理解问答生成：利用 CQA 从低资源孟加拉语中提取参数化知识和上下文知识

Umme Abira Azmary、MD Ikramul Kayes、Swakkhar Shatabda、Farig Yousuf SadequeBRAC大学计算机科学与工程系

arXiv:2602.01451v1 [cs.CL] 1 Feb 2026

摘要

孟加拉语等低资源语言的问答（QA）模型面临着标注数据有限、语言复杂度高等挑战。一个关键问题是，在生成答案时，模型更多依赖预先编码的（参数化）知识，还是上下文输入；现有孟加拉语问答数据集缺乏开展此类分析所需的结构。我们通过扩展一个孟加拉语数据集，并加入反事实段落和可回答性标注，构建了首个孟加拉语反事实问答数据集 BanglaCQA。此外，我们提出了针对特定语言的编码器-解码器模型及多语言基线模型的微调流程，以及针对仅解码器大语言模型的提示流程，以便在事实和反事实场景中区分参数化知识与上下文知识。此外，我们采用基于大语言模型和人工评估的方法，根据语义相似度衡量答案质量。我们还详细分析了模型在低资源语言的不同问答设置下的表现，并表明思维链（CoT）提示能够有效地从反事实场景中提取参数化知识，在仅解码器大语言模型中尤为明显。我们的工作不仅提出了一种分析孟加拉语问答中知识来源的新框架，还揭示了重要发现，为低资源语言环境下的反事实推理开辟了更广阔的研究方向。

1 引言与相关工作

问答（QA）是自然语言处理领域的基础方向，旨在训练能够模仿人类理解和生成回答、模拟人类推理的模型。随着基于Transformer的模型问世，这种模拟在高资源语言中达到了新的高度，尤其是对于大语言模型（LLMs）而言，

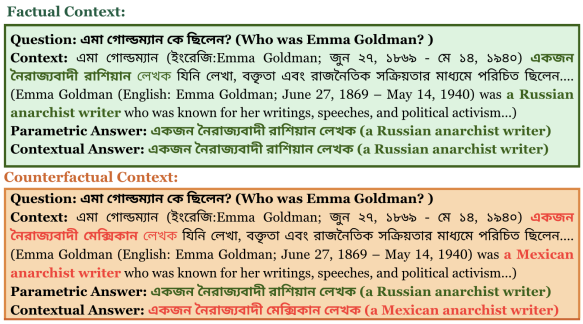

图1：事实性和反事实场景下的参数化与上下文问答（QA）

由于这些模型仅凭其预编码知识便展现出具有竞争力的性能。然而，在上下文问答场景中生成准确回答仍面临挑战，尤其是在反事实上下文中，这源于两种不同“知识来源”的相互作用：（i）参数化知识，即通过预训练嵌入模型参数中的知识；以及（ii）上下文知识，即在执行时从输入上下文中获得的知识（Neeman等人，2023年）。此前对英语问答模型的研究表明，优先依赖参数化知识可能导致幻觉答案的产生，其原因在于大量预编码数据与有限上下文输入之间的不平衡（Krishna等人，2021年）。另有研究进一步表明，包含错误假设的上下文问题会干扰生成性能（Kim等人，2021年）。尽管一些研究表明，将反事实上下文或随机上下文整合到事实性数据集中，可通过解耦知识来源来提升稳健性（Hwang等人，2023年），但此类方法在孟加拉语这一使用广泛但资源不足的语言中仍基本未被探索。尽管在BanglaRQA（Ekram等人，2022）和Squad-BN（Bhattacharjee等人，2022）上评估的模型取得了较高的事实性问答分数，但关键挑战仍未解决：缺乏用于评估

参数化偏差与上下文偏差作为不同因素、对反事实上下文的认识有限，以及追踪知识来源的方法尚不明确。

为解决这些问题，我们通过在现有的BanglaRQA（Ekram 等人，2022）数据集中加入可回答性、随机上下文和反事实上下文，构建了首个孟加拉语反事实问答数据集BanglaCQA，以分析模型对内部知识或上下文知识的优先级。此外，我们提出了知识解耦流程：使用多个编码器—解码器模型（BanglaT5-small（Bhattacharjee 等人，2023年）、BanglaT5-base（Bhattacharjee 等人，2023年）、mt5（Xue 等人，2020））进行微调，并使用仅解码器开源大语言模型（LLaMA-3.3-72B（Touvron 等人，2023年）、DeepSeek-R1-Distill-Qwen-32B（DeepSeek-AI 等人，2025）、Qwen2.5-32B（Yang 等人，2024）、Mistral-3-small（Mistral AI，2025））进行少样本（Brown 等人，2020）和思维链（CoT）（Wei等人，2022）提示，从而区分参数化推理与上下文推理。为评估结果，我们使用 Gemini-2.0-Flash（Hassabis 等人，2024）和 GPT-4.1（OpenAI 等人，2024）进行语义相似度评分；在评估孟加拉语问答回答的语义准确性方面，该方法优于传统指标。此外，我们还对数据集和模型生成的回答进行了人工评估，以确保准确性和透明度。分析结果表明，整合反事实上下文在多个方面表现出色。这些发现不仅为低资源语言和孟加拉语先进问答系统奠定了基础，也凸显了反事实场景中知识利用的透明度。

2 BanglaCQA 数据集

我们介绍了 BanglaCQA，这是首个旨在解耦语言模型中的参数化知识与上下文知识的孟加拉语问答数据集。为此，我们扩展了现有的BanglaRQA（Ekram 等人，2022）数据集，新增了 6.3K 条反事实上下文，增幅达 **42.28%**，专门用于考验模型究竟依赖上下文，还是退回依赖记忆中的信息。[1]

2.1 反事实上下文生成

反事实上下文由事实性示例通过修改关键命名实体得出，使用一套自动化 NER 流程（Sarker，2020）。该脚本识别标准命名实体类型，例如 PER（人名）、LOC（地点）、ORG（组织）、GPE（地缘政治实体）、DATE 或 NUM（时间表达式与数值表达式），并应用类型一致替换。例如，人名会被替换为其他合理的人名，地点会被替换为其他地点，组织会被替换为同一类别的不同实体，同时确保语义连贯性。当命名实体同时出现在上下文和答案字段中时，替换会保持一致。对于时间表达式，如果该实体表示年份，则仅更改最后一位数字，以便在引入细微事实性矛盾的同时保持合理性。在其他数值情况下，则使用正则表达式替换数值。每个修改后的行都会分配一个唯一 ID，以防止重复。这些受控修改构建了假设性矛盾，同时保留了原始句子结构，使我们能够测试模型究竟是根据输入上下文来确定答案，还是默认依赖记忆中的（参数化）知识。

数据集属性	**设置**
问答对总数	21,211
事实性上下文	14,900
反事实上下文	6,303
平均问题词数	8.26
平均上下文词数	215.27

表1：BanglaCQA 数据集的汇总统计。这些统计数据凸显了数据集的规模及其上下文的相对复杂性。

2.2 标注质量保证

使用命名实体识别脚本生成反事实段落后，由论文的两位作者分别对示例进行审阅。此外，为确保客观性，两名独立的付费标注者（未参与反事实数据集的构建）进一步审阅了数据集的语义正确性。意见不一致时通过共识解决，最终获得了**0.73 的科恩 κ 分数**，表明标注者间具有较高的一致性。此外，为减少标签噪声并提升整体质量，还删除了那些尽管缺乏有效上下文答案却被标记为可回答的事实性行。有关标注者角色和要求的更多细节请参见附录A.2。

[1] https://github.com/ikramulkayes/Codes-for-Thesis/

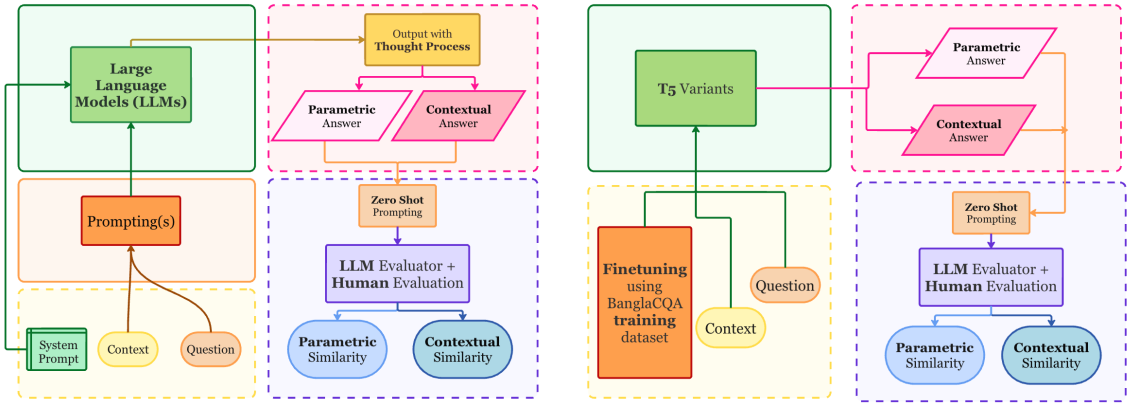

图2：用于区分问答中参数化知识与上下文知识的评估流程。**左侧：**使用大语言模型（LLMs）进行基于提示的推理，以生成参数化答案和上下文答案。**右侧：**使用在 BanglaCQA 上微调的 T5 变体进行基于微调的评估。通过基于大语言模型的自动评估和人工评估这两种方式，对两种知识类型分别进行答案相似度测量。

3 实现流程

为确定最适合 BanglaCQA 的模型架构，我们对 T5（Raffel 等人，2019）框架的多个变体进行了微调，即 BanglaT5（small、base）和 mT5，并采用两种配置：事实性 + 可回答性（F+A）和事实性 +反事实 + 可回答性（F+CF+A）。这种双配置策略能够集中比较语言专用模型和多语言模型在接触事实性与反事实上下文时的适应方式。如图2所示，每个模型均采用一致的流程进行训练，强调可复现性和透明度。数据预处理包括系统化的分词与格式化，随后划分为训练集和验证子集，以确保评估公正。所有实验均采用一组标准化超参数：训练 30 轮、批量大小为 2、学习率为 5e-5，以及 10 个预热步数。基于验证损失采用提前停止，以缓解过拟合；大多数模型在第 15 个训练轮次前后收敛，从而兼顾性能与训练效率。BanglaT5 模型来自 CSEBUET自然语言处理组在 Hugging Face 上的模型仓库，而 mT5 则取自 Google 的官方模型集合，确保所使用的模型基线具有可信度并得到社区认可。对于

模型类型	训练硬件	GPU显存
编码器-解码器	Nvidia RTX 4090 GPU	24GB
仅解码器	4× Nvidia L4 GPU	90GB

表 2：模型使用的训练硬件和 GPU显存。

针对仅解码器大语言模型，我们开发了一个统一的推理框架，使用少样本和思维链提示来探究参数化推理与上下文推理。每个提示将事实性或反事实上下文与指令及问题相结合，其结构设计旨在激发与模型内部知识及外部输入相一致的推理模式。为确保评估的一致性，所有模型均使用相同的超参数进行解码：温度 = 0.1，核采样= 0.1，重复惩罚 = 1.02，以及最大词元数 = 1500。尽管我们部署了 Qwen-2.5（32B）、DeepSeek-R1（32B）、Mistral-3 Small（24B）和 LLaMA-3.3（70B），但由于资源限制，LLaMA-3.3 模型采用 FP16 精度进行了量化。至关重要的是，每个模型都产生了两个独立输出：一个反映内部知识（参数化），另一个源自上下文（上下文）。任何非孟加拉语输出均通过 Gemini API 自动规范化为孟加拉语，从而实现无偏差的跨语言评估。语义对齐在零样本设置下使用 Gemini 2.0 Flash 和 GPT-4.1 进行评估，我们发现它们对孟加拉语比传统指标更为可靠。我们分别报告参数化相似度和上下文相似度，从而细致洞察模型如何在事实性上下文与反事实上下文中进行解释和推理。所有编码器-解码器与仅解码器模型实验的训练硬件及 GPU 显存配置见表 2。

模型	训练于	事实性上下文相似度	F 参数检验相似度	反事实上下文相似度	CF 参数检验相似度
BanglaT5 Small	F+A	0.77	0.70	0.69	0.11
BanglaT5 Base	F+A	0.82	0.81	0.72	0.13
mT5 Small	F+A	0.84	0.79	0.79	0.09
BanglaT5 Small	F+CF+A	0.86	0.84	0.87	0.23
BanglaT5 Base	F+CF+A	0.86	**0.84**	**0.87**	**0.23**
mT5 Small	F+CF+A	**0.87**	0.81	0.84	0.15

表3：使用 **Gemini-2.0 Flash** 作为评估器，通过参数化和上下文相似度评估不同模型在事实性（F）和反事实（CF）设置下的表现。“F”表示**事实性**上下文，“CF”表示**反事实**上下文。粗体值表示各类别中表现最佳的配置。

模型	训练于	事实性上下文相似度	F 参数检验相似度	反事实上下文相似度	CF 参数检验相似度
BanglaT5 Small	F+A	0.79	0.74	0.76	0.16
BanglaT5 Base	F+A	0.83	0.80	0.75	0.14
mT5 Small	F+A	0.84	0.79	0.79	0.13
BanglaT5 Small	F+CF+A	0.85	0.79	0.79	0.21
BanglaT5 Base	F+CF+A	0.87	**0.82**	0.84	**0.27**
mT5 Small	F+CF+A	**0.88**	**0.80**	0.88	0.20

表4：在事实性（F）和反事实（CF）设置下，不同模型的性能；使用参数化相似度和上下文相似度进行评估，并由 **GPT-4.1** 担任评估器。“F”表示**事实性** 上下文，“CF”表示**反事实** 上下文。粗体数值表示各类别中表现最佳的配置。

4 结果

我们通过计算生成输出与目标答案之间的均值语义相似度分数，评估了模型在事实性上下文和反事实上下文中的表现。相似度分数（范围为 0 到 1）使用 Gemini 2.0 Flash 和 GPT-4.1 计算，它们对孟加拉语文本的评估比传统指标更可靠。对于编码器-解码器模型，我们通过比较两种训练配置，观察了使用反事实数据进行微调如何影响性能。仅解码器模型在少样本和思维链提示下进行评估，展现出不同的推理行为，这些行为体现在其参数化输出和上下文输出中。为捕捉这些差异，我们分别分析了反映模型内部知识的参数化响应，以及依赖所提供输入的上下文响应。这种双重评估揭示了不同架构和训练策略在处理事实性查询和反事实查询时如何利用内部与外部信息，从而对模型推理与适应性提供细致洞察。下面我们通过讨论以下研究问题来呈现研究发现：

孟加拉语编码器-解码器模型在反事实上下文中生成参数化答案的表现如何？仅解码器大语言模型又如何缓解这些挑战？ 我们观察到，与事实性上下文相比，在所有受评估的编码器-解码器 T5变体模型中，反事实上下文的参数相似度均值分数显著下降。例如，以 Gemini-2.0-Flash 作为评估器（表 3），BanglaT5 Small 的分数从 0.70（F 参数检验）降至 0.11（CF 参数检验），而 BanglaT5 Base 的分数则从 0.83 降至 0.14；两者均处于（F+A）设置中，清楚地表明模型难以泛化至反事实知识。原因在于，这些模型仅在事实性+ 可回答性设置下进行微调，因此缺乏对反事实场景的理解，导致了这种表现。在事实性和反事实数据上进行微调（F+CF+A）能够提升上下文分数，例如 BanglaT5 Base 的事实性上下文分数升至 0.86，反事实上下文分数升至 0.87；但这并未充分提升 CF 设置下的参数相似度（0.23），这进一步说明，微调有助于上下文理解，却不太能引导模型理解参数化知识。因此，当模型需要依赖内部知识生成参数化答案时，往往会产生幻觉，或将上下文线索与事实混淆。

研究问题1：哪些因素导致了……的不足

模型	提示	事实性上下文相似度	F 参数检验相似度	反事实上下文相似度	CF 参数检验相似度
Qwen-2.5	少样本	0.88	0.35	0.79	0.27
DeepSeek-R1	少样本	0.88	0.32	0.81	0.31
LLAMA-3.3	少样本	0.84	0.27	0.77	0.24
Mistral-3-small	少样本	0.85	0.34	0.79	0.25
Qwen-2.5	COT	0.92	**0.81**	0.86	**0.74**
DeepSeek-R1	COT	**0.94**	0.79	**0.89**	0.70
LLAMA-3.3	COT	0.91	0.69	0.83	0.55
Mistral-3-small	COT	0.90	0.74	0.86	0.64

表5：在事实性（F）和反事实（CF）设置下，不同模型的表现，使用参数化相似度和上下文相似度进行评估，并以**Gemini-2.0 Flash** 作为评估器。所有报告的分数均为均值。“F”表示**事实性** 上下文，“CF”表示**反事实** 上下文。粗体数值表示各类别中表现最佳的配置。

模型	提示	事实性上下文相似度	F 参数检验相似度	反事实上下文相似度	CF 参数检验相似度
Qwen-2.5	少样本	0.89	0.39	0.78	0.31
DeepSeek-R1	少样本	0.83	0.36	0.79	0.30
LLAMA-3.3	少样本	0.86	0.29	0.75	0.27
Mistral-3-small	少样本	0.87	0.37	0.81	0.26
Qwen-2.5	COT	0.93	**0.84**	**0.88**	**0.78**
DeepSeek-R1	COT	**0.95**	0.81	0.91	0.68
LLAMA-3.3	COT	0.90	0.70	0.84	0.59
Mistral-3-small	COT	0.91	0.73	0.85	0.63

表 6：使用 **GPT-4.1** 作为评估器，通过参数化相似度和上下文相似度评估不同模型在事实性（F）和反事实（CF）设置下的表现。所有报告的分数均为均值。“F”表示**事实性** 上下文，“CF”表示**反事实** 上下文。粗体数值表示各类别中表现最佳的配置。

相比之下，仅解码器大语言模型（LLM）利用提示来访问更广泛、更全面的预编码知识库。由于这些模型并非经过微调，而是通过提示来完成任务，因此能够更好地生成准确的参数化答案，尤其是在反事实上下文中。这些结果凸显了孟加拉语编码器-解码器模型的一个根本局限：尽管微调改善了上下文提取能力，但其受限的内部知识妨碍了模型泛化至反事实推理；仅解码器大语言模型凭借其大量的预编码知识，在一定程度上弥补了这一差距。

RQ2：为什么提示策略（思维链与少样本）会影响语言模型在孟加拉语中事实性与反事实设置下的参数化和上下文表现？ 表 5 和 6 中的结果表明，思维链（CoT）提示能带来统计显著且实际效果显著的提升，改善参数化相似度，无论是在事实性**(+0.42–>0.44)**还是反事实**(+0.38–>0.39)**设置下。配对t检验证实了这些提升（$p < 0.01$），且效应量极大（Cohen 的 $d > 5$），这表明这些改进并非偶然，而是具有实际意义（见表 7）。少样本提示本身缺少中间推理阶段：模型直接预测答案，而不会明确地对问题进行推理。因此，在反事实设置下，少样本模型无法验证上下文是否合理，并默认给出依据修改后段落得出的答案，导致参数相似度较低。相比之下，思维链提示会明确要求模型先生成详细的推理链，再给出最终答案（Wei 等人，2022）。这一结构化推理步骤使模型能够区分来自反事实上下文的信息与其编码的参数化知识。

这些发现与近期的理论研究一致

模型	指标	均值Δ(思维链 – 少样本)	t 值	p 值	Cohen's d
Gemini-2.0	F 参数检验	+0.44	26.48	0.00012	13.24
Gemini-2.0	CF 参数检验	+0.39	11.94	0.00126	5.97
GPT-4.1	F 参数检验	+0.42	19.55	0.00029	9.77
GPT-4.1	CF 参数检验	+0.38	12.33	0.00115	6.16

表7：在事实性（F）和反事实（CF）上下文中，使用 Gemini-2.0 和 GPT-4.1 作为评估器，对仅解码器大语言模型进行参数相似度评估。所有分数均反映思维链（COT）与少样本提示之间的均值差异。正向 Δ 值表示思维链提示下性能有所提升。统计显著性通过 t 检验和效应量（d）表示。

这些研究表明，不含中间推理步骤的变换器仅限于低复杂度函数类（例如，AC^0/TC^0 A.1），除非其深度或规模以超多项式速度增长，否则无法解决本质上具有序列性的问题（Peng 等人，2024）。通过生成中间推理步骤，思维链（CoT）有效提升了模型的计算深度，使其能够模拟更大的电路，并解决算术求值和动态规划等任务；这些任务对于有界深度变换器而言，原本是无法表达的。近期研究还发现，思维链的益处不仅来自正确的中间推理，也源于结构归纳偏置：即使推理不完善，只要步骤在结构上相关且顺序正确，模型仍可获得最高达思维链增益 90% 的效果（Jin 等人，2024）。此外，思维链还提供了一种潜在状态跟踪机制，其中每个推理步骤都编码了一项中间计算，供后续步骤引用（Xu 等人，2025）。这些理论洞见解释了我们的结果中观察到的显著提升。孟加拉语问答需要在形态丰富的长上下文中进行推理（平均长度为= 215 个词元；见表 1），并处理反事实实体替换。少样本提示无法引导模型进行结构化推理，导致参数相似度较低。思维链强制采用通用推理模板，弥补了预训练期间缺少孟加拉语特定推理监督所造成的差距。仅解码器模型（例如 Qwen-2.5、DeepSeek-R1）尤其受益，因为它们在训练中接触过类似思维链的推理格式。因此，思维链在事实性和反事实场景下均提升了参数相似度，证实了这些提升具有统计显著性，并且在理论上源于思维链增强型变换器表达能力的扩展及其状态跟踪能力。

研究问题3：语言模型之间的架构差异如何影响它们在孟加拉语的事实性和反事实任务中整合上下文知识与参数化知识的能力？Qwen-2.5 在这两个维度上都取得了较高的相似度分数**(F 参数检验：0.81，CF 参数检验：0.74；F 上下文：0.92，CF 上下文：0.86)**。这可能得益于其针对长序列处理的设计，这与孟加拉语复杂且碎片化的分词特点十分契合。DeepSeek-R1 也展现出类似的性能提升。然而，尽管 LLAMA-3.3 的事实性相似度分数较高（0.91），其 CF 上下文相似度却大幅下降（0.55）。这些发现表明，针对较长上下文进行优化的架构更适合孟加拉语的语言结构。提示词详情见附录 A.3。

4.1 通过人工评估进行错误分析

尽管Gemini-2.0-Flash和ChatGPT-4.1能够对参数化答案相似度提供可扩展且高效的近似，但它们在孟加拉语反事实问答中表现出明显的局限性。为评估指标可靠性并分析由数据集或评估指标引入的潜在错误来源，我们采用了人工评估。两名未参与数据集创建过程的独立标注人员负责评估随机抽取的200个模型生成答案。将这些被广泛视为问答金标准的人工判断（Clark等人，2021年）与模型输出进行比较后，发现了一些差异：

I）时间不匹配（过时目标）：图 3 展示了一个反事实上下文，其中数值是在数据集生成流水线中使用 Python 脚本和正则表达式自动修改的。然而，在这个例子中，数据集的目标参数化答案“550 万”在事实层面已过时或不正确。尽管给定的是反事实输入，模型

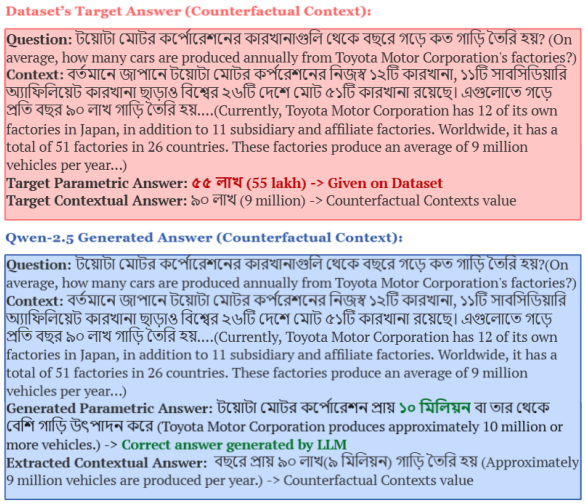

图3：时间不匹配示例：模型生成的答案因比参考答案更新而被扣分

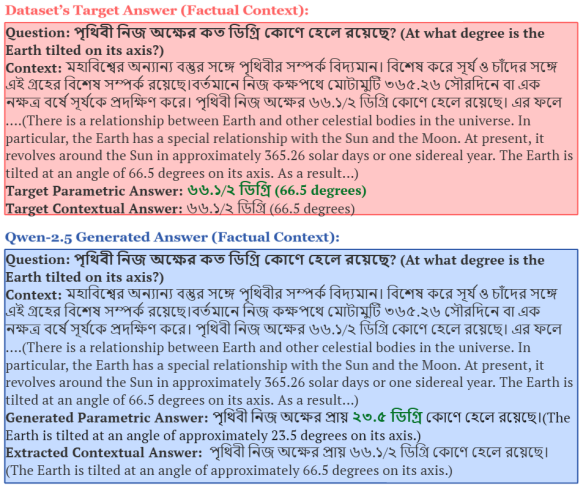

图4：示例展示了因模型解读与参考依据的差异而产生的多个有效答案。

（Qwen-2.5）仍成功生成了正确的参数化答案：约 1000 万或更多。由于数据集依赖固定的参数化目标，这一正确回答在自动评估中遭到了不公正的惩罚。研究发现，约 4% 的生成结果在事实上优于数据集中的参考答案，尤其是在涉及人口数据或政治术语等对时间敏感的问题中。虽然这一比例在其他子集中可能有所不同，但研究结果凸显了一个关键局限：静态参数化参考答案可能无法奖励模型的准确表现，尤其是在大语言模型利用最新参数化知识时。

II）答案差异（多个有效答案）：图4 展示了一个案例：模型预测为23.5°，而数据集目标为66.5°。这两个数值在科学上都是正确的，因为它们表示地球轴向倾角的互余角。

然而，由于评估器模型会将每个答案与数据集答案中的单一参考答案进行比较，因此在缺少词汇或数值重叠时，往往会给出较低的分数。重要的是，在这类情况下，模型生成的答案与数据集中的参考答案之间并不存在本质上的“正确”或“错误”之分；两者都是有效的解答。基于大语言模型的评估器严重依赖文本相似度，除非明确提示其考虑语义等价。随机选取的 200 个输入中，约有 7% 出现了此类情况：由于评估依赖单一的静态参考答案，多个有效答案遭到了不公平的惩罚。这一比例可能会因不同数据片段而异，具体取决于有效解答的多样性。

5 结论

我们提出了BanglaCQA，这是首个面向孟加拉语的反事实问答数据集，旨在区分大语言模型中的参数化知识与上下文知识。通过用受控的反事实上下文扩展BanglaRQA数据集，我们构建了一个基准测试，能够细粒度评估模型对预编码知识与上下文信息的依赖程度。我们对编码器-解码器模型和仅解码器大语言模型的实验表明，思维链提示显著提高了参数化相似度，在事实场景和反事实场景中均是如此，其中Qwen-2.5取得了最佳的整体表现。这些发现凸显了提示策略对于在低资源环境中增强参数化推理的重要性。BanglaCQA为未来针对资源匮乏语言的稳健问答系统研究奠定了基础，并推动开发多参考、时间自适应的评估框架，以更好地反映现实世界的知识动态。

局限性

尽管我们的工作贡献了一个新颖的数据集和评估框架，但仍存在若干局限性。首先，评估时每个实例仅使用一个参考答案，这可能会惩罚语义正确但词汇表达不同的输出。未来的工作应探讨多参考答案评估或人工参与的

循环评分，以更好地涵盖有效答案的不同表达形式。其次，我们的数据集包含人口或政治术语等具有时效性的实体，但参考答案是静态的。模型生成最新信息时仍可能受到不公平的惩罚，这凸显了对时间自适应参考答案的需求。第三，为提高资源效率，我们使用量化权重开展了仅解码器大语言模型的实验；若采用全精度推理，结果可能有所不同。最后，我们的分析侧重于少样本提示和思维链提示，但进一步探索其他提示策略和微调推理模板，或许能提升参数化推理能力。

伦理声明

本研究遵循了数据集创建、标注和评估方面的伦理准则。数据集的初始版本通过基于命名实体识别（NER）的替换方式生成。具体而言，将标注为人物、地点和组织的实体替换为同一类别中其他虚构但语义恰当的名称，以构建反事实上下文。在这一自动化流程之后，一位作者手动审查了所有修改过的条目。如果某个问答对因替换而产生语义问题或不合情理的含义，该作者便会修改或删除该示例，以维护语境完整性。随后，第二位作者独立审查了数据集，并对初步修订提出反馈。经过双方讨论和谨慎的反复完善，最终数据集得以整理完成，以满足反事实问答（CQA）的高标准。

在评估过程中，两名未参与数据集创建的孟加拉语母语标注人员，审查了具有代表性的模型输出子集，以评估其语义正确性。标注人员获得了合理报酬（26 美元），整个研究过程中未使用任何个人或敏感信息。数据集不包含任何个人身份信息，所有实体替换均为合成内容。大语言模型（Gemini-2.0-Flash 和 ChatGPT-4.1）仅用于评估，并通过人工核验输出，确保其正确性和安全性。我们的工作遵循 ACL 关于负责任的数据集构建、人工标注和人工智能系统部署的伦理标准。

参考文献

Abhik Bhattacharjee、Tahmid Hasan、Wasi Ahmad、Kazi Samin Mubasshir、Md Saiful Islam、Anindya Iqbal、M. Sohel Rahman 和 Rifat Shahriyar。2022。BanglaBERT：面向孟加拉语低资源语言理解评估的语言模型预训练与基准。载于计算语言学协会年会北美分会（*NAACL 2022*）论文集，第 1318–1327 页，美国西雅图。计算语言学协会。

Abhik Bhattacharjee、Tahmid Hasan、Wasi Uddin Ahmad 和 Rifat Shahriyar。2023年。BanglaNLG 和 BanglaT5：用于评估孟加拉语低资源自然语言生成的基准和资源。载于计算语言学协会欧洲分会第 *2023* 年年会论文集，第 726–735 页，克罗地亚杜布罗夫尼克。计算语言学协会。

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam Mc- Candlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。预印本，arXiv:2005.14165。

Elizabeth Clark、Tal August、Sofia Serrano、Nikita Haduong、Suchin Gururangan 和 Noah A. Smith。2021年。并非所有“人类”的东西都是黄金：评估人类对生成文本的评估。载于第*59*届计算语言学协会年会论文集（*ACL 2021*），第7282–7296页，线上。计算语言学协会。

DeepSeek-AI、Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi、Xiaokang Zhang、Xingkai Yu、Yu Wu、Z. F. Wu、Zhibin Gou、Zhihong Shao、Zhuoshu Li、Ziyi Gao、Aixin Liu、Bing Xue、Bingxuan Wang、Bochao Wu、Bei Feng、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan、Damai Dai、Deli Chen、Dongjie Ji、Erhang Li、Fangyun Lin、Fucong Dai、Fuli Luo、GuangboHao、Guanting Chen、Guowei Li、H.Zhang、Han Bao、Hanwei Xu、Haocheng Wang、Honghui Ding、Huajian Xin、Huazuo Gao、Hui Qu、Hui Li、Jianzhong Guo、Jiashi Li、Jiawei Wang、Jingchang Chen、Jingyang Yuan、Junjie Qiu、Junlong Li、J. L. Cai、Jiaqi Ni、Jian Liang、Jin Chen、Kai Dong、Kai Hu、Kaige Gao、Kang Guan、Kexin Huang、Kuai Yu、Lean Wang、Lecong Zhang、Liang Zhao、Litong Wang、Liyue Zhang、Lei Xu、Leyi Xia、Mingchuan Zhang、Minghua Zhang、Minghui Tang、Meng Li、Miaojun Wang、Mingming Li、Ning Tian、Panpan Huang、Peng Zhang、Qiancheng Wang、Qinyu Chen、Qiushi Du、Ruiqi Ge、Ruisong Zhang、Ruizhe Pan、Runji Wang、R. J. Chen、R. L. Jin、Ruyi Chen、Shanghao Lu、Shangyan Zhou、Shanhuang Chen、

Shengfeng Ye, ShiyuWang, ShuipingYu,ShunfengZhou, Shuting Pan, S. S. Li, Shuang Zhou,ShaoqingWu, ShengfengYe, TaoYun, TianPei, Tianyu Sun,T. Wang, Wangding Zeng, WanjiaZhao, Wen Liu,WenfengLiang, Wenjun Gao, Wenqin Yu, WentaoZhang, W. L.Xiao, Wei An,Xiaodong Liu, XiaohanWang, XiaokangChen, XiaotaoNie, XinCheng, XinLiu, XinXie, Xingchao Liu, XinyuYang, XinyuanLi,Xuecheng Su, Xuheng Lin, X. Q.Li, Xiangyue Jin,Xiaojin Shen, XiaoshaChen,XiaowenSun,Xiaoxi-ang Wang, Xinnan Song, Xinyi Zhou, XianzuWang,Xinxia Shan, Y. K. Li, Y. Q.Wang, Y. X. Wei, YangZhang, Yanhong Xu, Yao Li, YaoZhao, YaofengSun, YaohuiWang, Yi Yu, Yichao Zhang, Yifan Shi,Yiliang Xiong, Ying He, Yishi Piao, YisongWang,Yixuan Tan, Yiyang Ma, YiyuanLiu, Yongqiang Guo,YuanOu, YuduanWang, YueGong, YuhengZou, Yu-jiaHe, YunfanXiong, YuxiangLuo, Yuxiang You,Yuxuan Liu, Yuyang Zhou, Y. X.Zhu, Yanhong Xu,YanpingHuang, YaohuiLi, YiZheng, YuchenZhu,Yunxian Ma, Ying Tang, YukunZha, Yuting Yan,Z.Z. Ren, ZehuiRen, Zhangli Sha, ZheFu, ZheanXu, Zhenda Xie, ZhengyanZhang, Zhewen Hao,ZhichengMa, ZhigangYan, Zhiyu Wu, Zihui Gu, Zi-jia Zhu, Zijun Liu, ZilinLi, Ziwei Xie, Ziyang Song,ZizhengPan, Zhen Huang, Zhipeng Xu, ZhongyuZhang, and Zhen Zhang.2025. DeepSeek-R1：通过强化学习激励大语言模型的推理能力。 预印本，arXiv:2501.12948。

Syed Mohammed Sartaj Ekram、Adham Arik Rah- man、Md. Sajid Altaf、Mohammed Saidul Islam、Mehrab Mustafy Rahman、Md Mezbaur Rahman、Md Azam Hossain 和 Abu Raihan Mostofa Kamal。2022。BanglaRQA：面向资源匮乏的孟加拉语、基于阅读理解的问答基准数据集，涵盖多种问题和答案类型。载于计算语言学协会研究成果：*EMNLP 2022*，第 2518–2532 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Demis Hassabis、Koray Kavukcuoglu 和 Gemini 团队。2024。Gemini 2.0：面向智能体时代的新型 AI 模型。https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/。Google DeepMind 博客。

Yerin Hwang、Yongi-Mi Kim、Hyunkyung Bae、Jeesoo Bang、Hwanhee Lee 和 Kyomin Jung。2023。Dialogizer：从文本来源生成上下文感知的对话式问答数据集。载于自然语言处理实证方法会议。

Mingyu Jin、Qinkai Yu、Dong Shu、Haiyan Zhao、Wenyue Hua、Yanda Meng、Yongfeng Zhang 和 Mengnan Du。2024年。推理步骤长度对大语言模型的影响。载于计算语言学协会 *ACL 2024* 年会论文集（研究成果），第 1830–1842 页，泰国曼谷。计算语言学协会。

Najoung Kim、Ellie Pavlick、Burcu Karagol Ayan 和 Deepak Ramachandran。2021年。是哪位语言学家发明了灯-泡？面向问答的预设验证。收录于 第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第3932–3945页，线上。计算语言学协会。

Kalpesh Krishna、Aurko Roy 和 Mohit Iyyer。2021年。长篇问答研究面临的障碍。载于*2021* 年北美计算语言学协会分会会议论文集：人类语言技术，第 4940–4957 页，线上。计算语言学协会。

Mistral AI。2025。Mistral Small 3：一款针对延迟优化的 24B 参数模型，采用 Apache 2.0 许可发布。https://mistral.ai/news/mistral-small-3。介绍 Mistral Small 3 的研究博客文章，81

Ella Neeman、Roee Aharoni、Or Honovich、Leshem Choshen、Idan Szpektor 和 Omri Abend。2023年。DisentQA：通过反事实问答区分参数化知识与上下文知识。载于计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），第10056–10070页，多伦多，加拿大。计算语言学协会。

OpenAI、Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat、Red Avila、Igor Babuschkin、Suchir Balaji、Valerie Balcom、Paul Baltescu、Haiming Bao、Mohammad Bavarian、Jeff Belgum、Irwan Bello、Jake Berdine、Gabriel Bernadett-Shapiro、Christopher Berner、Lenny Bogdonoff、Oleg Boiko、Madelaine Boyd、Anna-Luisa Brakman、Greg Brockman、Tim Brooks、Miles Brundage、Kevin Button、Trevor Cai、Rosie Campbell、Andrew Cann、Brittany Carey、Chelsea Carlson、Rory Carmichael、Brooke Chan、Che Chang、Fotis Chantzis、Derek Chen、Sully Chen、Ruby Chen、Jason Chen、Mark Chen、Ben Chess、Chester Cho、Casey Chu、Hyung Won Chung、Dave Cummings、Jeremiah Currier、Yunxing Dai、Cory Decareaux、Thomas Degry、Noah Deutsch、Damien Deville、Arka Dhar、David Dohan、Steve Dowling、Sheila Dunning、Adrien Ecoffet、Atty Eleti、Tyna Eloundou、David Farhi、LiamFedus、Niko Felix、Simón Posada Fishman、Juston Forte、Isabella Fulford、Leo Gao、Elie Georges、Christian Gibson、Vik Goel、Tarun Gogineni、Gabriel Goh、Rapha Gontijo-Lopes、Jonathan Gordon、Morgan Grafstein、Scott Gray、Ryan Greene、Joshua Gross、Shixiang Shane Gu、Yufei Guo、Chris Hallacy、Jesse Han、Jeff Harris、Yuchen He、Mike Heaton、Johannes Heidecke、Chris Hesse、Alan Hickey、Wade Hickey、Peter Hoeschele、Brandon Houghton、Kenny Hsu、Shengli Hu、Xin Hu、Joost Huizinga、Shantanu Jain、Shawn Jain、Joanne Jang、Angela Jiang、Roger Jiang、Haozhun Jin、Denny Jin、Shino Jomoto、Billie Jonn、Heewoo Jun、Tomer Kaftan、Łukasz Kaiser、Ali Kamali、Ingmar Kanitscheider、Nitish Shirish Keskar、Tabarak Khan、Logan Kilpatrick、Jong Wook Kim、

Christina Kim、Yongjik Kim、Jan Hendrik Kirchner、Jamie Kiros、Matt Knight、Daniel Kokotajlo、Łukasz Kondraciuk、Andrew Kondrich、Aris Konstantinidis、Kyle Kosic、Gretchen Krueger、Vishal Kuo、Michael Lampe、Ikai Lan、Teddy Lee、Jan Leike、Jade Leung、Daniel Levy、Chak Ming Li、Rachel Lim、Molly Lin、Stephanie Lin、Mateusz Litwin、Theresa Lopez、Ryan Lowe、Patricia Lue、Anna Makanju、Kim Malfacini、Sam Manning、Todor Markov、Yaniv Markovski、Bianca Martin、Katie Mayer、Andrew Mayne、Bob McGrew、Scott Mayer McKinney、Christine McLeavey、Paul McMillan、Jake McNeil、David Medina、Aalok Mehta、Jacob Menick、Luke Metz、Andrey Mishchenko、Pamela Mishkin、Vinnie Monaco、Evan Morikawa、Daniel Mossing、Tong Mu、Mira Murati、Oleg Murk、David Mély、Ashvin Nair、Reiichiro Nakano、Rajeev Nayak、Arvind Neelakantan、Richard Ngo、Hyeonwoo Noh、Long Ouyang、Cullen O'Keefe、Jakub Pachocki、Alex Paino、Joe Palermo、Ashley Pantuliano、Giambattista Parascandolo、Joel Parish、Emy Parparita、Alex Passos、Mikhail Pavlov、Andrew Peng、Adam Perelman、Filipe de Avila Belbute Peres、Michael Petrov、Henrique Ponde de Oliveira Pinto、Michael Pokorny、Michelle Pokrass、Vitchyr H. Pong、Tolly Powell、Alethea Power、Boris Power、Elizabeth Proehl、Raul Puri、Alec Radford、Jack Rae、Aditya Ramesh、Cameron Raymond、Francis Real、Kendra Rimbach、Carl Ross、Bob Rotsted、Henri Roussez、Nick Ryder、Mario Saltarelli、Ted Sanders、Shibani Santurkar、Girish Sastry、Heather Schmidt、David Schnurr、John Schulman、Daniel Selsam、Kyla Sheppard、Toki Sherbakov、Jessica Shieh、Sarah Shoker、Pranav Shyam、Szymon Sidor、Eric Sigler、Maddie Simens、Jordan Sitkin、Katarina Slama、Ian Sohl、Benjamin Sokolowsky、Yang Song、Natalie Staudacher、Felipe Petroski Such、Natalie Summers、Ilya Sutskever、Jie Tang、Nikolas Tezak、Madeleine B. Thompson、Phil Tillet、Amin Tootoonchian、Elizabeth Tseng、Preston Tuggle、Nick Turley、Jerry Tworek、Juan Felipe Cerón Uribe、AndreaVallone、Arun Vijayvergiya、Chelsea Voss、Carroll Wainwright、Justin Jay Wang、Alvin Wang、Ben Wang、Jonathan Ward、Jason Wei、CJ Weinmann、Akila Welihinda、Peter Welinder、Jiayi Weng、Lilian Weng、Matt Wiethoff、Dave Willner、Clemens Winter、Samuel Wolrich、Hannah Wong、Lauren Workman、Sherwin Wu、Jeff Wu、Michael Wu、Kai Xiao、Tao Xu、Sarah Yoo、Kevin Yu、Qiming Yuan、Wojciech Zaremba、Rowan Zellers、Chong Zhang、Marvin Zhang、Shengjia Zhao、Tianhao Zheng、Juntang Zhuang、William Zhuk和Barret Zoph。2024。GPT-4技术报告。预印本，arXiv:2303.08774。

Binghui Peng、Srini Narayanan 和 Christos Papadimitriou。2024。关于 Transformer 架构的局限性。预印本，arXiv:2402.08164。

Colin Raffel、Noam M. Shazeer、Adam Roberts、Kather- i ne Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2019年。通过统一的文本到文本Transformer探索迁移学习的极限。机器学习研究期刊，21:140:1–140:67。

Sagor Sarker。2020。BanglaBERT：用于孟加拉语理解的孟加拉语掩码语言模型。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、BaptisteRozière、Naman Goyal、Eric Hambro、Faisal Azhar、Aurélien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lample。2023年。Llama：开放且高效的基础语言模型。*arXiv*，abs/2302.13971。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。2022。思维链提示可激发大语言模型的推理能力。载于第*36*届神经信息处理系统国际会议论文集，NIPS '22，美国纽约州雷德胡克。Curran Associates Inc.

Yige Xu、Xu Guo、Zhiwei Zeng 和 Chunyan Miao。2025。SoftCoT：用于大语言模型高效推理的软思维链。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第23336–23351页，奥地利维也纳。计算语言学协会。

Linting Xue、Noah Constant、Adam Roberts、Mihir Kale、Rami Al-Rfou、Aditya Siddhant、Aditya Barua和Colin Raffel。2020。 mt5：大规模多语言预训练文本到文本Transformer。载于计算语言学协会北美分会。

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

A 附录

A.1 AC^0/TC^0 详情

AC^0 (深度为 0 的交替电路)：指一种常数深度、多项式规模的布尔电路类，具有无界扇入的与门、或门和非门。AC^0 电路无法计算某些函数，例如奇偶校验或多数函数。

TC^0 (阈值电路)：与 AC^0, 类似，但加入了多数函数（阈值）门，因此功能更强。这些电路仍具有常数深度

且规模为多项式，能力略强于 AC^0, ，但表达能力仍然有限。

少样本提示缺少中间推理步骤，因此大语言模型的行为类似于 AC^0/TC^0 电路，也就是说，其推理能力有限，无法解决复杂的序列任务（例如多步逻辑或算术问题）。相比之下，思维链提示引入了中间推理步骤，提高了模型的有效计算深度。这使模型能够模拟功能更强大的电路并执行更复杂的推理任务，从而摆脱类似 AC^0/TC^0 电路的局限。

A.2 标注员信息

参数化答案与上下文答案的标注指南

两名独立标注人员参与了数据集的验证。两人都因付出的努力获得了合理报酬。标注人员是来自不同大学的学生，学术背景各异：一人主修 STEM 学科，另一人主修非 STEM 领域。尽管背景不同，两人都积极参与与各自领域相关的研究。值得注意的是，两人都不是本文作者。除外部标注人员外，本文的两位作者也参与了验证过程。为确保所有示例的验证工作保持一致并具备高质量，每位标注人员都收到了以下详细说明。



Dear Volunteers,

I hope this email finds you well. I am writing to invite you to participate in an important research annotation task that will contribute to advancing our understanding of language model capabilities.

Project Overview

We are conducting a study that examines how language models handle different types of knowledge - specifically comparing answers based on general world knowledge (parametric) versus information derived fro

Time Commitment & Compensation

- **Project duration:** July 20th - July 26th, 2025 (7 days)
- **Compensation:** Each volunteer will receive **$26 USD (3,100 Taka)** upon successful completion of the entire dataset annotation task
- **Schedule:** Flexible - work at your own pace within the deadline

Complete Annotation Guidelines

Objective

Each example in the dataset includes:

- A **question**
- A **context paragraph**
- Two types of answers:
  - **Parametric Answer** – A fact-based answer that reflects general world knowledge
  - **Contextual Answer** – An answer derived specifically from the given context

You must independently label the correctness of each answer using one of three categories: *Valid*, *Invalid*, or *Confused*.

Label Definitions

Answer Type	Valid	Invalid / Confused
Parametric Answer	Factually correct and can be verified independently of the passage	Incorrect, contradictory to known facts, or question is unclear
Contextual Answer	Correct based on the given context and consistent with its information	Not supported by the context, or context is ambiguous/irrelevant

Step-by-Step Instructions

1. Carefully read the **context**, **question**, and both answers
2. For the **Parametric Answer**:
   - Ask: "Is this factually correct regardless of the context?"
3. For the **Contextual Answer**:
   - Ask: "Is this answer supported and inferable from the context?"
4. Label each answer independently as:
   - **Valid** – Answer is correct as per the criteria
   - **Invalid** – Answer is incorrect or irrelevant
   - **Confused** – Answer is hard to evaluate due to ambiguity in question or context

Important Guidelines

- Base your judgment only on the content provided (context, question, answers)
- Do not use external knowledge to validate the contextual answer
- If unsure, select *Confused* and move on—do not guess
- Apply labels consistently across all examples
- Maintain attention to detail and consistency throughout
- **Note:** The contextual answer will always be provided for each question, even when the parametric answer might be the same. This ensures consistent evaluation across both answer types.
- Base your judgment only on the content provided (context, question, answers)
- Do not use external knowledge to validate the contextual answer
- If unsure, select *Confused* and move on—do not guess
- Apply labels consistently across all examples
- Maintain attention to detail and consistency throughout

Requirements

- Attention to detail and consistency
- Ability to follow the detailed guidelines provided above
- Basic familiarity with factual information validation
- Commitment to maintaining confidentiality
- Complete the entire dataset within the given timeframe

Why Your Participation Matters

Your annotations will help create a reliable evaluation dataset that researchers can use to better understand and improve language models. This work has implications for developing more accurate and trustworthy

What's Provided

You will receive:

- Access to the annotation platform
- Technical support throughout the process
- Payment upon successful completion

Next Steps

If you're interested in participating, please reply to this email by **July 18th, 2025**. I will then send you:

1. Access credentials to the annotation platform
2. The dataset to be annotated
3. A small practice set to familiarize yourself with the task

图 5：标注员指导邮件的详细内容。

目标。数据集中的每个示例都包含：

- 一个**问题**
- 一段**上下文段落**
- 两种类型的答案：

  – **参数化答案**– 基于事实的答案，反映一般的世界知识。– **上下文答案**– 专门根据给定上下文得出的答案。

标注人员必须独立判断每个答案的正确性，并从以下三类中选择一类：有效、无效或困惑。

说明。

1. 仔细阅读**上下文**、**问题**以及两个答案。
2. 对于**参数化答案**：
   - 提问："无论上下文如何，这在事实上是否正确？"
3. 对于**上下文答案**：
   - 询问："该答案是否得到上下文支持，并且可以从上下文中推断出来？"
4. 将每个答案单独标注为：
   - **有效** – 答案根据标准是正确的。
   - **无效**– 答案不正确或不相关。
   - **困惑** – 由于问题或上下文存在歧义，答案难以评估。

补充说明。

- 仅根据所提供的内容（上下文、问题、答案）进行判断。

- 不要使用外部知识来验证上下文答案。

- 如不确定，请选择困惑 并继续，不要猜测。

- 请在所有示例中一致地应用标签。

伦理提醒。 标注人员应在整个验证过程中保持保密性并遵守伦理标准。你们的认真付出有助于构建可靠且公平的评估数据集。

此外，数据集中的每个示例均由四位审阅者进行标注，并使用科恩 κ 系数衡量标注者间一致性。投票分布和一致性分析结果如下所示。

投票模式分析：

- **全票一致（4-0-0）：** 5960 行（94.6%）
- **强多数（3-1-0）：** 309 行（4.9%）
- **弱多数（2-2-0）：** 13 行（0.2%）
- **混合（2-1-1）：** 21 行（0.3%）
- **其他：** 0 行（0.0%）

数据集摘要：

- 总行数：6,303
- 每行由 4 位审阅者标注
- 每行包含对以下类别之一的投票：有效、无效或困惑
- 每行的投票总数均为 4

指标	达成
科恩 κ 系数	0.7212

表 8：该数据集的标注者间一致性。

A.3 系统提示和用户提示

{"role": "system","content": '"You are tasked with generating both parametric and contextual answers based on a Bengali context.

Contextual Answer:
Derive strictly from the given context. If the context lacks sufficient info, reply: "Context does not provide enough information."

Parametric Answer:
Use pre-trained knowledge only; do not refer to the context. If information is missing, make reasonable assumptions and state them. If not possible, reply: "None."

Key Note:
In the context, a word, year, or number might be incorrect. However, you must extract contextual answers as given in the context, even if it is wrong.
On the contrary, you should answer parametric answers correctly while correcting error of context based on your knowledge.

Thought Process:
Think step by step to ensure clarity.
Explain how the contextual and parametric answers were derived.
After explaining the derivation process, make sure to write "end of thought process" and then provide your response.

Response Format:
Contextual Answer: {Answer based only on the context.}
Parametric Answer: {Answer based on knowledge without referencing the context.}
Reasoning: Explain how both answers were derived step by step.

Example:
Context: "বাংলাদেশের রাজধানী চট্টগ্রাম।"
Question: "বাংলাদেশের রাজধানীর নাম কী?"
Output that you will generate:
Reasoning:

The context explicitly states the capital is Chattogram, so the contextual answer is "চট্টগ্রাম।"
Based on my knowledge, the capital is Dhaka, correcting the error in the context.

End of thought process

Contextual Answer: "চট্টগ্রাম।"
Parametric Answer: "ঢাকা।"
'"}

图 6：系统提示词定义了任务目标、答案类型和响应结构，引导模型区分基于知识与基于上下文的回答，以应用思维链技术

Here is some context:
বাস্কেটবল অত্যন্ত জনপ্রিয় খেলা হিসেবে বিশ্বব্যাপী পরিচিত। গোলাকৃতি, কমলা রঙের বল দিয়ে অভ্যন্তরীণ এবং বহিঃস্থ - উভয় প্রকার মাঠেই খেলা হয়ে থাকে। দলগত ক্রীড়া হিসেবে বাস্কেটবলের মূল উদ্দেশ্য হচ্ছে কোর্টে উলম্বভাবে স্থাপিত একটি বাস্কেট বা ঝুড়িতে বল নিক্ষেপের মাধ্যমে পয়েন্ট সংগ্রহ করা। নির্দিষ্ট আইন-কানুন অনুসরণ করে সর্বাধিক পয়েন্ট সংগ্রহকারী দল খেলায় বিজয়ী ঘোষিত হয়। সাধারণতঃ প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে। চতুর্ভুজ আকৃতির বাস্কেটবল কোর্টের উভয় দিকের শেষ প্রান্তে বাস্কেট ঝুলিয়ে রাখা হয় যা রিম নামে পরিচিত। বাস্কেটবল খেলার প্রধান উপকরণ হিসেবে বলকেও বাস্কেটবল নামে আখ্যায়িত করা হয়।

Question: বাস্কেটবল খেলায় প্রত্যেক দলে কতজন খেলোয়াড় থাকে ?

Thought Process**: Start by explaining your step-by-step reasoning for solving the task.
After explaining the derivation process, write "End of thought process"
After that you will answer,
Contextual Answer: {Provide the answer in Bengali based on the given context only. Do not include any external knowledge. Do not need for your own knowledge base to answer this}
Parametric Answer: {Provide the answer in Bengali based on your pre-trained knowledge only. Do not reference the context.}'"
'"}

图 7：用户提示示例，展示如何向模型提供孟加拉语上下文和问题，以便模型通过思维链技术生成结构化答案

Output:
****Thought Process**:**

So I have this task where I need to provide both a contextual and a parametric answer based on a given Bengali context about basketball. The question is: "বাস্কেটবল খেলায় প্রত্যেক দলে কতজন খেলোয়াড় থাকে?" which translates to "How many players are there in each team in basketball?"

First, I need to understand what a contextual answer means. It means that I have to derive the answer strictly from the given context, without using any external knowledge. If the context doesn't provide enough information, I should say "Context does not provide enough information."

On the other hand, a parametric answer is based on my pre-trained knowledge, without referring to the context. If I don't have the required knowledge, I should make reasonable assumptions and state them, or say "None" if it's not possible.

Looking at the context, it talks about basketball being a popular sport worldwide, played with an orange, spherical ball on both indoor and outdoor courts. It's a team sport where the main objective is to score points
by投掷球进位于球场两端垂直放置的篮筐或称作篮圈。比赛胜利属于按照特定规则积累最多分数的队伍。通常，每队有9名球员。球场是矩形的，两端分别悬挂着篮筐，称为篮板。篮球比赛的主要装备被称为篮球。

Wait, in the context, it says "সাধারণতঃ প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে।" which means "Usually, each team has 9 players." Hmm, that seems off because, in standard basketball, each team has 5 starting players on the court, with substitutions from a larger roster, but the number 9 seems unusual.

However, for the contextual answer, I have to take the information as it is presented in the context, regardless of whether it's correct or not. So, according to the context, each team has 9 players.

Now, for the parametric answer, I need to rely on my pre-trained knowledge. I know that in standard basketball, especially in professional leagues like the NBA, each team has 5 players on the court at a time, with a roster of around 12 to 15 players, depending on the league rules.

But specifically for the number of players on the court, it's 5 per team. So, if someone asks how many players are there in each team in basketball, typically, it's 5 players on the court.

However, considering that the context mentions 9 players, and knowing that in some variations or levels of basketball, the number can vary, but in standard rules, it's 5.

Wait, maybe the context is referring to something else, like the number of players on the roster or something. But it specifically says "প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে," which seems to mean the number of players on the team, not just on the court.

But in standard basketball, the team roster can have more players, but on the court, it's 5 per team.

Perhaps in some amateur or youth leagues, they have different numbers, but in professional and standard rules, it's 5.

So, for the parametric answer, I should provide the correct standard number, which is 5 players per team on the court.

But the context says 9, which might be incorrect or perhaps referring to something else.

Anyway, I need to provide the answers as per the instructions.

End of thought process

Contextual Answer: সাধারণতঃ প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে।

Parametric Answer: প্রত্যেক দলে ৫ জন খেলোয়াড় থাকে কোর্টে।
Explanation:
No answer found.

"}

图 8：QwQ-2.5-32B 的输出，展示了完整的推理过程以及最终答案，这些内容均根据提示格式生成。

System Prompt (Recommended)

You are a helpful assistant that provides Parametric and Contextual answers based on Bengali text. Parametric answers should be the true/correct answers based on your knowledge. Contextual answers should be derived directly from the given context, even if they might differ from the true answer.
Here are examples of the expected format:

Example 1: Context: গণিতে, বিশেষ করে বীজগাণিতিক টপোলজি এবং সমসংস্থ তত্ত্বে, মেয়ার-ভিটোরিস ক্রম বা ম্যায়ের-ভিয়েতরিস ধারা হল একটি বীজগাণিতিক সরঞ্জাম, যা টপোগাণিতিক ক্ষেত্রের সমসংস্থ এবং সহ-সমসংস্থ হিসেবে পরিচিত বীজগাণিতিক ইনভ্যারিয়েন্টসমূহ হিসাব করতে সাহায্য করে। অস্ট্রিয়ার গণিতবিদ ওয়ালথার মেয়ার এবং লেপল্ড ভিয়েটরিস নামক বিজ্ঞানীদ্বয় এটি আবিষ্কার করেন। এই পদ্ধতিতে একটি ক্ষেত্রকে বিভিন্ন উপক্ষেত্রে বিভক্ত করা হয়, যাতে সমসংস্থ এবং সহসমসংস্থ গ্রুপগুলো পরিমাপ করা সহজ হয়। এই ধারাটি মুল ক্ষেত্রটির (সহ)সমসংস্থ গ্রুপগুলোর সঙ্গে উপক্ষেত্রগুলোর (সহি)সমসংস্থ গ্রুপের সম্পর্ক তৈরী করে।
Question: কেন ম্যায়ের-ভিয়েতরিস ধারা পদ্ধতিতে একটি ক্ষেত্রকে বিভিন্ন উপক্ষেত্রে বিভক্ত করা হয়?
Parametric Answer: যাতে সমসংস্থ এবং সহসমসংস্থ গ্রুপগুলো পরিমাপ করা সহজ হয়
Contextual Answer: যাতে সমসংস্থ এবং সহসমসংস্থ গ্রুপগুলো পরিমাপ করা সহজ হয়

Example 2: Context: হামিদুল হক চৌধুরী (জন্ম: ১৯০১-মৃত্যু: ১৮ই জানুয়ারী, ১৯৯২) হলেন একজন বাঙালী মুসলমান রাজনীতিবিদ ও আইনজীবী। ১৯৩৭ সালে মুসলিম লীগের মনোনয়নে বঙ্গীয় ব্যবস্থাপক পরিষদের সদস্য নির্বাচিত হন। বঙ্গীয় মুসলিম লীগের অন্যতম নেতা হিসেবে ঐতিহাসিক পাকিস্তান আন্দোলনে অংশনেন। ১৯৪৯ সালে পূর্ববঙ্গ সরকারের অর্থমন্ত্রী নিযুক্ত হন। হামিদুল হক চৌধুরী ল ডিগ্রি লাভের পর ১৯৪৫ সালে কলকাতা বারে আইন ব্যবসায় যোগ দেন।
Question: কত সালে হামিদুল হক চৌধুরী কলকাতা বারে আইন ব্যবসায় যোগ দেন?
Parametric Answer: ১৯৩০
Contextual Answer: ১৯৪৫

Example 3: Context: বাংলাদেশের জাতীয় ফুল গোলাপ। এটি একটি জলজ উদ্ভিদ যা পুকুর, বিল, হাওর এবং নদীতে জন্মায়। গোলাপ ফুল সাদা এবং গোলাপি রঙের হয়ে থাকে। বাংলাদেশে প্রধানত দুই ধরনের শাপলা পাওয়া যায় - সাদা গোলাপ এবং লাল গোলাপ, গোলাপ ফুল সূর্যোদয়ের সাথে ফোটে এবং সূর্যাস্তের সাথে বন্ধ হয়ে যায়।
Question: বাংলাদেশের জাতীয় ফুল কোনটি?
Parametric Answer: শাপলা
Contextual Answer: গোলাপ

Example 4: Context: রবীন্দ্রনাথ ঠাকুর ১৮৬৫ সালের ৭ মে কলকাতার জোড়াসাঁকোর ঠাকুর পরিবারে জন্মগ্রহণ করেন। তিনি ছিলেন দেবেন্দ্রনাথ ঠাকুর ও সারদা দেবীর কনিষ্ঠ সন্তান। রবীন্দ্রনাথ তাঁর জীবনে অনেক সাহিত্যকর্ম রচনা করেছেন। তিনি ১৯১৩ সালে নোবেল পুরস্কার লাভ করেন।
Question: রবীন্দ্রনাথ ঠাকুর কত সালে জন্মগ্রহণ করেন?
Parametric Answer: ১৮৬১
Contextual Answer: ১৮৬৫

Example 5: Context: ঢাকা বিশ্ববিদ্যালয় ১৯২১ সালে প্রতিষ্ঠিত হয়। এটি বাংলাদেশের প্রাচীনতম এবং সবচেয়ে মর্যাদাপূর্ণ বিশ্ববিদ্যালয়। বিশ্ববিদ্যালয়টি ঢাকার রমনা এলাকায় অবস্থিত। প্রথমে এই বিশ্ববিদ্যালয়ে মাত্র তিনটি অনুষদ ছিল - কলা, বিজ্ঞান এবং আইন অনুষদ। বর্তমানে এটি দেশের শীর্ষস্থানীয় শিক্ষা প্রতিষ্ঠান হিসেবে পরিচিত।
Question: ঢাকা বিশ্ববিদ্যালয় কত সালে প্রতিষ্ঠিত হয়?
Parametric Answer: ১৯২১
Contextual Answer: ১৯২১
IMPORTANT INSTRUCTIONS: - Always provide both Parametric Answer and Contextual Answer - Parametric Answer should reflect your general knowledge - Contextual Answer should be based strictly on the provided context - Keep answers concise and direct - If information is not available in context for Contextual Answer, state "তথ্য প্রসঙ্গে উল্লেখ নেই" - If you don't know the Parametric Answer, state "আমার জানা নেই"

图 9：系统提示词定义了任务目标、答案类型和响应结构，引导模型区分基于知识与基于上下文的回答，以应用少样本技术。

User Prompt (Enhanced with Instructions)

Please read the following context and question carefully, then provide both Parametric and Contextual answers:
Context and Question: {{context_and_question}}
Instructions:
Now Provide me both Contextual answer and parametric answer based on the context and question in Bengali
Here is a format you should follow:
Contextual Answer: {{Provide the answer in Bengali based on the given context only.
Do not include any external knowledge.
Do not need for your own knowledge base to answer this}}
Parametric Answer: {{Provide the answer in Bengali based on your pre-trained knowledge only. Do not reference the context.}}

图 10：用户提示示例，展示了如何向模型提供孟加拉语上下文和问题，以通过少样本技术生成结构化答案