标准的信息抽取。值得注意的是，摘要中报告的研究发现与统计证据之间的联系通常是间接的。统计证据支持某项假设，而该假设又支持这一研究发现。

4 方法

4.1 抽取流程

我们将任务定义为一个依次进行的两步流程：首先抽取假设，然后抽取支持该假设的统计证据（见图 2）。

阶段 1：假设抽取。 我们使用摘要中的研究发现作为查询，以识别全文中最可能包含核心假设的段落。我们首先将文章正文转换为段落，然后使用第 4.2 节中的一种上下文配置构建特定于阶段的上下文集 C_1 。具体而言，基于检索的配置首先检索一个大小为 M 的候选池（例如，标准检索增强生成中为 M=k，或使用重排序器时为 M=30 ），并可选择使用大语言模型重排序器（RankGPT [36]）选出最终的前-k 上下文集。对于重排序后的配置，我们使用 M=30并保留 k=5 个段落，与表 1 中固定的上下文大小一致。全文配置跳过检索，并将 C_1 设为整篇文章正文（不包括摘要和参考文献）。给定由此得到的上下文 C_1 和研究发现 F_A，大语言模型抽取器会生成假设 $\hat{}H$。

阶段2：证据抽取。 我们将摘要发现与抽取出的假设$\hat{}$ q_2 = [F_A;H]拼接，形成复合查询，并使用相同的检索和重排序流程（或根据配置使用全文/Oracle 上下文）构建证据阶段上下文集 C_2 。随后，向抽取器提供 F_A、$\hat{}H$和 C_2 ，以抽取统计证据$\hat{}E$。所有抽取步骤中，我们使用直接根据人工标注指南改编的提示词 [2]。完整的提示模板（假设抽取、证据抽取和评估器解析）以及实验配置文件均可在随附的代码库中找到。[1]

4.2 上下文配置

我们比较以下上下文构建配置，每种配置均在顺序两阶段流水线中实例化：

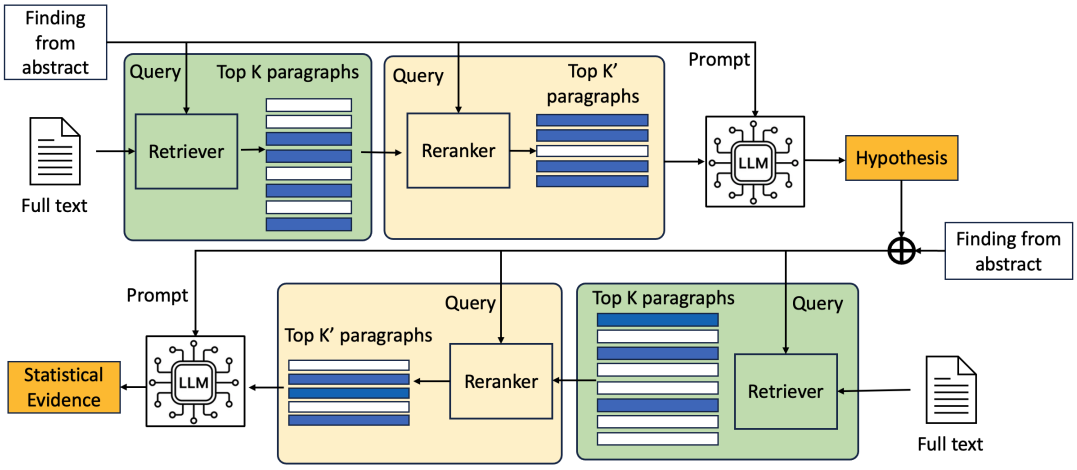


图 2：顺序检索与抽取流程示意图。阶段1（上）利用摘要发现抽取假设，随后将其作为复合查询的一部分，用于阶段2（下）抽取统计证据。

零样本（全文，两阶段）：此基线方法采用相同的两阶段抽取流程，但以全文文章正文替代检索到的上下文。阶段1中，我们提供全文（不包括参考文献）和 F_A，并提示大语言模型抽取 $\hat{}H$。阶段2中，我们再次提供全文，以及 F_A 和 H，并提示大语言模型抽取 $\hat{}E$。这样，在保持顺序式表述不变的同时，便可单独考察上下文选择的效果。

思维链（全文，两阶段）：此基线方法与全文两阶段基线方法完全相同，但在每个阶段中使用思维链提示 [21]（即模型在输出阶段1中的$\hat{}H$和阶段2中的$\hat{}E$之前，会先生成中间推理）。此方法用于测试，仅改进提示能否弥补缺少上下文选择的问题。

标准检索增强生成：在每个阶段中，我们使用 SciBERT 检索器从文章正文中选择排名前k的段落，并将其传递给大语言模型抽取器。我们评估 $k \in \{5、10、20\}$。

标准检索增强生成 +重排序器：在每个阶段中，我们使用 SciBERT 检索 M=30 个段落作为候选池，然后应用 RankGPT 重新排序并保留排名前k的段落。我们使用 k=5。

微调检索器 + 重排序器：这一流程由任务感知段落检索器驱动，我们对其进行了微调，使其能够区分段落的修辞功能。该检索器基于双编码器架构构建，使用 SciBERT [4] 作为共享双编码器，为查询和每个候选段落生成稠密表示。为了更好地区分两个段落的修辞功能，我们将 SciBERT 输出的嵌入传入两个独立的线性投影头。我们将假设检索（阶段1）和证据检索（阶段2）视为两个相关任务，并通过共享的 SciBERT 编码器上的独立投影头分别实现。模型使用多任务三元组损失进行微调，在缩小与正例段落之间距离的同时，增大与负例段落之间的距离。

为优化多任务三元组损失，我们在假设和证据任务中采用了一致的困难负样本识别策略。对于每个训练样例，将真实段落用作正样本点。为了识别困难负样本点，我们在文档的其余部分中搜索与正段落语义最相似的段落，并以它们的 SciBERT 嵌入的余弦相似度作为衡量标准（附录图7展示了一个选出的困难负样本示例）。为确保我们没有选取假设的未标注释义，我们过滤掉了其中任意组成句子与人工标注假设的相似度超过 0.89 的候选项。该阈值与假设评估中使用的校准语义匹配标准一致（第 5.1 节）。

由此产生了具有挑战性的训练样例，模型必须区分与查询相关、修辞作用不同但高度相似的段落。两个任务采用完全相同的流程，唯一的区别在于锚定查询：假设任务使用摘要发现，而证据任务使用由摘要中的发现与真实假设拼接而成的复合查询。在推理阶段，我们使用微调后的检索器，并采用与检索增强生成 + 重排序器基线方法相同的重排序设置。我们检索 30 个候选项，使用 RankGPT 进行重排序，并将排名前 5 的段落提供给抽取器。

Oracle 设置：为确定段落选择完美时的性能上限，我们向抽取器提供金标准段落。在阶段1，我们将 C_1 设为包含人工标注的假设的段落。在阶段2，我们提供真实假设（以避免错误传播），并将 C_2 设为包含人工标注的统计证据的段落。这样便能将抽取器局限与检索失败区分开来。

5 实验

我们在 Alipour-fard 等人介绍的数据集上进行实验，该数据集包含来自社会与行为科学（SBS）领域的 2,257 篇全文科学文章。 [2], 每篇文档都包含一条由领域专家标注的论断轨迹，其中包括摘要中的高层次概念性发现、正文中的解释性或假设层面陈述（假设），以及相关的统计结果（证据） [15]。预处理后，每篇文档平均包含 18,268 个词元和 367 个句子；预处理包括移除摘要和参考文献部分。

5.1 评估指标

我们使用针对各类输出特性设计的指标评估假设与证据提取。全文均报告基于提取单元（假设对应句子；证据对应结构化组件）计算的微平均精确率/召回率/F1，并对评估集中所有匹配单元进行汇总。

假设抽取（语义匹配）。论文中经常会以不同措辞转述假设，因此精确字符串重叠的判定标准过于严格。为此，我们采用经过校准的语义匹配标准来评估假设抽取。首先，我们从留出划分中抽取模型提取的假设样本，构建带标签的校准集，并使用四级评分标准，将每条抽取结果与真实假设进行对照评分：完全匹配（3）、大体相似（2）、部分相似（1）和不同（0）。完整的标注指南和示例可在随附的代码库中找到。[1] 随后，我们将评分2–3视为正确，0–1视为不正确，从而将这些标签二值化。

我们使用 Gemini 嵌入对预测假设文本和标准假设文本进行嵌入，采用针对语义文本相似度（例如推荐和重复项检测应用场景）优化的 SEMANTIC_SIMILARITY 任务类型，并计算余弦相似度。我们通过在带标签的校准集上进行网格搜索来选择相似度阈值 τ，搜索范围为 $\tau \in [0.1, 1.0]$，步长为0.01，并选取使 F1分数 最大的值（附录图5）。我们将得到的阈值 $\tau = 0.89$用作所有假设评估中的语义匹配截断值。在评估时，我们将预测的假设片段拆分为句子；如果某个预测句子与标准假设（或者当标准片段包含多个句子时，与其中任一标准句子）的余弦相似度不低于 τ，则将其计为匹配。随后，根据匹配与未匹配的预测句子和标准句子计算精确率、召回率和 F1分数。该流程既能奖励语义等价的转述，也会惩罚提取片段中包含的额外非假设内容。附录图6展示了一个具有代表性的抽取失败案例。

证据抽取（组件级评分）。统计证据通常是数值文本混合内容，因此评估时需要同时衡量语义对齐和字段级正确性（变量、检验、系数、区间）。精确字符串重叠对改写十分敏感，而嵌入相似度可能无法反映数值保真度。因此，我们采用组件级评分流程，与近期科学基准测试中的结构化评估范式一致 [29]。

组件模式。我们人工检查了 50 个随机抽取的金标准证据片段，以确定正文统计报告中表达最一致的字段，并定义一组固定的五种组件类型，与评估器提示相对应：(i) 变量（关键预测变量/结果变量及构念）；(ii) 效应关系（例如方向性和显著性陈述）；(iii) 置信区间；(iv) 统计检验/模型族；以及 (v) 样本量/自由度（如有说明）。

评估器辅助的组件解析。 给定一个预测证据片段及其对应的标准片段，我们使用固定的评估器大语言模型（GPT-4）和固定的组件专用提示，将两段文本解析为每种组件类型对应的结构化 JSON 字段。[1] 对于集合型组件（变量、关系、置信区间），评估器会返回标准陈述中的项目数（$|A|$）、提取的陈述中的项目数（$|B|$）以及模糊匹配交集计数（$|A \cap B|$）。对于二元组件（检验/模型类别；样本量/自由度），评估器会返回匹配指示值，以及从每段文本中提取的字符串。对于二元组件，我们将两者均缺失视为不适用：当评估器为标准字段和预测字段均返回空字符串时（例如，标准_检验 = 预测_检验 = ""），我们会将该组件实例排除在评分之外，而不将其计为错误。评估器仅用于解析；评分则根据返回的字段以确定性方式计算。

汇总精确率/召回率/F1。解析后，每种组件类型会生成一组抽取项（例如变量列表、置信区间列表）或单个属性（例如检验/模型类别、样本量/自由度）。我们通过统计这些组件项上的匹配数来计算精确率/召回率/F1。每个正确匹配的组件项计为一个真正例；系统生成但不受标准陈述支持的每个组件项计为一个假正例；未能恢复的每个金标准组件项计为一个假负例。对于集合型组件，这相当于统计解析后的金标准项列表与预测项列表之间的重叠（对变量和置信区间采用模糊匹配）。对于二元组件（检验/模型类别；样本量/自由度），若解析后的属性一致则计为匹配，不一致则计为错误。如果金标准证据和预测证据中均未出现某个组件，则不将其纳入评分。对于集合型组件，缺失对应空集（$|A| = |B| = 0$）。对于二元组件（检验/模型类别和样本量/自由度），缺失对应两个抽取字符串均为空（例如，金标准_检验 = 预测_检验 =""），此时不纳入该实例的评分。否则，将二元组件计为一个项：若双方均有内容且匹配 = 1，则计为一个真正例；若金标准有内容而预测为空，则计为一个假负例；若金标准为空而预测有内容，则计为一个假正例；若双方均有内容但匹配 = 0，则计为一个假正例和一个假负例。

这种基于组件的评估会在抽取结果正确还原了部分字段（例如变量和方向），但遗漏或错误表述了其他字段（例如置信区间）时给予部分得分。由于评分器依赖基于大语言模型的解析器来抽取组件，因此评估结果取决于评估器模型和提示词；因此，我们将其视为统计证据抽取在字段级（结构化）保真度方面的自动化代理指标。图3给出了一个具体示例，展示了解析出的组件以及它们如何计入部分得分。

6 结果

表 1报告了四种抽取器在所有上下文配置下的两阶段流水线的精确率/召回率/F1。我们分别评估阶段1和阶段2。除非另有说明，我们总结的是各抽取器的平均趋势。

我们围绕三个问题组织结果：在标准 RAG 中，检索到的**上下文数量**如何影响性能（$k \in \{5, 10, 20\}$）；在上下文大小固定时，检索到的**上下文质量** 如何影响性能（标准 RAG k=5 → RAG+重排序器 k=5 →微调检索器+重排序器 k=5）；以及**预言机上下文**揭示了哪些关于检索与抽取上限的信息。其他诊断分析，

包括逐篇论文转变分析，以及基于真实假设的证据抽取，详见附录（图 8，表 2）。所有配置都采用相同的两阶段表述；它们的区别仅在于如何构建 C_1 和 C_2（全文、top-k 检索、先检索后重排序至固定的 k=5，或预言机金标准段落）。

定向上下文选择能稳定提升假设抽取效果。对抽取器取平均后，全文提示的表现不及基于检索的上下文（平均 F1分数为 0.39，而标准检索增强生成在 k=5/10/20下分别为 0.43/0.47/0.49）；最佳非预言机配置（微调检索器+重排序器，k=5）的平均假设F1分数达到 0.50（比全文提示高约 +0.11）。这些提升与全文论文中的信号稀释效应相符：许多段落虽然与摘要发现主题相关，但在修辞上并非目标内容。

证据抽取对上下文呈现出不同的敏感性。平均而言，使用较小上下文（k=5）的标准检索增强生成表现不及全文提示（平均证据 F1为 0.20，而全文提示为 0.24），但增加k可提升性能（k=10时平均证据 F1为 0.23，k=20时为 0.25）。在固定规模下，改善检索质量带来的提升更为明显（RAG + k=5 + k=5 1重排序器 F1为 0.29 0.26，微调检索器重排序器 F）。预言机上下文显示出中等程度的性能上限（不同模型的预言机证据 F1为 0.47 0.55）。这表明，证据抽取不仅受检索限制，也受抽取器性能限制。

示例：证据抽取的组件级比较

DOI. 10.1016/j.jesp.2017.11.013

金标准证据（标注）。与内群成员相比，外群成员较少被描述为具有基本心理状态（估计边际均值 = 9.70%），而内群成员的估计边际均值为 = 27.60%，$F(1,68) = 14.948$, $p = .000$。

抽取出的证据。撰写有关假定外群成员内容的参与者，较少提及其复杂心理状态（估计边际均值 0 5 .%）；相比之下，撰写有关假定内群成员内容的参与者，估计边际均值为 = 3.6%；$F(1,68) = 8.457$, $p = .005$。

解析后的组件重叠（评估器解析；评分具有确定性）。

组件	重叠/匹配	备注
变量（集合）	$\|A \cap B\| = 2$（$\|A\| = 3$的 $\|B\| = 7$）	部分重叠
关系（集合）	$\|A \cap B\| = 2$（$\|A\| = 2$的 $\|B\| = 5$）	额外预测的关系
置信区间（集合）	NA（两者均不存在）	未评分
检验/模型族（二元）	1	F检验族
样本量/自由度（二元）	1	$df = 68$

图3：证据提取组件级评估示意图。该指标会给予部分得分：当提取的文本片段捕捉到了关系，但对应的是报告中数值细节不同的另一种效应时，即如此。

6.1 检索上下文数量（k）的影响

我们首先在标准检索增强生成设置下改变检索上下文数量：使用 SciBERT 检索排名前-k 个段落（$k \in \{5, 10, 20\}$），同时固定抽取器和提示。对于假设抽取，性能通常会随着 k 增加而提升，且某些抽取器的性能会趋于饱和（表 1）；这与假设经过改写并分布在多个位置的情况相符。对于证据抽取，增加 k 可提升各抽取器的性能（表 1），不过与假设抽取相比，性能增益较小，也更依赖模型。重要的是，全文提示并不等同于检索大量k。对于假设抽取，尽管全文输入具有最大召回率，其表现仍不及检索到的上下文数量最多的情况，这表明主题相似但修辞目的并非目标的段落会稀释信号。对于证据抽取，全文输入对某些抽取器仍具有竞争力，说明更广泛的周边上下文有时能帮助模型定位并解读统计报告。

6.2 固定 k=5时检索质量的影响

接下来，我们通过固定最终上下文大小（k=5）来隔离检索质量的影响，并比较三种检索方案：标准检索增强生成、检索增强生成+重排序器（使用 SciBERT 检索 30 条结果，然后用 RankGPT 选出前 5 条），以及微调检索器+重排序器（用微调检索器替换 SciB- ERT，同时保持相同的重排序流程）。在不同模型中，提升检索质量都能为两个子任务带来稳定的收益（表 1），但其影响有所不同。

对于假设抽取，重排序和微调的主要作用，是从大量主题相似的文档内候选项中选出修辞上恰当的假设陈述：平均 F1分数从 0.43（标准检索增强生成，k=5）提升至 0.47（检索增强生成+重排序器）以及 0.50（微调检索器+重排序器）。对于证据提取，检索质量的提升也有帮助，但收益与在固定上下文大小下保留更多与证据相关的段落相一致。值得注意的是，微调检索器也可能带来权衡，因为它会改变在固定上下文大小下保留哪些段落；我们在附录（图 8）中分析了这些逐篇论文的保留与丢失模式。

6.3 预言机上下文决定抽取上限

最后，我们评估一种预言机上下文设置：向抽取器提供包含人工标注假设或证据的金标准段落。这样可以排除检索失败，并在段落选择完美的情况下，为抽取性能提供上限。

对于假设抽取，预言机上下文显著提升了所有抽取器的性能（平均假设F1分数为 0.73），与表现最佳的非预言机配置（微调后的检索器+重排序器，平均 F1分数为 0.50）相比，仍存在较大差距。这表明假设抽取受到上下文选择的强烈制约，但即使提供正确段落，抽取性能仍有很大的提升空间。

对于证据抽取，预言机上下文揭示了另一处瓶颈：即便段落选择完美，性能仍然一般（各模型的预言机证据 F1 为 0。47–0。55；表 1）。这一上限可能反映了多种因素：抽取器对数值与文本混合陈述的处理性能，以及部分统计细节可能报告在正文之外（例如表格或图中），而这些内容在预处理时会被移除。因此产生的预言机差距（例如，微调检索器+重排序器的平均值为 0。29，而预言机为 0。52）表明，证据错误不仅由检索造成，也源于抽取器对结构化统计内容的保真度不足（例如，未能保留变量、方向性和数值字段）。

诊断分析。 为帮助理解表 1 中的汇总分数，我们在附录A中提供了详细的诊断分析。首先，我们分析相关句子比例（RSP），以量化假设相关信息的密度。我们发现，在不同配置下，信号密度越高，F1分数也越高（附录图4）。其次，对于以真实假设为条件的证据抽取，我们报告了逐篇论文的检索结果细分：检索变更带来的改进主要集中在新近检索到标准证据段落的情形；而当重排序或检索器变更未能保留该段落时，则会出现性能回退（附录图8、表 2）。

7 讨论

我们任务的一个显著特征是，检索在单篇论文内部进行。因此，候选段落通常在主题上与摘要发现相关，但在修辞功能上存在很大差异（例如背景框定、理论动机、假设关系、结果摘要和启示）。在这种情况下，重排序和微调检索器带来的改进，与检索变更影响固定大小的最终上下文中出现哪些段落这一点相一致。虽然我们不对模型行为提出机制性主张，但汇总模式表明，检索改进转化为假设抽取增益的可靠性高于其转化为证据抽取增益的可靠性。

为补充汇总指标，我们报告了两项描述性诊断（附录）。首先，我们分析 相关句子比例（RSP），将其作为假设上下文信号密度的代理指标：对于给定的假设阶段上下文，RSP 是其中语义相似度超过标准假设的句子所占的比例，所使用的阈值与假设评估中采用的同一校准阈值一致。在不同配置下，更高的 RSP 始终与更高的假设提取F1相关联（附录图4）。其次，对于以真实假设为条件的证据抽取，我们进行逐篇论文转变分析，根据检索变更对标准证据段落的处理情况，将论文分为 Gains、Keeps、Loses 或 Misses 四类，并报告相应的逐篇论文平均 ΔF1（附录图 8、表 2）。这些分析描述了检索变更如何改变检索到的上下文，以及增益集中出现在哪些位置（例如新检索到标准段落的情况），但不将性能变化归因于特定的模型内部机制。

预言机上下文进一步区分了检索失败与抽取局限。对于假设而言，最佳非预言机设置与预言机性能之间的剩余差距表明，即使提供正确的段落，性能仍有较大的提升空间。对于证据而言，各模型的预言机性能仍处于中等水平，这与证据抽取需要忠实处理数值与文本混合陈述相一致。在实践中，改进这一阶段可能需要更严格地强制结构的方法（例如模式约束输出），或结合检索与确定性解析的混合方法，以处理系数、置信区间和 p值等量值。

上述诊断和分析表明，构建主张轨迹抽取系统有两项最佳实践。首先，对于假设抽取，提升检索质量并检索范围适度更大的上下文，很可能显著提高性能。相反，对于证据抽取，检索质量和上下文大小固然重要，但如果不相应提升结构化统计内容的抽取器保真度，收益仍然有限。总之，这些研究发现强调，全文抽取流水线应将检索与抽取视为相互关联的组成部分：检索是影响假设抽取的主导因素，而证据抽取还取决于能否准确保留并对齐统计字段。

8 结论

我们研究了一种顺序式全文主张轨迹抽取问题：将摘要级发现关联到（i）论文正文中相应的假设陈述，以及（ii）支持该假设的统计证据。这一表述构成了具有挑战性的文档内检索设置：候选段落通常在主题上与该发现相关，但在修辞作用上有所不同，并且错误可能在各阶段之间传播。

我们采用一致的两阶段检索与抽取流水线，开展了一项受控研究，考察检索配置如何影响各个阶段。研究中，我们改变了检索上下文数量（top-$k \in \{5, 10, 20\}$）、检索质量（标准检索增强生成、重排序，以及与重排序配合的微调检索器），并采用预言机段落设置，以区分检索失败与抽取上限。在四种大语言模型抽取器中，假设抽取稳定地受益于有针对性的上下文选择：基于检索的上下文优于全文提示，而提升检索质量也能带来稳定的改进。检索变更对证据抽取的提升则较为有限；即使采用预言机上下文，准确率仍仅属中等，这表明对于结构化统计内容，仅靠段落选择并不能消除证据抽取错误。

总的来说，这些结果凸显了顺序式全文抽取中存在任务依赖型瓶颈：在文档内场景中，假设抽取会受到上下文选择的显著影响；而对于数值与文本混合陈述，即使能够获得正确的段落，证据抽取的效果仍取决于抽取保真度。未来的研究可以将分析范围扩展到社会与行为科学（SBS）论文之外，并开发能够更好保留统计证据中变量绑定关系和数值字段的抽取方法。

局限性与伦理考量

我们的实验仅限于从社会与行为科学学科的实证研究中提取的论断追踪链。不同学科的写作惯例各不相同，因此观察到的权衡可能无法直接推广到其他领域。我们的数据集中每篇论文仅包含一条标注的主张轨迹，并不全面，因为论文通常包含多个假设及其支持性结果。

我们的预处理仅使用纯正文段落，并排除表格、图（及图注）和方程；因此，证据抽取结果仅反映对正文报告的统计证据的性能。因此，Oracle 证据 F1 应被理解为正文内完美段落选择条件下的上限，而不是对恢复可能在正文文本之外报告的统计细节的上限。此外，该任务是顺序进行的，因此错误可能会从阶段1（假设抽取）传播到阶段2（证据抽取）；我们通过在给定

LLM	方法	假设			证据		
		P	R	F1	P	R	F1
Llama 3.2	零样本（全文）	0.34	0.40	0.36	0.18	0.20	0.19
	思维链（全文）	0.21	0.34	0.26	0.19	0.22	0.20
	检索增强生成（k=5）	0.45	0.36	0.40	0.17	0.22	0.19
	检索增强生成（k=10）	0.50	0.40	0.44	0.18	0.22	0.20
	检索增强生成（k=20）	0.50	0.41	0.45	0.20	0.21	0.21
	检索增强生成 + 重排序器（k=5）	0.45	0.41	0.43	0.19	0.24	0.21
	微调后的 SciBERT + 重排序器 (k=5)	**0.51**	**0.43**	**0.47**	**0.25**	**0.28**	**0.26**
	预言机上下文	0.67	0.55	0.61	0.43	0.52	0.47
GPT-OSS	零样本（全文）	0.48	0.40	0.44	0.23	0.27	0.24
	思维链（全文）	0.41	0.50	0.45	0.25	0.28	0.26
	检索增强生成 (k=5)	0.52	0.43	0.47	0.15	0.20	0.17
	检索增强生成 (k=10)	0.53	0.46	0.49	0.19	0.25	0.21
	检索增强生成 (k=20)	0.56	0.50	0.53	0.22	0.31	0.25
	检索增强生成 + 重排序器 (k=5)	0.54	0.50	0.52	0.24	0.28	0.25
	微调后的 SciBERT + 重排序器 (k=5)	**0.58**	**0.51**	**0.54**	**0.25**	**0.31**	**0.27**
	预言机上下文	0.91	0.73	0.81	0.50	0.57	0.53
GPT-4o-mini	零样本（全文）	0.33	0.40	0.36	0.17	0.22	0.18
	思维链（全文）	0.43	0.42	0.42	0.21	0.26	0.23
	检索增强生成（k=5）	0.50	0.46	0.48	0.23	0.29	0.26
	检索增强生成（k=10）	0.57	0.47	0.52	0.26	0.33	0.28
	检索增强生成（k=20）	0.58	0.47	0.52	0.28	0.35	0.29
	检索增强生成 + 重排序器（k=5）	0.56	0.46	0.50	0.30	0.39	0.33
	微调后的 SciBERT + 重排序器 (k=5)	**0.59**	**0.49**	**0.54**	**0.33**	**0.41**	**0.35**
	预言机上下文	0.88	0.70	0.78	0.49	0.63	0.55
gemini-2.5-flash	零样本（全文）	0.35	0.47	0.40	0.31	0.48	0.35
	思维链（全文）	0.35	0.48	0.40	**0.32**	**0.49**	**0.36**
	检索增强生成 (k=5)	0.40	0.37	0.38	0.15	0.20	0.16
	检索增强生成 (k=10)	0.42	0.41	0.42	0.19	0.25	0.21
	检索增强生成 (k=20)	0.43	0.45	0.44	0.22	0.30	0.25
	检索增强生成 + 重排序器 (k=5)	0.44	0.43	0.44	0.24	0.33	0.27
	微调后的 SciBERT + 重排序器 (k=5)	**0.46**	**0.47**	**0.46**	0.27	0.39	0.30
	预言机上下文	0.74	0.69	0.71	0.51	0.59	0.53

表 1：从科学文章全文中顺序两阶段抽取假设及其对应证据的结果。我们比较了四种大语言模型的假设和统计证据抽取性能（精确率、召回率、F1分数），并对比全文基线方法与采用不同上下文数量（k）和检索质量的检索增强策略。预言机设置表示在段落选择完美的情况下所能达到的性能上限。每种模型和子任务中，非预言机结果里 F1分数最高者以粗体标示；预言机结果则作为性能上限报告。

真实假设的条件下评估证据抽取，部分隔离了证据阶段（附录表2）。最后，评估依赖于自动化代理指标：假设评分使用经过校准的嵌入相似度阈值，证据评分使用基于大语言模型的组件解析器和确定性评分。我们将这些指标视为语义匹配（假设）和字段级性能（证据）的可扩展近似。

本研究不涉及人类受试者或个人数据。主要风险在于实际操作层面：全文科学文章可能受到版权和许可限制，因此我们仅发布代码、提示词和配置，而不重新分发全文。此外，假设与证据提取的结果看似合理却可能不正确——尤其是数值文本证据——因此应将输出视为辅助信息，并在高风险应用中进行审查。最后，我们的

证据评分依赖基于大语言模型的解析器，我们的系统依赖预训练模型；这两者都可能引入系统性错误或偏见，也可能偏向主流写作风格。

参考文献

[1] Titipat Achakulvisut、Chandra Bhagavatula、Daniel Acuna 和 Konrad Kording。2019。使用深度话语模型和迁移学习从生物医学出版物中抽取主张。*arXiv* 预印本 *arXiv:1907.00962*（2019）。[2] Nazanin Alipourfard、Beatrix Arendt、Daniel M Benjamin、Noam Benkler、Michael Bishop、Mark Burstein、Martin Bush、James Caverlee、Yiling Chen、Chae Clark 等。2021。系统化开放研究与证据中的信心评估（SCORE）。*SocArXiv*（2021）。[3] Chen Amiraz、Florin Cuconasu、Simone Filice 和 Zohar Karnin。2025。干扰效应：理解检索增强生成中的无关段落。载于第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编）。计算语言学协会，奥地利维也纳，18228–18258。doi:10.18653/v1/2025.acl-long.892

[4] Iz Beltagy、Kyle Lo 和 Arman Cohan。2019。SciBERT：面向科学文本的预训练语言模型。载于*2019*年自然语言处理经验方法会议与第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，Kentaro Inui、Jing Jiang、Vincent Ng 和 Xiaojun Wan（编）。计算语言学协会，中国香港，
3615 3620 10 18653 1 19 1371 –。doi: 。/v/D - Iz Beltagy、[5] Matthew E Peters 和 Arman Cohan。2020。Longformer：长文档转换器。*arXiv* 预印本 *arXiv:2004.05150*（2020）。[6] Lutz Bornmann、Robin Haunschild 和 Rüdiger Mutz。2021。现代科学的增长率：一种潜在分段增长曲线方法，用于对既有文献数据库和新建文献数据库的出版物数量进行建模。人文与社会科学传播 8, 1（2021），1–15。[7] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是少样本学习者。神经信息处理系统进展33（2020），1877–1901。[8] Haotian Chen、Qingqing Long、Meng Xiao、Xiao Luo、Wei Ju、Chengrui Wang、Xuezhi Wang、Yuanchun Zhou 和 Hengshu Zhu。2025。SciRerankBench：面向科学检索增强生成大语言模型的重排序器基准测试。*arXiv* 预印本 *arXiv:2508.08742*（2025）。[9] Arman Cohan、Waleed Ammar、Madeleine van Zuylen 和 Field Cady。2019。科学出版物中的引文意图分类结构框架。载于*2019*年计算语言学协会北美分会年会：人类语言技术论文集，第*1*卷（长篇与短篇论文），Jill Burstein、Christy Doran 和 Thamar Solorio（编）。计算语言学协会，明尼苏达州明尼阿波利斯，3586–3596。doi:10.18653/v1/N19-1361[10] Florin Cuconasu、Giovanni Trappolini、Federico Siciliano、Simone Filice、Cesare Campagnano、Yoelle Maarek、Nicola Tonellotto 和 Fabrizio Silvestri。2024。噪声的力量：重新定义检索增强生成系统中的检索。载于第*47*届*ACM SIGIR*信息检索研究与发展国际会议论文集。719–729。[11] Hao Cui、Zahra Shamsi、Gowoon Cheon、Xuejian Ma、Shutong Li、Maria Tikhanovskaya、Peter Norgaard、Nayantara Mudur、Martyna Plomecka、Paul Raccuglia 等。2025。CURIE：评估大语言模型在多任务科学长上下文理解与推理方面的表现。*arXiv* 预印本 *arXiv:2503.13517*（2025）。[12] Franck Dernoncourt 和 Ji Young Lee。2017。PubMed 200k RCT：医学摘要中句子序列分类数据集。载于第八届自然语言处理国际联合会议（*IJCNLP 2017*），中国台湾台北，*2017*年*11*月*27*日至*12*月*1*日，第*2*卷：短篇论文，Greg Kondrak 和 Taro Watanabe（编）。亚洲自然语言处理联合会，308–313。https://aclanthology.org/I17-2052/[13] Jay DeYoung、Sarthak Jain、Nazneen Fatema Rajani、Eric Lehman、Caiming Xiong、Richard Socher 和 Byron C Wallace。2020。ERASER：用于评估可解释化自然语言处理模型的基准测试。载于计算语言学协会第*58*届年会论文集。4443–4458。[14] Aris Fergadis、Dimitris Pappas、Antonia Karamolegkou 和 Harris Papageorgiou。2021。面向可持续发展的科学文献论证挖掘。载于第八届论证挖掘研讨会论文集。100–111。[15] Hannah Fraser、Martin Bush、Bonnie C. Wintle、Fallon Mody、Eden T. Smith、Anca M. Hanea、Elliot Gould、Victoria Hemming、Daniel G. Hamilton、Libby Rumpff、David P. Wilkinson、Ross Pearson、Felix Singleton Thorn、Raquel Ashton、Aaron Willcox、Charles T. Gray、Andrew Head、Melissa Ross、Rebecca Groenewegen、Alexandru Marcoci、Ans Vercammen、Timothy H. Parker、Rink Hoekstra、Shinichi Nakagawa、David R. Mandel、Don van Ravenzwaaij、Marissa McBride、Richard O. Sinnott、Peter Vesk、Mark Burgman 和 Fiona Fidler。2023。通过 repliCATS（可信科学协作评估）流程中的结构化专家引 elicitation预测可靠性。*PLOS ONE* 18, 1（2023年1月），1–16。doi:10.1371/journal.pone.0274429[16] Yunfan Gao、Yun Xiong、Xinyu Gao、Kangxiang Jia、Jinliu Pan、Yuxi Bi、Yixin Dai、Jiawei Sun、Haofen Wang 和 Haofen Wang。2023。大语言模型的检索增强生成：综述。*arXiv* 预印本 *arXiv:2312.10997* 2, 1 2023。（）。[17] Renee A Hesselbach、David H Petering、Craig A Berg、Henry Tomasiewicz 和 Daniel Weber。2012。科学论文写作指南：聚焦高中至研究生阶段的学生研究。斑马鱼9, 4（2012），246–249。[18] Sarthak Jain、Madeleine van Zuylen、Hannaneh Hajishirzi 和 Iz Beltagy。2020。SciREX：文档级信息抽取挑战数据集。载于计算语言学协会第*58*届年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，Dan Jurafsky、Joyce Chai、Natalie Schluter 和 Joel R. Tetreault（编）。计算语言学协会，7506–7516。doi:10.18653/V1/2020.ACL-MAIN.670[19] Tom Jansen 和 Tobias Kuhn。2016。从科学文章中抽取核心论断。载于*BNAIC 2016*：人工智能——第*28*届比荷卢人工智能会议，荷兰阿姆斯特丹，*2016*年*11*月*10*日至*11*日，修订精选论文（计算机与信息科学通信，第*765*卷），Tibor Bosse 和 Bert Bredeweg（编）。Springer，32–46。doi:10.1007/978-3-319-67468-1_3

[20] Bowen Jin、Jinsung Yoon、Jiawei Han 和 Sercan Ö. Arik。2025。长上下文大语言模型遇上检索增强生成：克服检索增强生成中长输入的挑战。载于第十三届国际学习表征会议，*ICLR 2025*，新加坡，*2025*年*4*月*24*日至*28*日。OpenReview.net。https://openreview.net/forum?id=oU3tpaR8fm[21] Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yutaka Matsuo 和 Yusuke Iwasawa。2022。大语言模型是零样本推理器。载于神经信息处理系统进展第*35*卷：神经信息处理系统年度会议 *2022*，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月9日，Sanmi Koyejo、S. Mohamed、A. Agarwal、Danielle Belgrave、K. Cho 和 A. Oh（编）。http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html[22] Thomas A Lang 和 Douglas G Altman。2015。生物医学期刊论文的基本统计报告：“已发表文献中的统计分析与方法”或 SAMPL 指南。*International journal of nursing studies* 52, 1（2015），5–9。[23] Anne Lauscher、Goran Glavas 和 Kai Eckert。2018。ArguminSci：用于分析科学写作中的论证和修辞方面的工具。载于第五届论证挖掘研讨会论文集，*ArgMining@EMNLP 2018*，比利时布鲁塞尔，*2018*年*11*月*1*日，Noam Slonim 和 Ranit Aharonov（编）。计算语言学协会，
22 28 10 18653 1 18 5203 – 。[24] 2021
doi: 。/V /W - Xiangci Li、Gully Burns 和 Nanyun Peng。 。用于证据抽取的科学话语标注。载于计算语言学协会欧洲分会第*16*届会议论文集：主会场，Paola Merlo、Jorg Tiedemann 和 Reut Tsarfaty（编）。计算语言学协会，线上，
2550 2562 10 18653 1 2021 218 – 。doi: 。/v / .eacl-main。[25] 2025
Xiangci Li 和 Jessica Ouyang。 。知识选择如何助力检索增强生成？载于计算语言学协会论文集：*EMNLP 2025*，Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，
4104 4121 10 18653 1 2025 218 – 。doi: 。/v / .[26] findings-emnlp。Nelson F. Liu、Kevin Lin、John Hewitt、Ashwin Paranjape、Michele Bevilacqua、Fabio Petroni 和 Percy Liang。2024。迷失在中间：语言模型如何利用长上下文。计算语言学协会会刊 12 2024 157 173 10 1162 _ _00638 ( )，– 。[27] doi: 。/tacl a Yi Luan、Luheng He、Mari Ostendorf 和 Hannaneh Hajishirzi。2018。面向科学知识图谱构建的实体、关系和共指多任务识别。载于*2018*年自然语言处理实证方法会议论文集，比利时布鲁塞尔，*2018*年*10*月*31*日至*11*月*4*日，- *11*月 Ellen Riloff、David Chiang、Julia Hockenmaier 和 Jun'ichi Tsujii（编）。计算语言学协会，
3219 3232 10 18653 1 18 1360 – 。doi: 。/[28] 2021
V /D - Ian H. Magnusson 和 Scott E. Friedman。 。抽取科学主张的细粒度知识图谱：数据集与基于 Transformer 的结果。载于*2021*年自然语言处理实证方法会议论文集，*EMNLP 2021*，线上活动/多米尼加共和国蓬塔卡纳，*2021*年*11*月7日至*11*日，Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih（编）。计算语言学协会，4651–4658。doi:10.18653/V1/2021.EMNLP-MAIN.381[29] Bodhisattwa Prasad Majumder、Harshit Surana、Dhruv Agarwal、Bhavana Dalvi Mishra、Abhijeetsingh Meena、Aryan Prakhar、Tirth Vora、Tushar Khot、Ashish Sabharwal 和 Peter Clark。2025。DiscoveryBench：借助大语言模型迈向数据驱动的发现。载于第十三届国际学习表征会议，*ICLR 2025*，新加坡，*2025*年*4*月*24*日至*28*日。OpenReview.net。https://openreview.net/forum?id=vyflgpwfJW [30] Iain Marshall、Joël Kuiper、Edward Banner 和 Byron C. Wallace。2017。自动化生物医学证据综合：RobotReviewer。载于*ACL 2017*论文集：系统演示，Mohit Bansal 和 Heng Ji（编）。计算语言学协会，加拿大温哥华，7–12。https://aclanthology.org/P17-4002/ [31] Iain J. Marshall 和 Byron C. Wallace。2019。迈向系统综述自动化：在研究综合中使用机器学习工具的实用指南。*Systematic Reviews*
8, 1 2019 163 10 1186 13643 019 1074 9 ( )，。doi: 。/s - - -[32] Ronak Pradeep、Xueguang Ma、Rodrigo Nogueira 和 Jimmy Lin。2021。使用 VerT5erini 验证科学主张。载于第*12*届健康文本挖掘与信息分析国际研讨会论文集，*LOUHI@EACL*，线上，*2021*年*4*月*19*日，Eben Holderness、Antonio Jimeno-Yepes、Alberto Lavelli、Anne-Lyse Minard、James Pustejovsky 和 Fabio Rinaldi（编）。计算语言学协会，94–103。https://www.aclweb.org/anthology/2021.louhi-1.11/[33] Freda Shi、Xinyun Chen、Kanishka Misra、Nathan Scales、David Dohan、Ed H Chi、Nathanael Schärli 和 Denny Zhou。2023。大语言模型很容易被无关上下文分散注意力。载于国际机器学习会议。PMLR，31210–31227。[34] W. Scott Spangler、Angela D. Wilkins、Benjamin J. Bachman、Meena Nagarajan、Tajhal Dayaram、Peter J. Haas、Sam Regenbogen、Curtis R. Pickering、Austin Comer、Jeffrey N. Myers、Ioana Stanoi、Linda Kato、Ana Lelescu、Jacques J. Labrie、Neha Parikh、Andreas Martin Lisewski、Lawrence A. Donehower、Ying Chen 和 Olivier Lichtarge。2014。基于科学文献挖掘的自动化假设生成。载于第20届 ACM SIGKDD 知识发现与数据挖掘国际会议，KDD '14，美国纽约，2014年8月24日至27日，

2014，Sofus A. Macskassy、Claudia Perlich、Jure Leskovec、Wei Wang 和 Rayid Ghani（编）。ACM，1877–1886。doi:10.1145/2623330.2623667

[35] Giulio Starace、Oliver Jaffe、Dane Sherburn、James Aung、Jun Shern Chan、Leon Maksin、Rachel Dias、Evan Mays、Benjamin Kinsella、Wyatt Thompson、Johannes Heidecke、Amelia Glaese 和 Tejal Patwardhan。2025。《PaperBench：评估人工智能复现人工智能研究的能力》。载于第四十二届国际机器学习大会（*ICML 2025*），加拿大不列颠哥伦比亚省温哥华，*2025*年*7*月*13*日至*19*日（《机器学习研究论文集》第*267*卷），Aarti Singh、Maryam Fazel、Daniel Hsu、Simon Lacoste-Julien、Felix Berkenkamp、Tegan Maharaj、Kiri Wagstaff 和 Jerry Zhu（编）。PMLR / OpenReview.net。https://proceedings.mlr.press/v267/ starace25a.html

[36] Weiwei Sun、Lingyong Yan、Xinyu Ma、Shuaiqiang Wang、Pengjie Ren、Zhumin Chen、Dawei Yin 和 Zhaochun Ren。2023。ChatGPT 擅长搜索吗？探究大语言模型作为重排序智能体的表现。载于*2023* 年自然语言处理实证方法会议论文集（*EMNLP 2023*），新加坡，*2023* 年 *12* 月 *6* 日至 *10* 日，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，14918–14937 . doi:10.18653/V1/2023.EMNLP-MAIN.923

[37] Simone Teufel 和 Marc Moens . 2002. 科学文章摘要：实验 - 验证相关性与修辞状态。计算语言学28，4（2002），409–445。doi:10.1162/089120102762671936

[38] James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018年。FEVER：用于事实抽取与验证的大规模数据集。载于《*2018*年北美计算语言学协会分会会议论文集：人类语言技术，*NAACL-HLT 2018*，美国路易斯安那州新奥尔良，*2018*年*6*月*1*日至*6*日，第*1*卷（长篇论文），Marilyn A. Walker、Heng Ji 和 Amanda Stent（编）。计算语言学协会，809–819。doi:10.18653/V1/N18-1074

[39] Rosni Vasu、Cristina Sarasua 和 Abraham Bernstein . 2024. SciHyp：一个细粒度 - 数据集，用于描述科学论文中的假设及其组件。载于《语义网——*ISWC 2024*——第*23*届国际语义网会议，美国马里兰州巴尔的摩，*2024*年*11*月*11*日至*15*日，会议论文集，第*III*部分（计算机科学讲义，第*15233*卷）》，Gianluca Demartini、Katja Hose、Maribel Acosta、Matteo Palmonari、Gong Cheng、Hala Skaf-Molli、Nicolas Ferranti、Daniel Hernández 和 Aidan Hogan（编）。Springer，134–152。doi:10.1007/978-3-031-77847-6 8_

[40] David Wadden、Shanchuan Lin、Kyle Lo、Lucy Lu Wang、Madeleine van Zuylen、Arman Cohan 和 Hannaneh Hajishirzi。2020。《事实还是虚构：科学主张验证》。载于《*2020*年自然语言处理实证方法会议论文集》（*EMNLP 2020*，线上，*2020*年*11*月*16*日至*20*日），Bonnie Webber、Trevor Cohn、Yulan He 和 Yang Liu（编）。计算语言学协会，7534–7550。doi:10.18653/V1/2020.EMNLP-MAIN.609

[41] David Wadden、Kyle Lo、Lucy Lu Wang、Arman Cohan、Iz Beltagy 和 Hannaneh Hajishirzi。2022。MultiVerS：通过弱监督和全文上下文改进科学主张验证。载于《计算语言学协会北美分会第 *2022* 届年会论文集：*NAACL 2022* 研究发现》，美国华盛顿州西雅图，*2022*年*7*月*10*日至*15*日（*ACL*会议论文集，*NAACL 2022* 卷），Marine Carpuat、Marie-Catherine de Marneffe 和 Iván Vladimir Meza Ruíz（编）。计算语言学协会，61–76。doi:10.18653/V1/2022.FINDINGS-NAACL.6

[42] Xin Wei、Md Reshad Ul Hoque、Jian Wu 和 Jiang Li。2023。ClaimDistiller：基于监督式对比学习的科学主张抽取。载于第*4*届科学文献知识实体抽取与评估联合研讨会（*EEKE2023*）与第*3*届人工智能 + 信息计量学（*AII2023*）联合举办，会议与 *JCDL 2023* 同期举行，地点为美国新墨西哥州圣菲及线上，*2023*年*6*月*26*日（*CEUR*研讨会论文集，第*3451*卷），Chengzhi Zhang、Yi Zhang、Philipp Mayr、Wei Lu、Arho Suominen、Haihua Chen 和 Ying Ding（编）。CEUR-WS.org，65–77。https://ceur-ws.org/Vol-3451/paper11.pdf

[43] Elizabeth K 从科学期刊文章全文中提取假设与证据。载于*2011*年生物医学自然语言处理研讨会论文集，*BioNLP@ACL*，美国俄勒冈州波特兰，*2011*年*6*月*23*日至*24*日，Kevin Bretonnel Cohen、Dina Demner- Fushman、Sophia Ananiadou、John Pestian、Jun'ichi Tsujii 和 Bonnie L. Webber（编）。计算语言学协会，134–135。https://aclanthology. org/W11-0217/

[44] Jian Wu 和 Sarah Rajtmajer。2025。《什么是研究假设？》*arXiv* 预印本 *arXiv:2509.00185*（2025）。[45] Siye Wu、Jian Xie、Jiangjie Chen、Tinghui Zhu、Kai Zhang 和 Yanghua Xiao。2024。《无关输入多容易扭曲大语言模型的响应？》*arXiv* 预印本 *arXiv:2404.03302*（2024）。[46] Wenhan Xiong、Jingyu Liu、Igor Molybog、Hejia Zhang、Prajjwal Bhargava、Rui Hou、Louis Martin、Rashi Rungta、Karthik Abinav Sankararaman、Barlas Oguz、Madian Khabsa、Han Fang、Yashar Mehdad、Sharan Narang、Kshitiz Malik、Angela Fan、Shruti Bhosale、Sergey Edunov、Mike Lewis、Sinong Wang 和 Hao Ma。2024。《基础模型的有效长上下文扩展》。载于《*2024*年北美计算语言学协会年会论文集：人类语言技术》（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，4643–4663。doi:10.18653/v1/2024.naacl-long.260

[47] OriYoran、TomerWolfson、Ori Ram 和 JonathanBerant。2024。《让检索增强语言模型免受无关上下文干扰》。载于第十二届国际学习表征会议，*ICLR 2024*，奥地利维也纳，*2024*年*5*月*7*日至*11*日。OpenReview.net。https://openreview.net/forum.id= ZS4m74kZpH [48] Haozhen Zhang、Tao Feng、Pengrui Han 和 Jiaxuan You。2025。《AcademicEval：实时长上下文大语言模型基准测试》。机器学习研究汇刊 2025（2025）。https: //openreview.net/forum.id=LjQ4voE5bs

A 诊断分析：上下文信号密度（假设抽取）

对于每个假设阶段上下文，我们计算相关句子比例（RSP），其定义为：余弦相似度超过 τ=0.89 的句子所占比例。相似度使用 Gemini 嵌入计算，阈值与第 5.1 节假设评估所用阈值相同。我们按 RSP 对论文分箱（0–1%、1–2%、2–3%、3–4%、4–5% 和 > 5%），并在图 4 中报告每个分箱内的平均假设提取F1分数。点的大小与该分箱中的论文数量成正比。

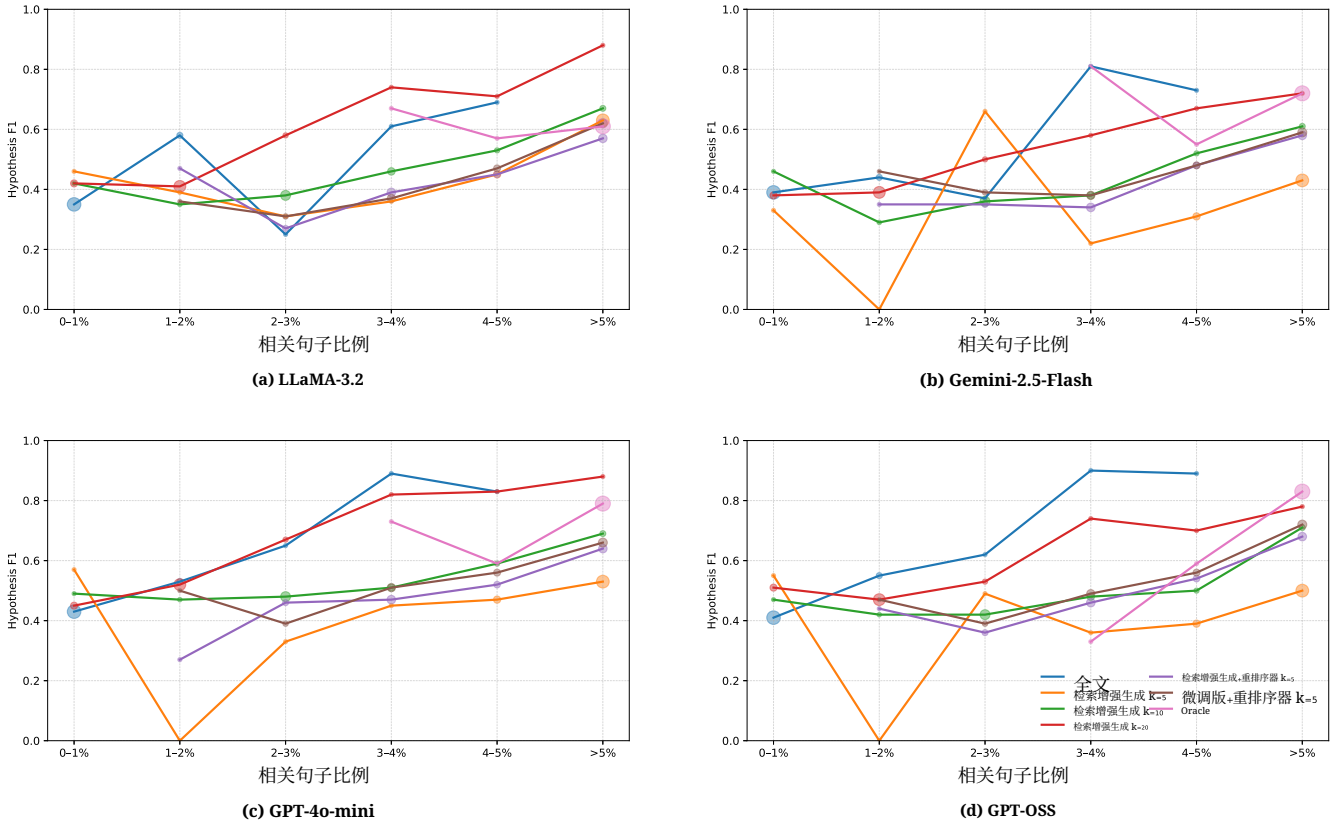
B 诊断分析：逐篇论文转变分析：数量与质量（证据抽取）

为隔离检索到的上下文对证据抽取的影响，排除假设阶段错误的干扰，我们以真实假设为条件，仅改变证据阶段上下文的构建方式。我们比较两类转变：一类改变上下文数量（标准检索增强生成，使用 $k \in \{5, 10, 20\}$ 和全文），另一类在固定最终上下文大小的同时改变上下文质量（稠密检索 k=5 → 重排序 k=5 → 微调检索器+k=5）。

对于每种转换（源 → 目标），我们根据标准证据段落 g 是否出现在检索到的证据阶段上下文 C_{src} 和 C_{tgt}中，对每篇论文进行分类。如果 g 同时出现在两个上下文中（$g \in C_{\text{src}}$ 和 $g \in C_{\text{tgt}}$），则将该论文标记为 GT_保留；如果 g 在源上下文中缺失，但出现在目标上下文中（$g \notin C_{\text{src}}$ 和 $g \in C_{\text{tgt}}$），则标记为 GT_新增。对于目标上下文不保证包含源上下文的转换（例如重排序或切换检索器），我们还会跟踪 GT_丢失案例，即g 存在于源上下文中，但未保留在目标上下文中（$g \in C_{\text{src}}$和 $g \notin C_{\text{tgt}}$）；并将其与 GT_缺失案例区分开来，后者指g 在两个上下文中均不存在（$g \notin C_{\text{src}}$ 和 $g \notin C_{\text{tgt}}$）。在每个类别中，我们计算每篇论文的证据 F1变化均值， $\Delta\text{F1} = \text{F1}_{\text{tgt}} - \text{F1}_{\text{src}}$。我们在图 8 中汇总了每篇论文的转换结果，并在表 2 中报告相同真实值—假设设置下相应的汇总精确率/召回率/F1。

我们发现，证据抽取性能的变化主要取决于检索变更是否将金标准证据段落加入上下文。在数量消融实验中，更大的 k所带来的改进主要出现在 GT_Gained 情况下（即扩展 k后新检索到了 g）。当金标准段落已经存在时（GT_Kept），扩展上下文（例如 k=20 → 全文）带来的收益有限，甚至可能导致性能下降。

相比之下，检索质量改进（重排序和微调）带来的收益更加稳定，因为在固定上下文大小下，这些改进提高了 GT_Gained 结果的发生率。然而，质量变更并非单调改进：在 GT_Lost 情况下，微调可能带来小幅权衡，即 g 在先前的方法下能够被检索到，但在检索器/重排序器变更后未能保留。



(a) LLaMA-3.2 (b) Gemini-2.5-Flash

(c) GPT-4o-mini (d) GPT-OSS

图 4：假设提取 n F1分数与相关句子比例（RSP）的分箱关系。各点表示每个分箱内的平均 F1分数，点的大小与每个箱中的论文数量成正比。

C 语义相似度校准

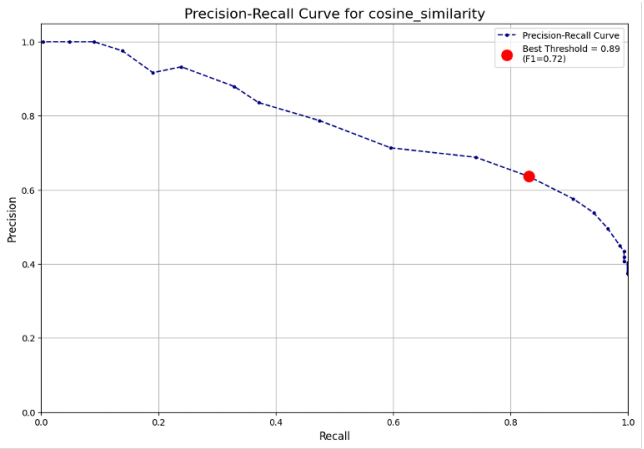


图5：使用精确率-召回率曲线校准余弦相似度阈值。该图展示了 Gemini嵌入模型在不同阈值下的数值。所选阈值 0.89 对应于人工标注集上 F1分数的最大值（0.72）。

示例：嵌入相似度较高时的假设抽取错误 **余弦相似度：**0.875 (87.5%)

DOI 10.1016/j.jesp.2017.07.008

人工标注。存在替代选项会对参与者产生竞争性的、隐蔽的影响。

抽取结果。 我们假设，参与者更可能做出与政治阴谋论陈述 字面内容相反的回应，因为威胁防护需求和准确性偏见等相互竞争的隐性影响，会在阴谋论中吸引人和不吸引人的方面之间造成张力，从而使回应与基线陈述相比呈现相反倾向。

诊断。 词汇／语义重叠较高（“隐性影响”），但该预测将核心构念和机制从**备选选项** 转变为**对阴谋论陈述的回应**，并引入了人工标注假设中未提及的其他构念。

图6：嵌入相似度高但假设抽取不正确的定性示例。

示例：用于文档内检索器训练的困难负例构造

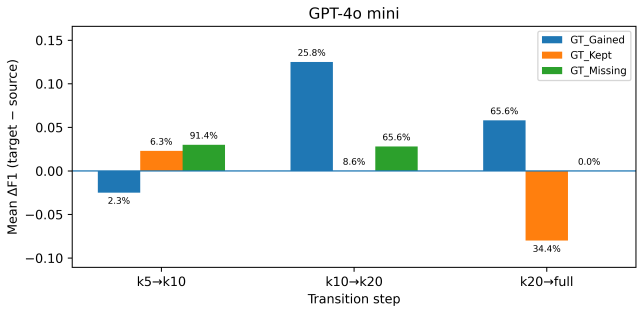
DOI. 10.1016/j.jesp.2016.01.014

人工标注。 我们预期，经历因果不确定性的参与者在中断任务与较抽象（相较于具体）思维相关联时，会表现出更高的任务恢复率。

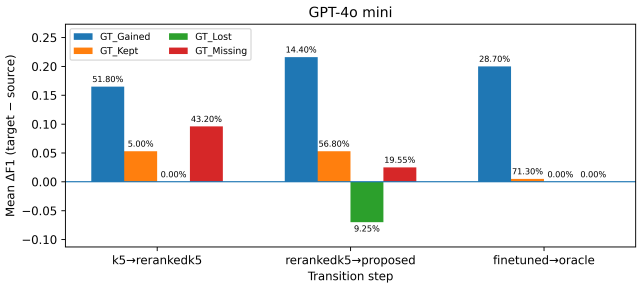
正向段落（包含人工标注假设）。回到我们的主要预测，我们预期，经历因果不确定性的参与者在被中断的任务与更抽象（相对于更具体）的思维方式相关联时，会表现出更高的任务恢复率。为了衡量任务恢复，我们让参与者有机会参与一项更令人愉悦的任务，即阅读幽默故事。先导研究证实，参与者确实认为幽默评估任务是更令人愉悦的替代选择。...

困难负例段落（主题相似，修辞作用不同）。本实验采用的任务中断与恢复范式排除了这一替代解释，因为参与者可以选择的替代任务是阅读幽默故事。先导研究证实了这一点，结果显示，参与者认为幽默评估任务是最有趣、最令人愉悦的。如果参与者的目标只是改善情绪状态，而不是进行更抽象的思考，那么我们就不会发现所展示的效应...

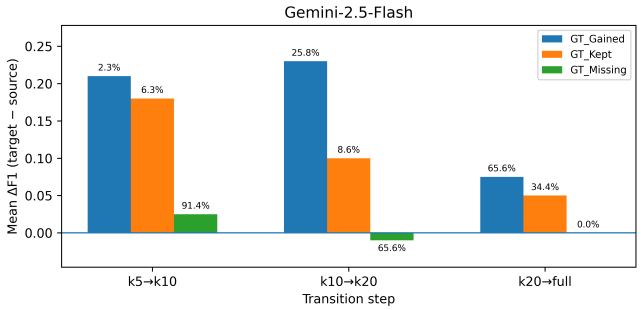
图7：用于训练检索器的困难负例选择示例。困难负例在语义上与金标准段落相似，但相对于提取目标，其修辞功能有所不同。



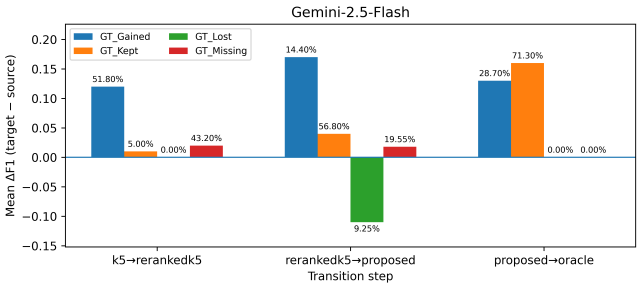
(a) GPT-4o-mini（数量）



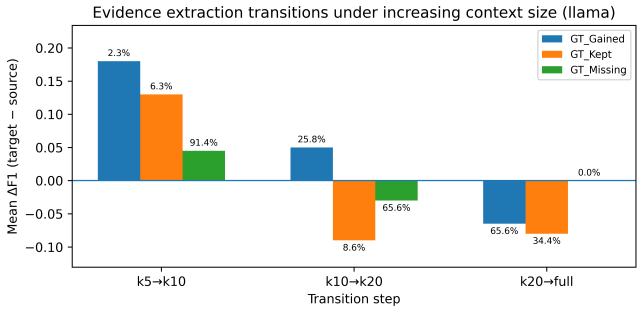
(b) GPT-4o-mini（质量）



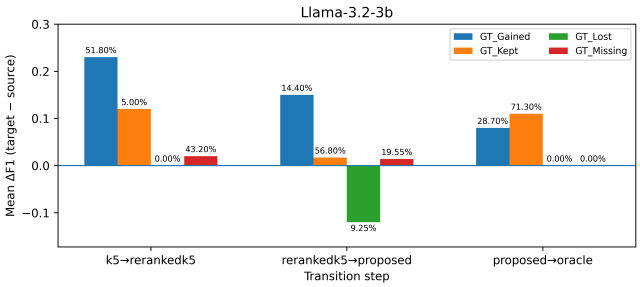
(c) Gemini-2.5-Flash（数量）



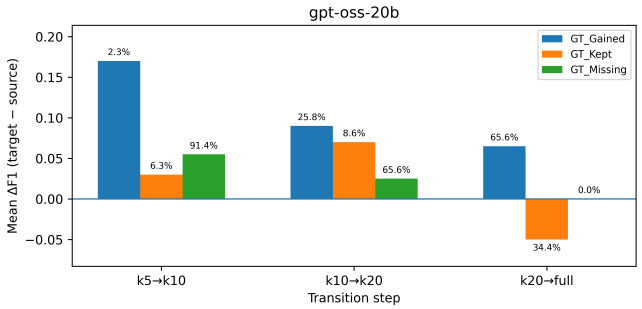
(d) Gemini-2.5-Flash（质量）



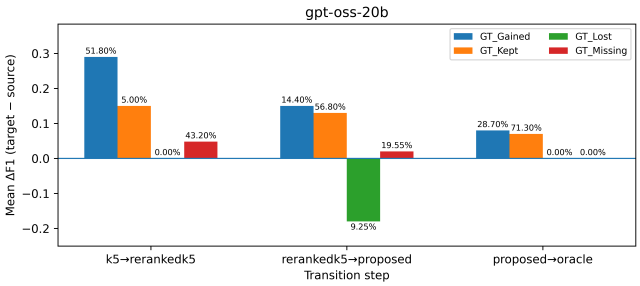
(e) LLaMA-3.2（数量）



(f) LLaMA-3.2（质量）



(g) GPT-OSS（数量）



(h) GPT-OSS（质量）

图 8：证据抽取在数量与质量变化下的逐篇论文转变分析。左列改变检索到的上下文数量（k=5 → 10 → 20 → 全文）。右列在上下文大小固定的情况下改变检索质量（k=5）（检索增强生成 →检索增强生成+重排序器 → 微调检索器+重排序器 → Oracle）。柱形表示每个检索结果分桶中的逐篇论文平均 ΔF1分数；柱形上方的标签显示该分桶中的论文比例。对于质量转变，当金标准证据段落存在于源上下文中、但不存在于目标上下文中时，我们还报告 GT_Lost。

LLM	方法	证据		
		P	R	F1
Llama 3.2	检索增强生成 (k=5)	0.19	0.25	0.22
	检索增强生成 (k=10)	0.24	0.32	0.27
	检索增强生成 (k=20)	0.25	0.29	0.26
	全文	0.21	0.28	0.24
	检索增强生成 + 重排序器 (k=5)	0.29	0.43	0.35
	微调后的 SciBERT + 重排序器 (k=5)	0.31	0.48	0.38
	预言机上下文	0.43	0.52	0.47
GPT-OSS	检索增强生成 (k=5)	0.19	0.20	0.20
	检索增强生成 (k=10)	0.23	0.28	0.25
	检索增强生成 (k=20)	0.25	0.34	0.29
	全文	0.28	0.36	0.31
	检索增强生成 + 重排序器 (k=5)	0.34	0.44	0.38
	微调后的 SciBERT + 重排序器 (k=5)	0.43	0.49	0.46
	预言机上下文	0.50	0.57	0.53
gpt-4o-mini	检索增强生成 (k=5)	0.22	0.40	0.28
	检索增强生成 (k=10)	0.24	0.45	0.31
	检索增强生成 (k=20)	0.28	0.49	0.36
	全文	0.35	0.40	0.37
	检索增强生成 + 重排序器 (k=5)	0.39	0.45	0.41
	微调后的 SciBERT + 重排序器 (k=5)	0.43	0.51	0.47
	预言机上下文	0.49	0.63	0.53
gemini-2.5-flash	检索增强生成 (k=5)	0.21	0.39	0.27
	检索增强生成 (k=10)	0.24	0.44	0.31
	检索增强生成 (k=20)	0.29	0.50	0.37
	全文	0.33	0.48	0.39
	检索增强生成 + 重排序器 (k=5)	0.27	0.48	0.34
	微调后的 SciBERT + 重排序器 (k=5)	0.31	0.53	0.38
	预言机上下文	0.51	0.59	0.53

表 2：基于真实假设的证据抽取。我们向抽取器提供人工标注的（真实）假设，并仅改变证据阶段上下文配置（检索增强生成 $k \in \{5, 10, 20\}$、重排序和微调检索），从而隔离证据阶段。报告的精确率/召回率/F1 使用与主要结果相同的基于组件的评估器计算。