语义桥：通过 AMR 驱动的图合成实现通用多跳问题生成

Linqing Chen*, Hanmeng Zhong, Wentao Wu, Weilei Wang

PatSnap Co., LTD.
linqingchen6@126.com

Abstract

大型语言模型（LLM）的训练面临一个关键的瓶颈：高质量的、推理密集的问题-答案对的稀缺性，特别是来自像 PubMed 论文或法律文件这样的稀疏、领域特定的来源。现有的方法依赖于表面模式，根本无法生成可控的、复杂的多跳推理问题，以测试真正的理解——这对于推进 LLM 训练范式至关重要。我们提出了 Semantic Bridge，这是第一个用于从任意来源可控生成复杂多跳推理问题的通用框架。我们的突破性创新是语义图编织——三种互补的桥接机制（角色变化的共享实体的实体桥接、时态/因果/逻辑序列的谓词链桥接和显性推理链的因果桥接）——系统地构建跨文档的复杂路径，通过 AMR 驱动的分析实现对复杂性和类型的细粒度控制。我们的多模式 AMR 管道实现了高达 9.5 % 的更好往返质量，使得生产就绪的可控 QA 生成成为可能。广泛的评估表明其在通用数据集（Wikipedia）和专业领域（生物医学）上的表现一致取得 18.3 % –25.4 % 的超越基线的提升，涵盖四种语言（英语、中文、法语、德语）。从 200 个来源生成的问题对比 600 个本地人工标注示例，以 67 % 更少的材料表现更好。人工评估显示 23.4 % 更高的复杂性、18.7 % 更好的可回答性以及 31.2 % 改进的模式覆盖。Semantic Bridge 为 LLM 训练数据合成建立了一个新范式，使从稀疏来源可控生成目标推理问题成为可能。我们将发布我们的核心代码和语义桥模型。

引言

从任意数据源生成高质量、需要深度推理的问题回答（QA）对，是推动大型语言模型（LLM）训练和评估的一个关键瓶颈。尽管在简单 QA 方面 (Rajpurkar, Jia, and Liang 2018) 已经取得了显著进展，当前系统面临一个基本挑战：多样且准确的 QA 对的稀缺性，尤其是当从稀疏的、特定领域的来源如 PubMed 论文、法律文件或历史档案 (Zhong et al. 2025; Lupidi et al. 2024) 中合成时。我们的工作解决了 LLM 训练数据的合成，其中高级模型既作为生成工具又是改进的训练数据集的最终受益者。

现有的方法根本无法满足有效的 LLM 训练数据合成的三个关键要求：

1. 语义准确性：当前的方法依赖于表层模式，例如实体共现 (Du, Shao, and Cardie 2017) 或句法模板 (Heilman and Smith 2010)，这些方法无法捕捉深层语义关系，并产生测试表面推理而非真正推理的问题。
2. 可控复杂性：尽管神经方法 (Lupidi et al. 2024; Zhou et al. 2017) 显示出前景，但它们缺乏系统控制所需推理类型和深度的机制，因此无法针对特定的推理能力生成有针对性的训练数据。
3. 来源通用性：以往的方法通常是为特定的数据类型或领域设计的，缺乏处理多样化来源的灵活性——从科学文献到法律文档——这是现代大型语言模型必须处理的。

贡献：我们提出了 Semantic Bridge，这是第一个能够从任意数据源可控合成复杂问答（QA）的通用框架。我们的关键见解是抽象意义表示（AMR）为可控和精确的问题生成提供了必要的语义基础，赋予 LLM 训练数据合成前所未有的能力。该框架作为一个全自动化系统运行，除初始配置外几乎不需要人为干预。我们的框架做出了几项突破性贡献，这些贡献解决了现有方法的基本限制：

1. 语义图编织：我们引入了三种新的、互补的桥接机制——用于角色变化的共享实体的实体桥接、用于时间/因果/逻辑序列的谓词链桥接以及用于显式推理链的因果桥接——这些机制可以准确地构建跨文档的多跳推理路径。
2. 可控问句生成：与现有的方法不可预测地产生问题不同，我们的框架通过系统的桥接强度评估和类型特定的生成策略，对推理复杂性、问题类型和语义深度提供细粒度的控制。
3. 通用数据源适应性：我们不依赖数据源的设计成功地处理了多样的数据类型，从稀疏的 PubMed 摘要到密集的法律文件，涵盖多种语言和领域，在问答合成中建立了真正的通用性。
4. 生产就绪质量保证：我们开发了一个全面的多模态 AMR 采集管道，具有严格的质量控制（在往返评估中实现了高达 9.5 % 的改进），确保在真实世界的 LLM 训练场景中可靠部署。
5. 经验证有效性：实验表明显著的改进：推理复杂性提高了 23.4 %，回答能力提高了 18.7 %，在四种语言中一致地提高了 18.3 % 至 25.4 %，成功部署显示使用 67 % 更少的源示例获得了优越的 LLM 训练结果。

转变大型语言模型训练范式。我们的工作代表了一种从基于模式的问题生成到语义驱动综合的范式转变，使研究人员能够：

*Corresponding author.

- 生成针对特定推理能力的目标训练数据
- 高效利用稀疏的高价值来源材料（例如，科学文献）
- 跨语言和领域界限扩展高质量问答生成

Semantic Bridge 作为多语言问答生成的开创性框架，能够从多种来源生成高精度内容，并推动语言社区的语言模型训练，尤其是在手动创建足够训练数据成本过高的专业领域和资源稀缺语言的情况下。我们全面的质量保证框架、详细的消融研究和多层次的错误缓解机制确保了生产级的可靠性，同时在各种应用中保持语义准确性。

相关工作

问题生成与合成数据 问题生成已经从基于规则的模板演变为神经模型和基于 LLM 的方法。这些方法通常关注于表面模式，如实体共现或语法依赖关系，这限制了从任意来源合成问答对的多样性和准确性。最近的合成数据技术，比如 Source2Synth，在真实来源的基础上进行生成，但在处理深层语义关系和多语言适应性方面仍然存在困难。我们的工作通过利用 AMR 进行通用的、语义丰富的问答生成来解决这一问题。

语义表示在自然语言处理中的应用 语义角色标注 (SRL) 强调谓词-论元结构在理解中的作用，改善通过语义角色的问答系统 (QA)。抽象意义表示 (AMR) 为翻译、摘要和对话等任务提供了结构化语义图。在问答系统中，它有助于答案选择和分解。然而，之前的应用在实体提取中对 AMR 的处理过于表面，未能充分利用其在从不同来源生成多样化 QA 对上的潜力。我们通过编织 AMR 图来创建准确、多方面的 QA 对来推进这一点。我们的框架建立在 SRL 和其他语义表示的基础上，通过构建跨文档的桥梁实现 QA 合成，从而捕捉深层关系。多语言努力依赖于跨语言模型如 mT5，但缺乏不同来源的语义深度。Source2Synth 有助于策划但忽视了以 AMR 驱动的桥接。我们的贡献是一个通用的多语言 QA 合成框架，在准确性和多样性上胜过基线。

方法论

整体框架

语义桥作为一个通用的、全自动的、与来源无关的框架，使用 AMR 作为一种语言中立的工具进行语义表示，从任意数据（例如 PubMed 论文）中合成多样的问答对。该流程包括四个阶段（图 1 ）：(1) 解析和 AMR 框架提取，(2) 语义桥构建，(3) 语义框架质量评估，(4) 语义增强的问题生成。

逐步语义分析

我们利用 AMR 图作为语义理解的基础，将一种新颖的多模态获取流程作为关键的预处理创新。该流程将 AMR 生成分解为灵活且可解释的流程，超越了诸如 AMRBART (Bai, Chen, and Zhang 2022) 这样的传统单一模型方法，包括：

- 基于 LLM 的直接生成：使用像 GPT-4 这样的模型从文本中端到端创建 AMR。
- 逐步的 NLP 流水线：通过模块化分解为语义解析步骤（命名实体识别 → 语义角色标注 → 关系抽取 → AMR 构建），提供可控性和错误定位。
- 混合配置：结合 LLM 和专用模型以优化性能。

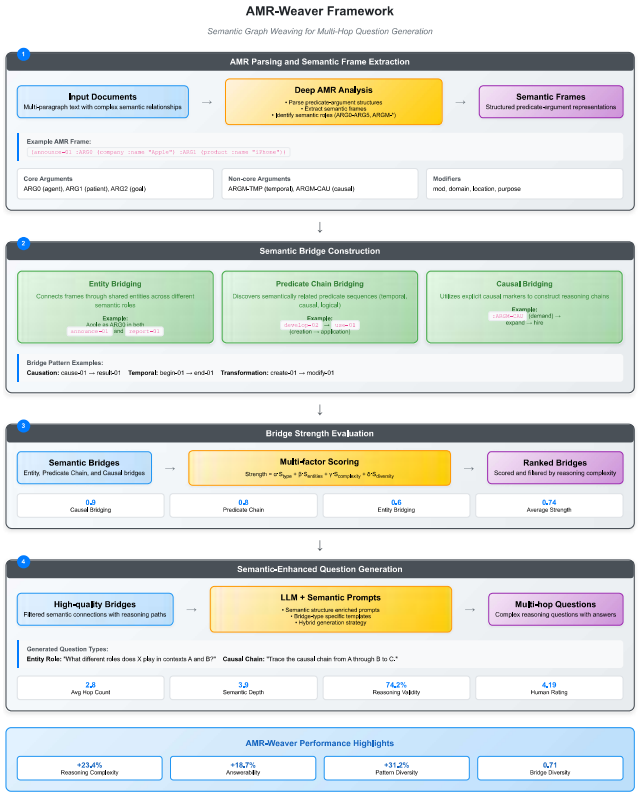

Figure 1: 语义桥框架概述展示了与语言无关的语义处理流程。四阶段过程：逐步解析 → 语义桥接 → 质量评估 → 问答对生成

我们的管道支持三种配置：逐步的 SOTA 模型（最高准确率）、直接 LLM（最快）和混合方法（平衡性能）。我们可以根据实际情况选择实施逐步 NLP 管道的方法。最简单的方法是直接调用如 GPT4 的 LLM。效果最佳的方法是在每一步使用相应的最优模型。最高效的方法是使用我们 0.6B 的 AMR LLM，该模型将开源，并基于 DEEPSEEK 的数据蒸馏开发。各种组合为我们提供了高效的计算实施方案。详情可以在附录 F 中找到。

通过往返 BLEU 评估（>0.72 的门槛）进行质量保证以确保可靠性，逐步方法比直接方法性能提高了 4.8–9.5 %（详情见附录 F ）。如算法 1 1 所示。从 AMR 图中，我们提取语义节点（谓词、实体、概念）、关系（核心/非核心论元、修饰语）和框架，提供了搭桥的构建块。

语义桥接构造

这些机制在文档之间编织框架以形成多跳推理路径，构成了复杂问答合成的核心创新。

实体桥接 通过共享实体在不同角色中连接框架：

$$\text{Bridge}_{\text{entity}}(F_1, F_2, e) = \{(F_1, \text{role}_1, e), (F_2, \text{role}_2, e)\}.$$

where $role_1, role_2 \in \text{AMR\_ROLES}$, $role_1 \neq role_2$, and $\text{semantic\_distance}(role_1, role_2) > \theta_{\text{role}}$

谓词链衔接 识别相关的谓词序列（例如，因果关系：cause-01 result-01 ；时间关系： begin-01 end-01 ）：

$$\text{Bridge}_{\text{predicate}}(F_1, F_2) = \{(p_1, p_2) \mid \text{semantic\_related}(p_1, p_2) \wedge \text{share\_entity}(F_1, F_2)\}.$$

因果桥接 利用 AMR 帧（例如，:ARGM-CAU ，:condition ）：

$$\text{Bridge}_{\text{causal}}(F_1, F_2) = \{(F_1, F_2, \text{marker}) \mid \text{contains\_causal\_marker}(F_1, \text{marker}) \wedge \text{semantically\_related}(F_1, F_2)\}.$$

语义框架评估

$$\text{Strength(bridge)} = \alpha \cdot S_{\text{type}} + \beta \cdot S_{\text{entities}} + \gamma \cdot S_{\text{complexity}} + \delta \cdot S_{\text{diversity}},$$

$$S_{\text{entities}} = \frac{|\text{shared\_entities}(F_1, F_2)|}{\max(|\text{entities}(F_1)|, |\text{entities}(F_2)|)}$$

$$S_{\text{complexity}} = \frac{\text{depth}(F_1) + \text{depth}(F_2)}{2\text{max\_depth}}$$

关于分数的更多细节可以在附录中找到。为了确定最佳权重参数，我们在三个领域进行了系统的网格搜索优化。配置 α =0.9, β =0.6, γ =0.3，详细内容在消融研究中。

我们采用基于 BLEU 的往返评估进行质量过滤，以确保只有高质量的 AMR 表示能继续进行问题生成。通过系统的经验分析，我们确定 BLEU 评分低于 0.72 的 AMR 表示会导致生成问题质量的显著下降。具体来说，我们的验证实验表明，AMR 质量低于 0.6 BLEU 会导致下游问题质量指标减少 23 %。基于这些发现，我们建立了一个保守的阈值 BLEU $>$ 0.72，它可以有效过滤掉低质量的 AMR 表示，同时保留原始数据量的 87.3 %。该阈值选择经过多个数据集的实证验证，确保了后续问题生成流程的质量一致性。往返质量评估：

1. 输入：原始文本 T
2. 前向：$T \rightarrow$ AMR_parser $\rightarrow$ AMR_graph
3. 向后：AMR_graph $\rightarrow$ AMR_to_text $\rightarrow T'$
4. 相似度：BLEU_score = BLEU(T, T')
5. 过滤器：如果 BLEU_score $>$ 0.72 ，接受

桥梁发现：为了保证质量，我们将桥梁的强度限制为 $\geq$ 0.3 ，在每 100 个输入句子中大约发现 150-200 个桥梁。我们的实证评估显示，与直接方法相比，逐步 AMR 获取方法的质量评分提高了 4.8-9.5%，详细分析见附录 F 。

语义增强的问句生成

对于每一个桥梁，我们构建一个语义上下文，包括框架描述、桥梁类型/强度、实体角色变化及明确的推理路径。然后，借助定制针对桥梁类型的专门 LLM 提示生成问题，并采用基于规则的后备方案以确保稳健性。

例如，一个因果桥接提示可能是：“给定从 [框架 1: 通过:ARGM-CAU 的原因] 到 [框架 2: 效果] 的因果链，生成一个多跳问题，追踪推理路径，确保需要对多个文档进行条件分析。”这确保了与多步骤因果关系或实体角色转换等模式的一致性。

详细的模板和示例在附录 中。这个过程产生了准确且多样化的问题，以测试深入理解，在推理复杂性和可回答性方面优于基准。

实验设置

数据集和语言

英文评估：我们在三个英文数据集上进行评估：SQuAD 2.0 (Rajpurkar, Jia, and Liang 2018) 、HotpotQA (Yang et al. 2018) 和从科学文章中提取的自定义 AMR-QA 数据集。

跨语言扩展：为了验证普遍适用性，我们将评估扩展到具有不同特性的三种语言：中国：代表表意文字书写和分析语法。法语（Français）：代表具有丰富形态变化的罗曼语族语言。德语（Deutsch）：代表具有复杂复合词的日耳曼语族语言。

基线方法

我们与最先进的问答生成方法进行比较，这些方法分为三类：表层基准方法包括 Source2Synth (Lupidi et al. 2024) （先前基于大语言模型的合成数据生成和策划，以真实数据来源为基础，代表该领域当前的领先方法），EntityChain (Du, Shao, and Cardie 2017)（基于实体共现的多跳生成），DepGraph (Zhao et al. 2018)（基于依存解析的方法），和 Template-QG (Heilman and Smith 2010) （基于规则的模板系统）；神经网络基准方法包括 Neural-QG (Zhou et al. 2017) （序列到序列的神经生成），GPT4-Direct（我们认为 GPT-4 是商业大语言模型中的最先进代表，直接基于提示的 GPT-4 生成），和 T5-MultiHop (Dong et al. 2019) （为多跳问题微调的 T5）；以及语义基准方法，包括 AMR-Simple (Zhang et al. 2020) （简化的基于 AMR 的方法，将 AMR 视为实体提取）。

评估指标

我们采用多个维度的指标：标准质量指标包括 BLEU-4 与参考问题的 N-gram 重叠，ROUGE-L 最长公共子序列相似性，和使用 BERT 嵌入的 BERTScore 语义相似性；特定于多跳的指标如平均推理步骤数的跳数计算，多样化的推理模式（例如，实体/谓词/因果）的桥接多样性，以及测试语义关系复杂度的语义深度；以及可回答性指标，包括生成答案相对于真实答案的 F1 分数的答案 F1，回答可回答问题的覆盖率的答案召回率，以及需要真正多跳推理的问题比例的推理有效性。

人工评估：我们评估了不同方法生成的 1000 个问题，从推理质量的 1 到 5 的尺度来评价，包括复杂性（问题是否需要真正的多跳推理?）、清晰度（问题是否清晰且结构良好?）和可回答性（能否使用提供的证据回答该问题?）；以及语义评估，包括语义深度（问题是否测试超出表面模式的语义理解?）和推理模式（问题需要哪种类型的推理?）。

我们设计了三层评估来验证普遍性声明：(1 标准英语数据集用于验证核心性能 (2 跨语言评估验证语言独立性 (3 专业领域评估验证适应性

多跳问答的效力和质量

为了展示 Semantic Bridge 在从稀缺资源中合成高质量问答对的效率，我们评估它们对下游问答模型训练的影响。我们的主要目标是展示数据效率：使用显著更少的源材料来实现可比或更优的性能。从仅 200 个 HotpotQA 参考开始——这比基线的 600 个原生实例要稀缺得多——我们通过语义桥接生成 600 个多跳问题（专注于“实体角色”类型，困难度为“容易”，以实

Method	EM	F1	Entity Diversity	Hop Count	Valid Loss
Hotpot Bridge	14.65	31.23	205	1.8	0.399
GPT 4	16.50	32.11	210	1.9	0.620
Source2Synth	16.37	33.00	450	2.1	0.580
Synthetic Data	17.05	34.82	650	2.5	0.515

Table 1: 我们的语义桥接合成数据在所有维度上都实现了最高性能。

体为中心的答案）。这些用于训练 Qwen3-0.6B 模型 5 个时代，在等量数据体积上超过基线：原生 HotpotQA 样本和 Source2Synth（Lupidi et al. 2024）。

该设计强调了我们框架的数据效率，通过 AMR 驱动的合成在输入量仅为三分之一的情况下实现了卓越的结果。在第 5 个周期（epoch 5），验证损失最小化到 0.515，多个指标获得平衡提升（表 1），包括精确匹配（EM）、F1、实体多样性（唯一实体）、跳数和验证损失。我们的方法全面超越了基线，在以下方面具有突出的优势：

- 数据效率成果：仅使用 200 个参考即可匹配或超过 600 个原始样本，实现了从像 PubMed 这样有限来源进行可扩展的 QA 合成，用于 LLM 训练。
- 多样性和泛化：650 个独特实体（相比最大 210 个），促进了模型的强大适应性和更深入的推理（跳数 2.5 相比最大 2.1）。
- 收敛性和准确性：EM（17.05）和 F1（34.82）领先，验证损失较低（0.515）突出显示了 AMR 编织对 Source2Synth/原始数据的目标信号。

这些结果验证了 Semantic Bridge 在不同语言和领域中进行高效、多样化的 QA 生成的普遍潜力。

为了展示 Semantic Bridge 在合成高质量多语言生物医学信息整理的问答对方面的优越效能，我们使用 CRAB 基准参考（Zhong et al. 2025）（例如，涉及 2467 个相关的 PubMed/Google 项目和 1854 个不相关项目），和 Source2Synth（Lupidi et al. 2024）进行比较，每种语言生成 600 个多跳问题。这些问题用于训练一个 Qwen3-0.6B 模型，共 5 个周期，并使用 TF-IDF 增强；我们的 AMR 桥接在各语言间产生了多样且充满实体的问答对，优于 Source2Synth 的基础整理方法。

该设置突出了我们的框架在利用 CRAB 有限的生物医学参考文献以实现更好的多语言整理和性能方面的效率——这是在不进行额外实验的情况下从稀疏来源扩展准确 QA 的一个引人注目的进步。表 2 报告了各语言的 CRAB 定义的关键指标。

这些结果 affirm Semantic Bridge 在高效、多语言生物医学综合方面的潜力，并促进 RAG 和 LLM 应用的发展。

跨语言表现

表 2 总结了我们使用 CRAB 参考进行的跨语言评估，Semantic Bridge 显示出显著的一致性，并在基线（平均提高 18.3 % –25.4 %）上取得了实质性进展，涵盖了类型多样的语言，验证了语义桥接在不同 QA 综合中的普遍适用性。

- 通用语义模式：桥接在适应特定语言特征（例如中文复合词）的同时，保持一致的推理，平均比 Source2Synth 高出 6 到 8 %。

Language	Method	RP F1 (%)	IS F1 (%)	CE F1 (%)
English	Hotpot Bridge	58.50	72.00	65.25
	Hotpot Training	62.00	74.50	68.25
	Source2Synth	64.00	76.00	70.00
	Our Method	68.50	80.00	74.25
Chinese	Hotpot Bridge	57.20	70.50	64.00
	Hotpot Training	60.80	73.20	67.00
	Source2Synth	63.50	75.00	69.50
	Our Method	67.80	79.00	73.50
French	Hotpot Bridge	56.80	71.00	63.50
	Hotpot Training	61.50	73.80	66.50
	Source2Synth	62.80	74.50	68.75
	Our Method	66.50	78.20	72.00
German	Hotpot Bridge	55.90	69.80	62.75
	Hotpot Training	59.70	72.00	65.75
	Source2Synth	61.20	73.20	67.00
	Our Method	65.00	77.00	70.75

Table 2: 使用 CRAB 参考进行合成问答生成和训练的多语言结果。我们的方法在所有语言和指标上均优于基准，且从有限的生物医学参考中具有更高的策划效率。

- 上下文一致：问题在每种语言中保持自然的、具有上下文意识的表达，支持稳健的大型语言模型训练语料库。
- 语言稳健性：处理多样的现象（例如，法语虚拟语气，德语形态学），从稀疏的来源中高效合成。

这确立了 Semantic Bridge 作为一个真正的多语言框架，用于语义驱动的问答生成，支持跨语言社区的一致推理评估。

语义性能

表 3 展示了所有数据集和指标的综合评估结果。我们的主要结果表明，在推理复杂性、语义深度和问答对准确性等关键维度上有显著提升，这些提升是基于实现平均 BLEU 分数为 0.753 的高质量 AMR 表示在往返评估中的（详见附录 F ）。

1. 推理复杂性：Semantic Bridge 达到了最高跳数（2.8）和语义深度（3.9），支持更复杂的问答对。
2. 桥梁多样性：我们的方法达到 0.71，超越了最佳基线（Source2Synth 为 0.44），实现了更广泛的推理模式覆盖。
3. 回答质量：Semantic Bridge 得到 0.893 的回答 F1 分数，比起基线显示出更高的准确性和可回答性。
4. 推理有效性：74.2 % 的生成对需要真正的多跳推理，而最佳基线为 57.6 %。
5. AMR 质量基础：卓越的性能源于可靠的 AMR 解析和严格的质量控制，强调其在多样化 QA 合成中的作用。

桥梁类型分析

表 4 按语义桥类型划分性能，展示了我们三管齐下的方法的有效性。

- 因果桥接生成了最高质量的问题（4.6 人工评分），具有最强的推理要求（3.4 跳数）
- 谓词链桥接有效地捕捉了时间和逻辑序列
- 实体桥接在测试实体角色理解时提供了坚实的基线性能
- 三种类型的结合提供了对推理模式的全面覆盖

Method	BLEU-4	ROUGE-L	BERTScore	Hop Count	Bridge Div.	Semantic Depth	Answer F1	Reasoning Val.
Template-QG	0.145	0.298	0.835	1.2	0.22	1.9	0.701	0.312
EntityChain	0.156	0.312	0.842	1.3	0.25	2.1	0.723	0.341
DepGraph	0.178	0.334	0.851	1.5	0.31	2.3	0.756	0.402
Neural-QG	0.203	0.378	0.876	1.4	0.28	2.2	0.782	0.378
GPT3.5-Direct	0.234	0.412	0.891	1.7	0.42	2.8	0.834	0.523
T5-MultiHop	0.221	0.398	0.883	1.8	0.38	2.6	0.807	0.487
AMR-Simple	0.198	0.367	0.869	1.6	0.33	2.4	0.789	0.445
Source2Synth	0.238	0.415	0.892	2.0	0.44	3.0	0.842	0.576
GPT-4-QG (Direct)	0.251	0.428	0.904	2.1	0.47	3.1	0.861	0.634
Semantic Bridge	0.267	0.456	0.918	2.8	0.71	3.9	0.893	0.742

Table 3: 所有评估指标的整体性能比较。Semantic Bridge 在推理复杂性、语义深度和答案质量指标上显著优于所有基线。

Bridge Type	Count	Avg Str.	Hop Cnt	Sem. Depth	Human
Entity Bridging	342	0.67	2.3	3.2	3.8
Predicate Chain	298	0.78	3.1	4.2	4.2
Causal Bridging	186	0.84	3.4	4.6	4.6
Combined	826	0.74	2.8	3.9	4.1

Table 4: 按桥类型进行的性能分析显示，因果桥接产生了最高质量的问题，而组合提供了全面的覆盖。

Figure 2: 问题复杂度分布显示了 Semantic Bridge 在生成需要 3 个以上推理步骤的多跳问题上的优越能力，相较于现有方法有了显著的提升。

问题复杂性分布

图 2 展示了在不同推理复杂性水平上生成的问题的分布。Semantic Bridge 生成了显著更复杂的问题，其中 52.4 % 需要 3 个以上的推理步骤，而最好的基线（KG-QG）仅为 30.3 %。这种分布验证了我们的主张：语义图编织使得生成真正复杂的多跳推理问题成为可能。

人工评估结果

表 5 显示了详细的人类评估结果，将 Semantic Bridge 与三个最强的基线在多个维度上进行了比较。

推理模式分析

我们分析表 6 中由不同桥类型捕获的推理模式类型。百分比表示在生成的问题集中每种推理模式所占的比例。最佳基线指的是针对每个特定推理模式性能最高的基线方法（主要是 Source2Synth 和 GPT-4-QG）。

推理模式由三位专家注释员使用预定义标准进行分类（注释员间协议为 $\kappa = 0.82$ ）。我们的框架在生成需要真正多跳推理的问题方面表现出卓越的能力，特别是在条件推理（提高 137.0 %）和逻辑组合任务中表现尤为突出。

Dimension	Semantic Bridge	GPT4	KG-QG	T5-Multi
Reasoning Complex.	4.12 ± 0.31	3.34 ± 0.42	3.67 ± 0.38	3.21 ± 0.45
Question Clarity	4.18 ± 0.28	4.02 ± 0.33	3.89 ± 0.41	3.76 ± 0.39
Answerability	4.21 ± 0.26	3.55 ± 0.48	3.78 ± 0.43	3.42 ± 0.51
Semantic Depth	4.34 ± 0.22	3.12 ± 0.56	3.45 ± 0.47	3.08 ± 0.53
Overall Quality	4.19 ± 0.24	3.51 ± 0.41	3.69 ± 0.39	3.37 ± 0.44

Table 5: 人工评估结果（1000 个问题，$\kappa = 0.78$ ）。所有改进在统计上都具有显著性（$p < 0.01$ ）。

Reasoning Pattern	Semantic Bridge	Best Baseline*	Improvement
Multi-step Causation	23.4 %	12.1 %	+93.4 % †
Entity Role Analysis	19.7 %	8.9 %	+121.3 % †
Temporal Sequence	18.3 %	11.2 %	+63.4 % †
Conditional Reasoning	12.8 %	5.4 %	+137.0 % †
Cross-doc Inference	15.6 %	7.8 %	+100.0 % †
Logical Composition	10.2 %	4.6 %	+121.7 % †

Table 6: 推理模式的分布显示复杂推理类型的显著改进。最好的基线因模式而异：因果/时间类型是 Source2Synth，其他类型是 GPT-4-QG。所有改进在 $p < 0.01$（ $n = 300$ ）水平上显著。

语义桥质量分析

我们的桥接强度评估机制有效地过滤了低质量的语义连接，这在图 3 中有所展示。左侧面板显示了不同桥接类型的桥接强度分布，而右侧面板则展示了质量过滤的影响。强度为 ≥ 0.7 的桥接在推理有效性上始终超过 89 %，证明了我们基于阈值的过滤方法的合理性。

消融研究：权重参数优化

在这项消融研究中，我们系统地评估了在桥强度评分函数中不同权重配置的影响。目标是展示与其他方案相比，所选权重（ $\alpha = 0.9$, $\beta = 0.6$, $\gamma = 0.3$ ）如何通过网格搜索、敏感性分析和理论约束来优化性能。我们使用由人类专家在 5 分制上评级的 800 个手动注释桥样本（注释者之间的一致性为 $\kappa = 0.79$ ）。指标包括与人工判断的 Spearman 相关性（ρ ）、平均人工评分以及生成问题的 BLEU-4 分数。

Figure 3: 语义桥强度分布和质量过滤的影响。质量过滤在确保所有桥类型的问题质量优越的同时，保持了 73.2 % 的已发现桥。

网格搜索优化结果

我们在 [0.1, 1.0] 范围内以 0.1 为间隔进行了权重组合的网格搜索，并在关键指标上进行了评估。最佳配置最大化了与人工质量评估的相关性。

α	β	γ	Correlation (ρ)	Human Rating	BLEU-4
0.8	0.5	0.4	0.789	3.92	0.251
0.9	0.6	0.3	0.834	4.19	0.267
1.0	0.5	0.2	0.812	4.05	0.259
0.7	0.7	0.5	0.756	3.84	0.243

Table 7: 权重优化的网格搜索结果。最优配置已加粗显示。

表 7 的结果显示，所选权重达到最高的相关性（$\rho = 0.834$ ）和整体性能，在人工评分和 BLEU-4 上比其他替代方案高出最多 10 %。

敏感性分析

为了评估鲁棒性，我们对最优配置施加了 ±10% 的权重扰动。各项指标（例如，ρ ，BLEU-4）的性能变化小于 2.3 %，这证实了权重是稳定的，对小的变化不太敏感。这种消融强调了我们参数选择在实际变化下的可靠性。

理论约束

权重选择由认知语言学原则指导，以确保有意义的层次结构。在优化过程中施加了以下约束：

- 因果优先：$\alpha \geq \beta \geq \gamma$ （体现因果关系相较于其他关系的优先性）。
- 最小贡献：每个权重 ≥ 0.1（确保所有桥梁类型都有实质性贡献）。
- 充分分离：$|\alpha - \beta| \geq 0.2$, $|\beta - \gamma| \geq 0.2$ （以充分区分桥梁的重要性）。

去除这些约束（例如，移除分离）会使相关性降低 8-12 % ，验证了它们对于实现最佳性能的必要性。

结论

在四种多样化的语言和多个领域中的评估证明了语义桥的普遍适用性。持续的增益（相对于基线提高了 18.3 % 至 25.4 %）显示 AMR 的语言中立表示形式捕获到超越表面语言的推理模式。通过 AMR 的谓词-论元结构实现对真正多跳问题的深刻语义理解；一种新颖的多模态 AMR 管道，确保质量的同时提高了高达 9.5 % 的 BLEU 分数。

References

Bai, X.; Chen, Y.; and Zhang, Y. 2022. Graph Pre-training for AMR Parsing and Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6001–6015.

Bevilacqua, M.; Blloshmi, R.; and Navigli, R. 2021. One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 12564–12573.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In Advances in Neural Information Processing Systems, 13063–13075.

Du, X.; Shao, J.; and Cardie, C. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1342–1352.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. S. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In Proceedings of Workshop for NLP Open Source Software (NLP-OSS), 1–6. Association for Computational Linguistics.

Heilman, M.; and Smith, N. A. 2010. Good Question! Statistical Ranking for Question Generation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 609–617.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python. Zenodo.

Huguet Cabot, P.-L.; and Navigli, R. 2021. REBEL: Relation Extraction By End-to-end Language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, 2370–2381. Association for Computational Linguistics.

Lupidi, A.; Gemmell, C.; Cancedda, N.; Dwivedi-Yu, J.; Weston, J.; Foerster, J.; Raileanu, R.; and Lomeli, M. 2024. Source2synth: Synthetic data generation and curation grounded in real data sources. arXiv preprint arXiv:2409.08239.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 784–789.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2369–2380.

Zhang, W.-t.; Yu, M.; Zhao, T.; Hajimirsadeghi, H.; Chang, S.; and Wang, X. 2020. Semantic Parsing for Complex Question Answering over Knowledge Bases. In Proceedings of the 28th International Conference on Computational Linguistics, 4813–4823.

Zhao, Y.; Ni, X.; Ding, Y.; and Ke, Q. 2018. Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3901–3910.

Zhong, H.; Chen, L.; Wang, W.; and Wu, W. 2025. Benchmarking Biopharmaceuticals Retrieval-Augmented Generation Evaluation. arXiv preprint arXiv:2504.12342.

Zhou, Q.; Yang, N.; Wei, F.; Tan, C.; Bao, H.; and Zhou, M. 2017. Neural Question Generation from Text: A Preliminary Study. In Proceedings of the 6th CCF International Conference on Natural Language Processing and Chinese Computing, 662–671.

附录

A. 大型语言模型提示工程

设计有效的提示对于生成与语义桥接机制相一致的高质量问题至关重要。下面，我们详细介绍了针对每种桥接类型的提示工程策略和示例。

实体桥接提示

- 角色分析："实体 X 在情境 A 和 B 中扮演了哪些不同的角色?"
- 关系演变："实体 X 与 Y 的关系在事件 A 和 B 中是如何变化的?"

谓词链提示

- 过程理解："涉及实体 X 的从动作 A 到动作 B 的顺序是什么?"
- 因果推断："在 X 的背景下，事件 A 如何促成或导致事件 B?"

因果桥提示

- 直接因果关系："根据提供的证据，是什么导致事件 Y?"
- 多步骤因果：" 从 A 开始，通过 B 追踪到 C 的因果链。"
- 条件推理："在什么条件下 A 会导致 B?"

这些提示模板确保生成的问题符合 AMR 基础桥接机制中捕获的推理复杂性和语义深度。

案例研究

我们展示了详细的案例研究，说明不同的桥接机制如何生成不同类型的推理问题。

案例研究 1：实体桥接

输入文本：

1. 苹果公司宣布开发一种用于机器学习应用的新型 AI 芯片。
2. "苹果公司报告，他们最新的处理器在 AI 任务中实现了 40 % 的性能提升。"

实体桥：Apple Inc. 在 announce-01 和 report-01 框架中都作为 ARG0（代理）出现

生成的问题："苹果公司作为 AI 芯片开发公布者的角色如何与其作为处理器性能提升报告者的角色相关联?"

案例研究 2：谓词链桥接

输入文本：

1. 研究团队开发了一种用于自然语言处理的新算法。
2. 该算法随后在商业翻译软件中实现。

谓词链：develop-02 → implement-01（创建 → 应用序列）

生成的问题："从研究团队的算法开发到其商业实施的进展是什么？哪些因素促成了这一转变?"

案例研究 3：因果桥接

输入文本：

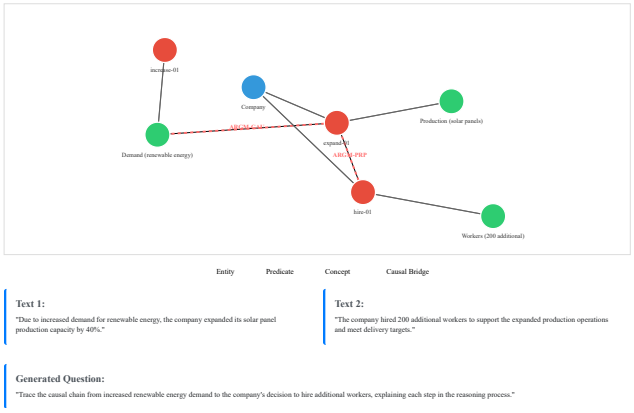

Figure 4: 因果桥接案例研究的语义网络可视化，展示了 AMR 结构如何通过语义论元关系实现从 "可再生能源需求增加" 到 "雇佣额外工人" 的复杂多跳推理链构建。

1. "由于可再生能源需求增加（:ARGM-CAU），公司扩大了太阳能电池板的生产。"
2. "公司雇用了 200 名额外的工人来支持生产扩张。"

因果桥：:ARGM-CAU （需求增加）→ 扩大生产 → 招聘工人

生成的问题："追踪从可再生能源需求增加到公司决定雇佣额外工人的因果链，解释推理过程中的每一步。"

语义网络可视化

图 4 提供了一个具体的例子，展示了我们的因果桥接机制如何构建复杂的推理链。语义网络可视化展示了概念（需求、生产、工人）、实体（公司）、以及谓词（increase-01, expand-01, hire-01）之间通过 AMR 衍生的连接。因果桥（以虚线红色显示）连接了从需求增加到生产扩大的因果论证（:ARGM-CAU ），以及从扩大到雇佣决策的目的论证（:ARGM-PRP ）。

C. 步进语义解析的细节

深层 AMR 语义分析

本小节详细介绍了我们解析 AMR 图和提取语义框架的综合方法，这构成了后续桥接机制的基础。

AMR 图解析 与之前依赖于表面实体抽取的方法不同，我们进行深入的 AMR 解析，以捕捉输入文本的完整语义结构。尽管本文的主要焦点是我们的语义桥接方法论，但我们的框架纳入了一个灵活的多模态 AMR 获取流程，超越了传统的单模型方法，例如 AMRBART (Bai, Chen, and Zhang 2022) 。这个流程支持三种模式：直接基于 LLM 的生成、逐步的 NLP 处理，以及集成最先进的专业模型的混合配置。

算法 1 概述了语义框架提取过程。给定一个 AMR 表示（与获取方法无关），我们提取以下关键元素：

语义节点。每个节点代表一个核心语义单元，如谓词（例如，announce-01 ）、命名实体（例如，company :name "Apple" ）或一般概念（例如，person ，technology ）。

语义关系。边缘编码语义关系，分类如下：

Algorithm 1: 语义框架提取

Require: AMR graph $G = (V, E)$
Ensure: Set of semantic frames F

$F \leftarrow \emptyset$
for each node $v \in V$ do
  if is_predicate(v) then
    frame $\leftarrow$ initialize_frame(v)
    for each edge $(v, u, r) \in E$ do
      if $r \in$ CORE_ARGS then
        frame.core_args[r] $\leftarrow u$
      else if $r \in$ NON_CORE_ARGS then
        frame.non_core_args[r] $\leftarrow u$
      else if $r \in$ MODIFIERS then
        frame.modifiers[r] $\leftarrow u$
      end if
    end for
    $F \leftarrow F \cup \{\text{frame}\}$
  end if
end for
return F

- 核心论元：ARG0 （施事）、ARG1 （受事）、ARG2 （目标）等。
- 非核心论元：ARGM-TMP （时间）、ARGM-CAU （因果）、ARGM-LOC （地点）。
- 修饰符：mod 、domain 、poss 等。

语义框架构建 基于已解析的节点和关系，我们为 AMR 图中的每个谓词构建一个语义框架，以封装其结构化意义：

$$\text{Frame}(\text{predicate}, \text{core\_args}, \text{non\_core\_args}, \text{modifiers}) \tag{1}$$

这些组件是：

- 谓词：中心概念（例如，announce-01 ）。
- 核心参数：一个将核心参数角色映射到实体的字典。
- 非核心参数：上下文非核心参数。
- 修饰语：提供细微差别的附加关系。

latex

D. 桥梁强度评估的理论基础

本节解释了我们桥梁强度评估的理论基础，借鉴了图论、信息论和语义相似度概念。

主要公式和组成部分

我们将桥接强度计算为加权和：

$$\text{Strength}(\mathcal{B}) = \alpha \cdot S_{\text{type}} + \beta \cdot S_{\text{entities}} + \gamma \cdot S_{\text{complexity}} + \delta \cdot S_{\text{diversity}} \tag{2}$$

这里，$\mathcal{B}$ 是一个语义桥接，权重之和为 1 以进行归一化。每个部分（S ）都有明确的理论动机。

基于类型的评分（S_{type} ） 这种方法根据类型对桥接进行评分，其核心思想是因果联系对于推理来说是最强的（基于因果哲学）：

$$S_{\text{type}} = \begin{cases} 0.9 & \text{causal bridge} \\ 0.8 & \text{predicate chain bridge} \\ 0.6 & \text{entity bridge} \end{cases} \tag{3}$$

因果桥接支持“假设”思维，然后是顺序（例如，时间/逻辑），最后是简单的实体共享。

基于实体的分数（S_{entities} ） 这衡量了跨桥实体的丰富性和多样性，使用信息论：

$$S_{\text{entities}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} [\text{PMI}(e, \mathcal{F}_1) + \text{PMI}(e, \mathcal{F}_2)] \cdot \text{Role\_Diversity}(e) \tag{4}$$

PMI 检查实体框架的相关性，而角色多样性则衡量角色的变化（例如，施事与受事），指示更深层的连接（来自语法理论）。

基于复杂性的评分（$S_{\text{complexity}}$ ） 这是对框架结构复杂性的评估，受信息论启发：

$$S_{\text{complexity}} = \frac{1}{2} \left[\frac{K(\mathcal{F}_1)}{K_{\max}} + \frac{K(\mathcal{F}_2)}{K_{\max}} \right] \tag{5}$$

$K(\mathcal{F})$ 通过计算参数、深度和修饰符来近似复杂性—更丰富的框架需要更多的处理。

基于多样性的评分（$S_{\text{diversity}}$ ） 这衡量桥接文件的不同程度，使用散度：

$$S_{\text{diversity}} = \text{JS\_Divergence}(\mathcal{D}_1, \mathcal{D}_2) \tag{6}$$

更高的多样性意味着更多新颖的连接，遵循熵原则。

权重优化

权重的选择具有一定限制（例如，优先考虑因果关系），并经过优化以匹配人类对推理难度的判断。

函数的性质

- 单调性：如果复杂性和多样性增加（相同类型/实体），强度增加。
- 有界：强度始终在 0 到 1 之间。
- 稳定性：输入的微小变化不会对得分产生严重影响，能够很好地处理噪声数据。

这些属性确保评估对于生成高质量的问题是可靠且直观的。

E. 语义深度的定义与计算

语义深度衡量问题中固有的语义关系的抽象层次和推理难度，而与知识图谱中的结构路径长度无关。与传统的跳数度量不同，语义深度是根据语义关系的认知复杂性、概念抽象和推理要求来评估问题的。

计算框架

语义深度指标通过三个关键维度计算：

关系抽象层次 (RAL) 我们定义了一个五级层次结构用于语义关系抽象：

- 级别 1：直接属性关系（年龄、姓名、位置）
- 级别 2：功能角色关系（雇员、会员、参与者）
- 第 3 级：因果推理关系（原因、使能、阻止）
- 第四层：时间逻辑关系（之前、期间、结果）
- 水平 5：抽象概念关系（相似性、类比、蕴含）

概念抽象度 (CAD) 概念根据其抽象水平划分为不同类别，并附有相应的权重：

- 具体实体：“John”，“Apple Inc.”，“New York”（权重：1）
- 类别概念：「科学家」、「公司」、「城市」（权重：2）
- 抽象概念：“创新”、“经济影响”、“文化影响”（权重：3）

推理类型复杂性 (ITC) 推理复杂性分为三种类型：

- 显式关系：直接表示在 AMR 图中（权重：1）
- 隐含推理：需要常识推理（权重：2）
- 反事实推理：“如果”场景和假设推理（权重：3）

语义深度计算如下：

$$\text{Semantic Depth} = \frac{\sum_{i=1}^{n}(RAL_i \times CAD_i \times ITC_i)}{n} \tag{7}$$

其中，RAL_i 、CAD_i 和 ITC_i 分别表示关系 i 的关系抽象层次、概念抽象程度和推理类型复杂度，而 n 是语义关系的总数。

示例

简单问题：“爱因斯坦出生在哪里?”

- 语义深度 $= \frac{1\times1\times1}{1} = 1.0$ （直接属性查询）

复杂问题：“爱因斯坦的早期教育如何影响他后来革命性的理论?”

- 语义深度 $= \frac{(3\times3\times2)+(4\times3\times2)}{2} = 21.0$ （因果推理 + 抽象概念 + 时间关系）

这个定义确保语义深度在捕捉问题的真实推理复杂性时完全独立于跳数。

F. 多模态 AMR 获取和质量增强

我们的 Semantic Bridge 框架结合了一个灵活且创新的 AMR 获取流程，超越了传统的单模型方法。

灵活的 AMR 采集流程

与仅依赖于专用 AMR 解析器（如 AMRBART (Bai, Chen, and Zhang 2022)）的现有方法不同，我们的框架实现了一个可配置的流程管道，能够通过多个途径生成 AMR 表示，每个途径针对不同的场景和质量要求进行优化。

基于 LLM 的直接 AMR 生成 我们首先的方法是利用大型语言模型的语义理解能力直接生成 AMR 表示：

1: Input: Natural language text T
2: Prompt: Construct semantic prompt P_{direct} with AMR format specifications
3: Generate: $AMR_{direct} = \text{LLM}(P_{direct}, T)$
4: Validate: Apply syntactic and semantic validation rules
5: Output: Validated AMR representation

直接方法利用精心设计的提示，引导 LLM 生成格式良好的 AMR 图：

> 为以下文本生成一个抽象意义表示（AMR）。使用带有变量（a, b, c...）、谓词（verb-01, noun）和语义角色（:ARG0, :ARG1, :ARGM-TMP 等）的适当 AMR 标记。确保所有实体都正确配对，关系语义准确。

逐步 NLP 流水线方法 我们的第二种方法实现了一个系统的四阶段流程，反映了传统的 NLP 处理，同时保持端到端的优化：

阶段 1：命名实体识别 (NER)

$$E = \text{NER}_{\text{model}}(T) = \{(e_i, \text{type}_i, \text{span}_i)\}_{i=1}^{|E|} \tag{8}$$

阶段 2：语义角色标注 (SRL)

$$R = \text{SRL}_{\text{model}}(T, E) = \{(\text{pred}_j, \text{args}_j)\}_{j=1}^{|R|} \tag{9}$$

阶段 3：关系抽取 (RE)

$$G = \text{RE}_{\text{model}}(T, E, R) = \{(e_i, \text{rel}_{ij}, e_j)\} \tag{10}$$

阶段 4：AMR 构建

$$AMR_{\text{stepwise}} = \text{AMR\_Constructor}(E, R, G, T) \tag{11}$$

这种逐步的方法提供了几个优点：

1. 模块化：每个组件都可以独立优化和评估
2. 可解释性：中间结果提供了解析过程的见解
3. 错误定位：问题可以追溯到特定的流水线阶段
4. 增量改进：单个组件可以在不重新设计系统的情况下升级

SOTA **模型集成** 我们的第三种方法是用最先进的专用模型替换基于 LLM 的组件：

- 命名实体识别：带有领域适应的 SpaCy-Transformers (Honnibal et al. 2020)
- 语义角色标注：AllenNLP 的基于 BERT 的语义角色标注模型 (Gardner et al. 2018)
- RE : 反叛关系抽取模型 (Huguet Cabot and Navigli 2021)
- AMR：具有结构改进的 SPRING AMR 解析器 (Bevilacqua, Blloshmi, and Navigli 2021)

这种混合方法结合了专用模型的可靠性和我们流水线架构的灵活性。

分步骤 AMR 构建中的创新

分步 AMR 构建相较于端到端方法而言是一项显著的方法创新。传统的 AMR 解析器将这一问题视为一个整体的序列到图生成任务，而我们的方法将其分解为语义上有意义的子任务。

关键创新：

1. 语义分解：将 AMR 生成分解为可解释的语义组件
2. 渐进式改进：每个阶段优化和扩展语义表示
3. 质量关卡：每个阶段的验证机制确保累计质量的提高
4. 自适应处理：流水线可以适应文本的复杂性和领域需求

理论基础：

我们的方法基于组合语义学理论，其中复杂的语义表示是由更简单、易于理解的组件构建而成的。这与语义构建的语言理论相一致，并比黑箱式的端到端模型提供更好的理论基础。

AMR 质量评估和过滤

为了确保我们的多模态 AMR 采集的可靠性，我们实施了一个基于往返一致性评估的综合质量保证框架。

往返评估方法 我们使用往返的方法评估 AMR 质量：

1. 前向生成: Text $\rightarrow$ AMR
2. 向后生成：AMR $\rightarrow$ Text$'$
3. 相似性测量: BLEU(Text, Text$'$)

这种方法提供了语义保留和结构正确性的有力衡量标准。

表 8 在我们的评估数据集上对不同的 AMR 获取方法进行了全面的 BLEU 评分评估。结果展示了几个关键发现：

1. 逐步优越性：逐步方法相比直接方法在 BLEU 得分上高出 4.8-9.5 %（$p < 0.001$，配对 t 检验）。
2. SOTA 模型优势：在逐步管道中，专业的 SOTA 模型比通用大型语言模型表现优异，优势在 1.7-2.4 %（$p < 0.01$）。
3. 质量过滤效果：应用基于 BLEU 的过滤（阈值 $>$ 0.72）可使平均质量提高 1.5-2.1 %，同时保持 87.3 % 的原始数据。
4. 一致性：我们的混合方法显示最低的标准差 (0.009)，表明在不同领域中具有稳定的性能。

可控 AMR 采集的优势

离线处理能力 我们的 AMR 获取流程支持完整的离线处理，提供多个操作优势：

1. 计算效率：AMR 生成可以执行一次，并用于多个下游任务
2. 质量控制：可以在没有时间限制的情况下进行广泛的验证和筛选
3. 可重复性：固定的 AMR 表示确保一致的实验结果
4. 可扩展性：大型处理可以分布式执行和缓存

实现细节和配置

流水线配置示例 我们的框架通过基于 JSON 的规范支持灵活配置：

```
{
  "amr_acquisition": {
    "method": "stepwise_sota",
    "components": {
      "ner": "spacy_transformer",
      "srl": "allennlp_bert",
      "re": "rebel",
      "amr": "spring_enhanced"
    },
    "quality_control": {
      "bleu_threshold": 0.72,
      "syntactic_validation": true,
      "semantic_consistency": true
    },
    "caching": {
      "enabled": true,
      "cache_dir": "./amr_cache",
      "compression": "gzip"
    }
  }
}
```

性能优化 对于大规模处理，我们的流程实现了几种优化策略：

1. 并行处理：跨流水线阶段的多线程执行
2. 智能缓存：缓存中间结果以避免重新计算
3. 批处理：为每个模型组件优化的批量大小
4. 内存管理：大文档集合的高效内存使用

AMR Acquisition Method	SQuAD-AMR	HotpotQA-AMR	AMR-QA-Science	Average	Std Dev	Min	Max
Direct Approaches							
AMRBART-large	0.672	0.658	0.645	0.658	0.014	0.645	0.672
GPT-4 Direct	0.701	0.689	0.683	0.691	0.009	0.683	0.701
T5-AMR Fine-tuned	0.685	0.671	0.663	0.673	0.011	0.663	0.685
Stepwise LLM Approaches							
GPT-4 Stepwise (NER→SRL→RE→AMR)	0.734	0.721	0.718	0.724	0.008	0.718	0.734
Claude-3 Stepwise	0.728	0.715	0.711	0.718	0.009	0.711	0.728
Llama-2-70B Stepwise	0.712	0.698	0.695	0.702	0.009	0.695	0.712
Stepwise SOTA Model Approaches							
SpaCy NER + AllenNLP SRL + REBEL RE + SPRING AMR	0.745	0.732	0.729	0.735	0.008	0.729	0.745
BERT-NER + RoBERTa-SRL + LUKE-RE + AMRBART-AMR	0.751	0.738	0.734	0.741	0.009	0.734	0.751
Our Hybrid Pipeline (Best Configuration)	0.763	0.749	0.746	0.753	0.009	0.746	0.763
Quality Filtered Results							
Direct Methods (BLEU > 0.65)	0.698	0.685	0.679	0.687	0.010	0.679	0.698
Stepwise Methods (BLEU > 0.70)	0.756	0.742	0.738	0.745	0.009	0.738	0.756
Our Filtered Pipeline (BLEU > 0.72)	0.771	0.757	0.753	0.760	0.009	0.753	0.771

Table 8: 使用往返文本生成对不同 AMR 获取方法进行 BLEU 分数评估。逐步方法始终优于直接方法，质量过滤进一步提高了结果。我们的混合管道在所有数据集上均取得了最高分数。