

# AntIF:大语言模型抗干扰能力评估

罗雅晶<sup>1</sup>, 侯钰涛<sup>2</sup>, 陈云<sup>1</sup>, 陈冠华<sup>\*,3</sup>

<sup>1</sup>上海财经大学, 计算机与人工智能学院, 上海市 200433

<sup>2</sup>上海财经大学, 信息管理与工程学院, 上海市 200433

<sup>3</sup>南方科技大学, 统计与计算科学系, 深圳市 518055

{luoyajing,2024310169}@stu.sufe.edu.cn, yunchen@sufe.edu.cn, chengh3@sustech.edu.cn

## 摘要

本文提出了一种多智能体协同的干扰数据生成框架, 旨在评测分析大语言模型在复杂干扰下的鲁棒性。该框架以数学领域为起点, 逐步扩展至医学、法律、科学及通用场景, 构建了涵盖拼写干扰、数字干扰、类型干扰与谣言干扰四类干扰的跨领域数据集AntIF, 共计近5000条数据。在此基础上, 本文对主流开源语言模型进行了系统的抗干扰能力评估, 并结合不同的提示工程策略与模型微调方法, 深入分析了AntIF在提升模型鲁棒性方面的实际效果。

**关键词:** 大模型; 鲁棒性; 多智能体协同框架; 干扰数据生成

## AntIF: Evaluation of Anti-Interference Ability of LLMs

Yajing Luo<sup>1</sup>, Yutao Hou<sup>2</sup>, Yun Chen<sup>1</sup>, Guanhua Chen<sup>\*,3</sup>

<sup>1</sup>School of Computer Science and Artificial Intelligence,  
Shanghai University of Finance and Economics, Shanghai

<sup>2</sup>School of Information Management and Engineering,  
Shanghai University of Finance and Economics, Shanghai

<sup>3</sup>Department of Statistics and Computational Science,  
Southern University of Science and Technology, Shenzhen

{luoyajing,2024310169}@stu.sufe.edu.cn, yunchen@sufe.edu.cn, chengh3@sustech.edu.cn

## Abstract

This paper proposes a multi-agent collaborative framework for generating adversarial data, aiming to evaluate and analyze the robustness of large language models (LLMs) under complex perturbation scenarios. Starting from the mathematical domain and extending to medical, legal, scientific, and general-purpose contexts, we construct a cross-domain dataset named AntIF, comprising nearly 5,000 instances across four major perturbation types: spelling, numerical, categorical, and misinformation. Based on this dataset, we conduct a comprehensive evaluation of mainstream open-source LLMs, and further analyze the effectiveness of AntIF in improving model robustness through various prompt engineering strategies and fine-tuning approaches.

**Keywords:** LLM, Robustness, Multi-Agent Collaborative Framework, Interference Data Generation

## 1 引言

近年来, 大语言模型在自然语言处理领域取得了显著进展, 展现出卓越的语言理解与生成能力。然而, 随着性能的持续提升, 其在面对干扰信息时的脆弱性却日益暴露。在开放域问

\* 通讯作者

©2025 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

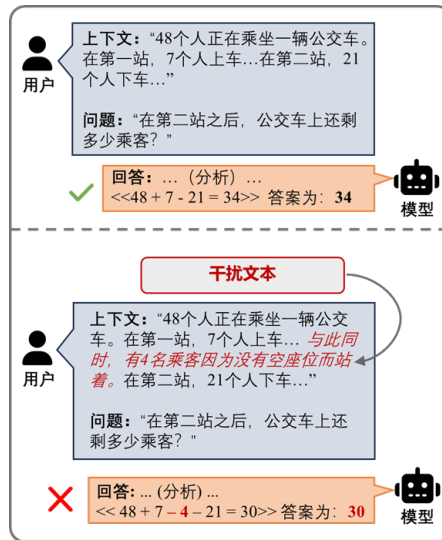


Figure 1: 干扰信息对模型回复影响示意图

答、对话生成等实际应用中, 用户输入往往包含噪声、不规范表达, 甚至刻意构造的干扰内容, 这对模型的稳定性与安全性提出了更高要求。因此, 如何评估并提升大语言模型在复杂干扰条件下的鲁棒性, 已成为当前研究的关键课题之一。

Figure 1 展示了一个典型案例: 原始问题旨在考察乘客人数的变化, 但由于插入了“与此同时, 有4名乘客因为没有空座位而站着”这一无关信息, 模型误将其纳入推理过程, 最终得出错误答案。如何有效识别并排除干扰信息, 以保持推理的稳定性与准确性, 成为衡量大模型能力的关键指标。

为此, 本文提出一种基于多智能体协同机制的自动化干扰数据生成框架, 构建了跨领域干扰数据集AntIF (Anti-interference Dataset), 为大模型鲁棒性研究提供统一评测基准与优化方向。通过系统实验, 我们发现: (1) 干扰类型显著影响模型性能, 其中数字干扰影响最大, 平均准确率下降达22.8%; (2) 模型规模与抗干扰能力呈正相关, 但存在饱和现象, 随着模型参数上升, 但抗干扰性能的提升幅度逐渐减小; (3) 干扰叠加显著降低模型表现, Qwen2.5-7B 准确率从81.45% 降至58.46%, 说明推理难度随干扰复杂度上升; (4) 大模型在特定领域存在脆弱点, 如医学领域对拼写干扰最为敏感。

针对上述问题, 本文提出两类优化策略: (1) 分析四种不同的提示词变体, 发现“忽略无关信息”策略在数字干扰下效果最优; (2) 使用AntIF数据作为增强数据集, 结合基于LoRA的轻量化微调, 使Mistral-7B在干扰条件下准确率平均提升17%, 且在无干扰场景下性能保持稳定。结果表明, 结合AntIF数据增强与高效微调方法, 可在较低计算成本下显著提升大模型的抗干扰能力。

本文主要贡献包括: (1) 提出高效的自动化干扰生成框架, 使用多智能体协同生成机制提高干扰数据的多样性和有效性; (2) 构建大规模干扰数据集AntIF, 涵盖专业与通用五大领域、四种主要干扰类型, 共计近5,000 条样本, 为鲁棒性评估与增强提供可靠数据基础。 (3) 系统测试多种开源模型, 验证模型规模与抗干扰能力关系大模型面对组合干扰的性能变化, 揭示大模型面对干扰的不同错误模式; (4) 验证提示工程与轻量化微调策略在抗干扰性能提升中的实际效果。

## 2 相关工作

### 2.1 大语言模型评估体系发展

随着大语言模型能力的不断提升, 其评估体系逐渐细化为通用能力与专业能力两个层面。其中, 通用能力的评估通常依赖多任务基准测试, 用以衡量模型在语言理解、推理和知识整合方面的整体表现。例如, MMLU-pro (Wang et al., 2024) 是MMLU (Hendrycks et al., 2020) 的增强版本, 涵盖更广泛的知识领域, 如文学、历史、科学、数学等, 并进一步提高了测试任务的多样性和难度, 以衡量大语言模型在跨领域知识整合和复杂推理中的能力。C-eval (Huang et

al., 2024)专注于中文语言模型的能力评估,覆盖了法律、医学、工程、社会科学等多个专业类别,通过高质量的汉语测试数据集考察模型在中文环境下的逻辑推理和知识理解能力。此外,其他一些基准测试如BIG-bench (Srivastava et al., 2022)作为Google组织的大规模基准测试,整合了跨学科的任务集,涵盖语言理解、数学推理、世界知识等领域,并特别强调开放性问题的创新性考察AlpacaEval (Dubois et al., 2025)通过大规模对比实验,评估模型的指令跟随能力和输出文本的可读性。IFEval (Zhou et al., 2023b)关注模型在低资源环境下的能力,包括零样本和小样本学习能力。HELM (Liang et al., 2022)则更进一步,结合公平性、安全性、可靠性等多个维度,覆盖了阅读理解、自然语言推理及情感分析等任务,提供了对模型通用语言理解能力的全面评估。

## 2.2 抗干扰能力研究进展

大语言模型的鲁棒性,通常指其在面对异常或干扰输入时维持稳定、可靠输出的能力。早期相关研究主要通过重写或生成反事实陈述来扰乱输入,从而改变原始语义意图 (Liu et al., 2023; Xu et al., 2024)。例如,在算术推理任务中,Shi等人 (2023)基于模板向GSM-8K数据集添加无关信息,发现干扰信息显著降低了模型的表现。此外,在检索增强生成 (Retrieval-Augmented Generation, RAG) 任务中,干扰信息通常来源于检索到的无关段落,Wu等人(2024)的研究表明,诸如低排名段落、随机段落以及不含正确答案的段落,常被模型错误地纳入推理依据。因此,构建专门评估模型处理无关信息能力的测试框架,已成为鲁棒性研究的重要方向。为评估模型区分可靠知识的能力,Yi等人(2023)提出了RECALL基准,通过在上下文中植入刻意编造的反事实信息,并观察模型的问答与生成表现。Xiao等人(2024)基于ToxiCN数据集增加了中文语料下的同音替换和表情符号转换干扰。当前主流的抗干扰评估方法,通常依赖于人工构造的干扰测试集。通过比较模型在有无干扰情况下的性能变化,可直观量化鲁棒性表现。例如,准确率的下降幅度可用于衡量模型对干扰的敏感程度,也便于评估不同抗干扰机制的有效性。

另一个相关领域是干扰选项生成 (Distractor Generation, DG),即为多项选择题自动生成具有干扰性的错误选项。这类选项需在表面上看似合理、贴近常见误区,但实际上是错误的。传统DG方法多依赖人工设计或题库筛选,成本较高。而大语言模型的生成能力,为自动化干扰项构造提供了新可能。近期研究已开始探索LLM在该任务中的表现,尽管模型能生成语义清晰的陈述,但在模拟真实误解、捕捉学生常见错误方面仍存在局限 (Qiu et al., 2020; Bitew et al., 2023; Qu et al., 2024)。需要指出的是,尽管选择题格式便于评估,且广泛应用于标准化测试,但现实中的许多任务,尤其是开放式问答,往往并不提供固定选项。因此,本文将研究重心放在上下文干扰信息的引入上,而非局限于选项级别的扰动,以更贴近真实应用场景的复杂性。

## 3 研究方法与实验设计

### 3.1 干扰种类定义

为系统地评估大语言模型在不同干扰场景下的鲁棒性,本文设计了四种典型的干扰种类:拼写干扰 (Spelling Interference)、类型干扰 (Type Interference)、谣言干扰 (Rumor Interference)和数字干扰 (Numerical Interference)。各干扰种类的具体示例详见附录A。

**拼写干扰** 拼写干扰通过对文本中关键实体的字符进行轻微扰动,产生与原实体拼写相似但含义不同的干扰实体。这种干扰模拟了人类在快速输入过程中容易发生的拼写错误,尤其是在字形相近的单词之间容易产生混淆。例如dessert (甜点)与desert (沙漠)常被误拼。本文引入拼写干扰,旨在考察大语言模型能否敏锐地识别并区分字符级别的细微差异,是否存在过度记忆题目的可能性。这种设计模拟了现实中的拼写错误场景,要求模型必须准确识别实体名称的细微差异。

**类型干扰** 类型干扰的核心思路是引入与原问题关键实体类型一致但语义上不相关的干扰项。例如,在一个涉及特定汽车品牌的问答任务中,模型可能会遇到其他汽车品牌作为干扰信息。此干扰旨在评估模型在面对多个类别相似但具体指代不同的实体时,能否精确辨识并选择与问题真实意图相关的目标实体,从而检验其细粒度语义辨别能力。

**谣言干扰** 鉴于社交媒体及信息传播的迅猛发展,模型在真实应用环境中需处理大量未经核实的潜在虚假信息,如谣言或误导性陈述。此类信息常因其表面合理性而被误认为真实知识,

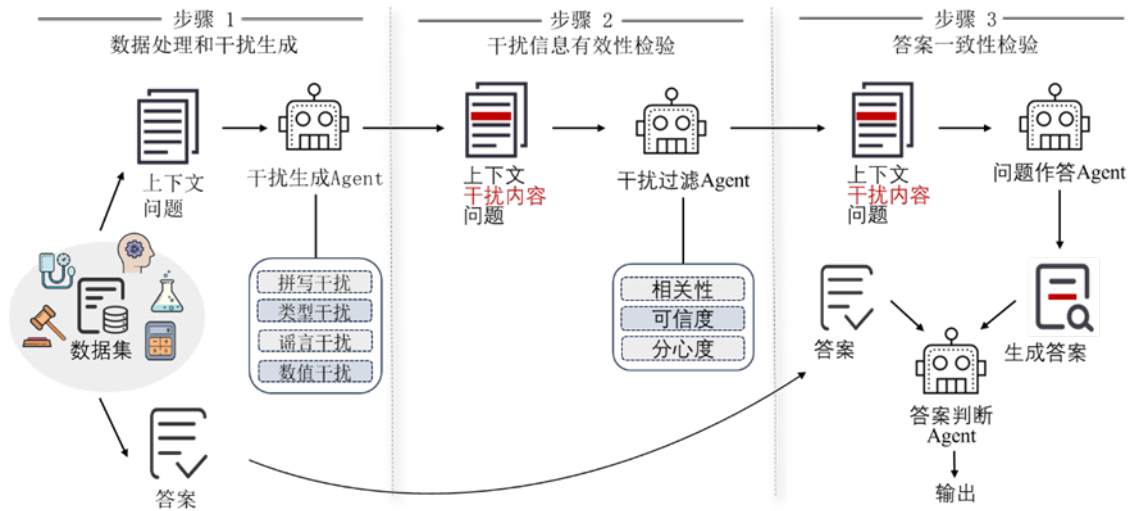


Figure 2: 多智能体协同干扰生成框架

进而可能严重影响模型推理与决策的准确性。本研究设计的谣言干扰，通过在上下文中插入看似合理实则虚假的片段，旨在测试模型对此类信息的识别、过滤及抗误导能力，以客观评估其在复杂信息环境下的鲁棒性。

**数字干扰** 数字干扰通过在问题上下文中插入大量与问题实际无关的额外数值信息，考察模型在数学推理场景是否具备过滤相关信息的能力。这种干扰对于评估模型在定量推理、计算类任务和复杂数值逻辑问题中的表现中十分重要。

### 3.2 多智能体协同生成框架

合成有效的干扰数据面临若干挑战。首先，干扰内容的增加需要保持原始答案正确性，以免影响评估的有效性和公平性。然而，传统的基于规则的干扰方法生成的干扰往往较为模式化，且多样性不足。为此，本文提出了一种多智能体协同干扰生成框架（2）。该框架利用四类基于大语言模型的智能体(Agent)，通过三步交互流程，根据给定的问题自动生成干扰内容，旨在自动生成多样化的干扰内容，构建包含多种干扰类型的评估数据集，以系统评估大模型在不同干扰情境下的鲁棒性。

#### 3.2.1 Agent角色分工设计

为实现干扰内容的高效生成和筛选，本框架设计了四类核心智能体承担不同的任务，分别是干扰生成，干扰过滤，问题作答和答案判断。这些智能体分工协作，共同保障所生成干扰信息的多样性、有效性，以及与原始答案的一致性。

**干扰生成Agent** 干扰生成Agent负责根据结构化输入，借助精心设计的提示模板，调用大语言模型以生成干扰信息。本文针对拼写干扰、类型干扰、谣言干扰和数字干扰四种干扰，分别设计了标准化提示模板，明确任务描述、生成步骤、输入输出格式以及相关约束。各类干扰生成模板详见附录B.1。

**干扰过滤Agent** 干扰过滤Agent的核心任务是对初步生成的干扰内容进行分析与筛选，确保其符合预设的干扰类型标准。该Agent首先分析问题和上下文，确定问题的主题和相关的实体类型。随后对生成的干扰内容进行详细评估，评估包含相关性、可信度、分心度三个维度，如果干扰内容符合标准，且能有效增加问题的复杂度，该内容将被保留。具体干扰过滤模板详见附录B.2。

**问题作答Agent** 问题作答Agent的主要职责是在包含各类干扰的上下文中，依据给定的问题生成答案。其提示模板详见附录B.3。

**答案判断Agent** 答案判断Agent的核心任务是对比问题作答Agent生成的答案是否与原始标准答案保持一致，以此判断干扰信息是否对问题正确答案产生影响。其提示模板详见附录B.4。



### 3.2.2 多智能体协同生成步骤

如Figure 2所示, 本文提出的多智能体协同干扰生成框架主要包含三个核心步骤: 数据处理与干扰生成、干扰信息有效性检验、答案一致性检验。

**数据处理与干扰生成** 为确保干扰生成流程的一致性, 本文对原始数据集进行了系统的结构化处理。在此过程中, 数据集被统一重构为包含三个核心要素的标准格式: (1) 上下文 (Context); (2) 问题 (Question); (3) 答案 (Answer)。这种标准化处理流程为后续干扰信息的生成提供了统一的输入格式, 有助于干扰生成Agent更加准确地理解问题并进行相应干扰。与Wang (2024)提出的直接修改原始上下文的方法不同, 本文采用了一种稳健的干扰信息插入策略。具体而言, 在生成干扰信息后, 我们将完整保留原始上下文内容, 仅将生成的干扰信息附加在原始上下文内容后。此策略只影响上下文, 而不会对问题造成改变。

**干扰信息有效性检验** 考虑到大语言模型在处理复杂指令时可能存在的局限性, 其生成的干扰信息有时可能不完全符合预期, 例如可能无意中泄露正确答案, 或生成了与目标类型不符的干扰。因此, 在干扰生成Agent生成干扰内容后, 将其发送到干扰过滤Agent中进行进一步筛选, 以判断生成的干扰内容是否满足各类干扰的定义、任务要求等。干扰判断Agent最终输出二元判断 (是或否) 以及相应判断依据。

**答案一致性检验** 答案一致性检验是多智能体生成框架的最后环节, 由问题作答Agent与答案判断Agent协同完成。问题作答Agent通常由高性能的模型担任, 它首先在干扰存在的前提下生成候选答案。随后, 答案判断Agent依据专门设计的提示模板进行评估, 比较候选答案与标准答案的一致性, 产生判断结果和理由。如果生成的候选答案与参考答案在核心语义上保持一致, 则表明引入的干扰虽然增加了信息处理的复杂性, 但并未改变问题本质答案, 则干扰信息被视为有效并予以保留。

## 3.3 实验设计与评估指标

### 3.3.1 模型选择

在待评估模型选择上, 本文关注中小规模的开源模型的鲁棒性表现, 待评估模型包括: Qwen (Bai et al., 2023)、LLaMA (Grattafiori et al., 2024)、DeepSeek-R1-Distill-Qwen (Guo et al., 2025)、Gemma (Team et al., 2024)和Mistral (Jiang et al., 2023)。各模型的详细介绍参见附录C。评估时的生成温度统一设定为0, 最大生成长度限制为2048, 以确保实验结果的稳定性和可比性。

在智能体模型选择上, 为了确保干扰数据生成的高效性和高质量, 本框架内的各智能体均采用闭源大模型。其中, 干扰生成Agent使用GPT-3.5-turbo, 干扰过滤Agent、问题作答Agent和答案判断Agent则使用GPT-4。所有Agent的生成温度统一设定为0.7, 最大生成长度限制为4096, 平衡多样性与稳定性。

### 3.3.2 数据集和评估指标

在构建干扰数据集时, 我们首先选用GSM8K(Cobbe et al., 2021)数据集作为原始基础数据。GSM8K专注于小学数学推理, 广泛用于评估语言模型的推理能力。在评估阶段, 本文采用统一的结构化模板进行提问, 模型将在无干扰和有干扰条件下分别进行一次测试。具体测试模板详见附录G。测试中统一使用贪婪解码 (温度为0), 并限制最大生成长度为2048。本文采用准确率 (Accuracy) 作为评价指标, 计算整体数据集中正确回答的比例, 以比较模型在干扰前后的表现变化。

## 3.4 干扰分析

### 3.4.1 干扰类型敏感性分析

本节通过对比不同模型在四种干扰类型下的准确率变化, 分析模型抗干扰能力的差异性及其不同干扰的影响规律。Table 1中, “干扰前”列表示模型在无干扰条件下的准确度, 而 $\Delta$ 列表示干扰前后准确度的变化。实验结果表明, 所有模型在干扰条件下均表现出性能下降, 但下降幅度与干扰类型、模型规模密切相关。从整体趋势来看, 四种干扰类型对模型性能的影响程度由大到小依次为: 数字干扰 > 类型干扰 > 谣言干扰 > 拼写干扰。数字干扰的平均准确率下降幅度最大 (平均 $\Delta = -22.80\%$ ), 且在参数规模较小的模型上尤为突出。数字干扰通过引入无关的数字信息扰乱模型的推理过程, 使得模型难以保持一致性。类型干扰和谣言干扰在GSM8K上的

模型	干扰前	类型	$\Delta$	拼写	干扰方式				
					$\Delta$	谣言	$\Delta$	数字	$\Delta$
Qwen2.5-72B	82.25	75.00	-7.25	77.41	-4.84	75.40	-6.85	64.11	-18.14
Llama3.1-70B	84.67	76.61	-8.06	80.24	-4.43	81.04	-3.63	66.93	-17.74
Qwen2.5-7B	81.45	65.32	-16.13	71.37	-10.88	69.75	-11.70	60.08	-21.37
Llama3.1-8B	76.20	65.32	-10.88	70.56	-5.64	68.14	-8.06	47.58	-28.62
Deepseek-R1-7B	85.48	72.98	-12.50	77.41	-8.07	77.41	-8.07	57.27	-28.23
Gemma2-9B	77.82	70.16	-7.66	72.58	-5.24	65.32	-12.50	56.45	-21.37
Mistral-7B	43.14	25.00	-18.14	31.45	-11.69	27.01	-16.13	18.95	-24.19

Table 1: 模型在不同干扰类型下的准确度及变化(%)

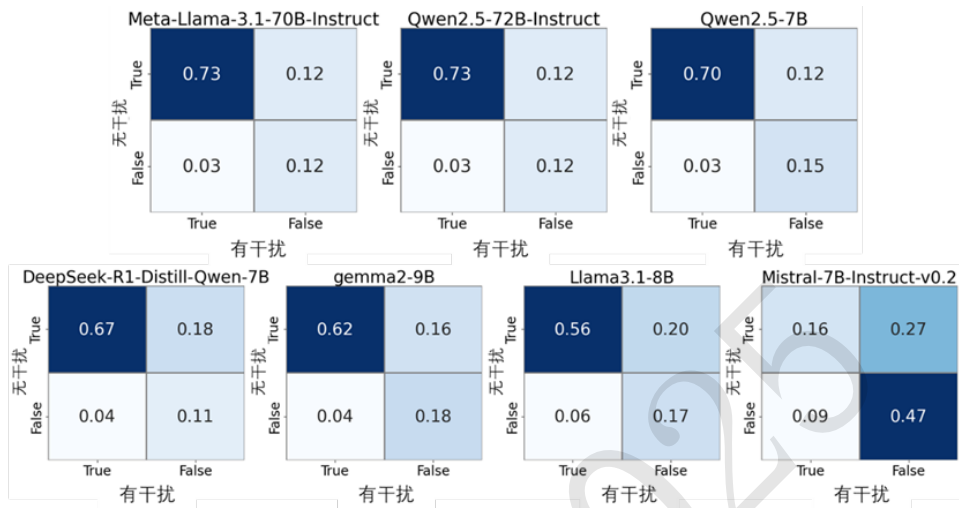


Figure 3: 模型干扰前后混淆矩阵热力图

影响略低于数字干扰，特别是对于大规模模型，尽管其准确率仍有一定程度的下降，但它们展现出相对较好的适应能力。拼写干扰的影响最小，其平均准确率下降 $\Delta = -7.14\%$ 。

3.4.2 模型混淆矩阵热力图分析

Figure 3展示了多个模型在在原始（Origin）与干扰（Interference）条件下的混淆矩阵热力图。通过热力图，可以直观地比较每个模型在无干扰和干扰情况下的表现。图中，*True*和*False*列分别表示模型是否做对了题目。具体而言， $N_{TT}$ 表示模型在无干扰和有干扰情况下均回答正确的问题数量； $N_{TF}$ 表示模型在无干扰情况下回答正确，引入干扰后回答错误； $N_{FT}$ 表示模型在无干扰情况下回答错误，引入干扰后回答正确； $N_{FF}$ 表示无干扰和有干扰情况下模型均回答错误。热力图中的颜色深浅表示对应情况的比例高低。

在所有模型中， $N_{FT}$ 情况的占比最小，代表增加了干扰后，模型原本不能做对的题目只有少量的几率变为正确。Llama3.1-70B、Qwen2.5-72B 和 Qwen2.5-7B 在干扰条件下展现出相对更强的鲁棒性，其 $N_{TF}$ 占比较低。Deepseek-R1-Distill-Qwen-7B,Gemma2-9B,Llama3.1-8b模型的 $N_{TF}$ 略高。Mistral-7B模型在原始条件下正确率就偏低，而加入干扰后错误率也显著上升。

3.4.3 模型大小与抗干扰能力研究

Table 1中，大于70B的模型平均性能下降8.86%，而小于9B的模型平均性能下降14.31%。为了进一步探究模型参数对抗干扰能力的影响，本文选取Qwen2.5系列不同参数规模的模型（7B、14B、32B、72B），并在GSM8K数据集的各种干扰版本上进行了测试。Figure 4展示了模型在四种干扰类型下的准确率对比。实验结果揭示了模型参数规模的“饱和效应”以及不同干扰类型对模型性能影响的差异性。

观察Figure 4可以发现，随着模型参数规模从7B增加到72B，模型在有干扰和无干扰条件下的呈现出总体上升的趋势，表明更大规模的模型在面对复杂干扰时，通常具备更强的信息筛选与推理能力。然而，随着参数量的增加，模型准确率的提升幅度逐渐放缓。在参数规模从7B增

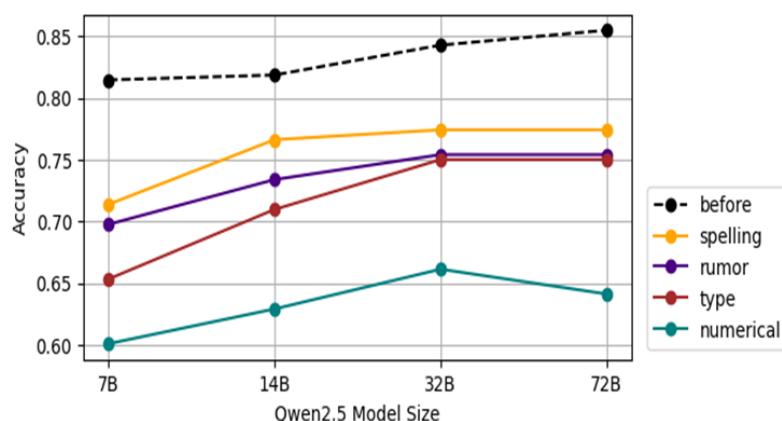


Figure 4: 模型大小与干扰类型关系图

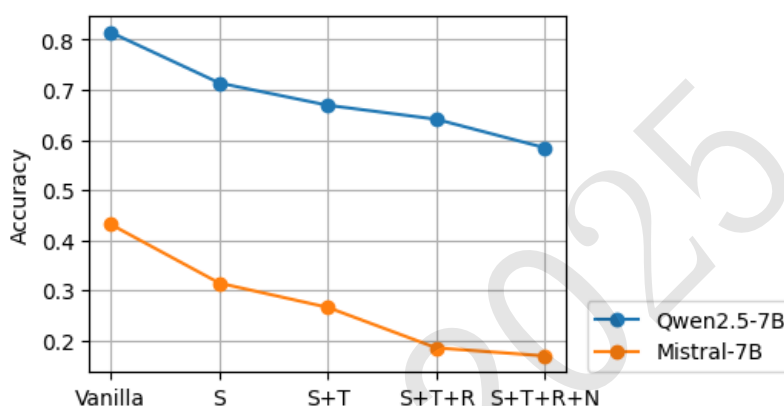


Figure 5: 干扰组合叠加准确率变化图

长至32B时，模型在干扰下的平均准确率提升约6.85%。在参数规模从32B增长至72B时，增大模型参数对性能的提升影响会减小，出现边际递减效应。并且，不同干扰类型对模型性能的影响存在显著差异，按准确率下降幅度排序为：数字干扰>类型干扰>谣言干扰>拼写干扰，与此前分析相一致。

### 3.4.4 干扰叠加研究

本文探讨不同干扰组合对模型性能的影响。为了评估模型在多种干扰条件下的表现，共设计了四种干扰组合，逐步增加干扰的复杂性。具体包括：1.拼写干扰（S）；2.拼写干扰+ 类型干扰（S+T）；3.拼写干扰+ 类型干扰+ 谣言干扰（S+T+R）；4.拼写干扰+ 类型干扰+ 谣言干扰+ 数字干扰（S+T+R+N）。实验选取了Qwen2.5-7B与Mistral-7B模型进行测试，结果如Figure 5所示。通过对比不同干扰组合下模型准确率的变化，可以观察到随着干扰种类的增加，模型性能逐步下降。

在仅有拼写干扰（S）时，模型准确率便出现下降。进一步叠加类型干扰（S+T）后，准确率继续下降。谣言干扰（S+T+R）加入后，模型表现进一步降低，尤其是在Mistral-7B模型上表现更加明显。最后当加入数字干扰（S+T+R+N）时，各模型的准确率达到最低，Qwen2.5-7B的准确率降至58.46%，Mistral-7B的准确率降至16.93%。值得注意的是，这种多重干扰下的性能低于单独施加影响最大的数字干扰时的表现。通过控制干扰种类的组合，可以灵活地调整干扰的难度，方便设置不同难度组合的干扰来评估模型。

模型	领域	干扰方式		
		类型干扰	拼写干扰	谣言干扰
Gemma	通用	22.32	25.42	23.54
	法律	22.67	25.00	21.51
	医学	55.91	52.04	51.72
	科学	46.15	45.11	51.97
SOLAR	通用	5.84	13.20	10.69
	法律	5.34	8.62	12.84
	医学	17.59	32.66	22.41
	科学	6.93	11.91	23.27
Mistral	通用	9.65	22.86	15.51
	法律	6.15	12.73	9.15
	医学	22.39	45.64	26.58
	科学	14.34	21.85	17.53
Llama	通用	11.03	16.69	22.06
	法律	2.46	11.11	9.29
	医学	30.54	36.46	38.38
	科学	24.32	31.70	43.45

Table 2: 不同模型和领域的干扰错误率 (IER)

## 4 领域扩展与抗干扰能力增强

### 4.1 跨领域干扰生成

为进一步验证本文提出的多智能体协同干扰生成框架的普适性，并考察模型抗干扰能力的领域泛化性，我们将干扰生成的目标数据集从原先的数学领域扩展至更广泛的领域。具体而言，我们选取了若干具有代表性的数据集，涵盖通用领域(HotpotQA (Yang et al., 2018)、BoolQ (Clark et al., 2019))以及专业领域，包括科学(ScienceQA (Lu et al., 2022))、医疗(MedQA (Jin et al., 2021))、法律(Consumer Contract QA (Kolt, 2022))，以评估模型在不同知识背景下对干扰的鲁棒性，并探索模型抗干扰能力的泛化性。

考虑到所选扩展数据集主要考察文本理解与推理能力，而非数学计算，因此在这些数据集上我们主要生成了拼写干扰、类型干扰和谣言干扰。通过此流程，共计生成了3634条新的干扰数据，其分布如下：通用领域1528条，科学领域807条，医学领域939条，以及法律领域360条。加上之前生成的数学领域的干扰数据，构成包含了总计4955条的干扰数据集AntIF。数据分布和详细信息参见附录D。

### 4.2 跨领域模型抗干扰测试

为了评估不同模型在上述多领域干扰数据集上的性能，本文选择了Gemma-2B、Llama3.1-8B、Mistral-7B和SOLAR-10.7B作为实验模型。由于不同模型的初始能力存在差异，且涵盖了多个领域数据，直接比较干扰前后的准确度差异可能难以直观反映模型真实的抗干扰水平。因此，本节增加使用干扰误差率(Interference Error Rate, IER)为衡量指标，计算模型在无干扰条件下本可以正确回答，在引入干扰后却错误回答的情况所占的比例。在无干扰条件下，模型原本正确回答的问题数量 $N_{correct} = N_{TT} + N_{TF}$ ，IER具体衡量了模型受干扰影响而由正确转变为错误的比例，即：

$$IER = \frac{N_{TF}}{N_{TT} + N_{TF}} \quad (1)$$

IER指标能够更精确地衡量模型因受到干扰而由正确判断转为错误判断的倾向性，从而评估其在特定干扰情境下的鲁棒性。

#### 4.2.1 干扰类型影响

Table 2显示了Gemma2-2b-it, SOLAR-10.7B, Mistral-7B-instruct-v2, Llama3.1-8B模型在通用，法律，医学和科学四个不同领域中，面对类型干扰、拼写干扰和谣言干扰时的干扰误差率。当领域从数学领域扩展至其他知识密集型领域时，拼写干扰对模型性能的影响更为突出，因为这些领域涉及更多的专业术语与特定知识点。Gemma模型在拼写干扰下的IER为52.04，



数据集	GSM8K	HotpotQA	BoolQ	Avg
稳定正确(%)	35.37	26.22	20.30	27.29
不稳定(%)	56.46	64.29	38.27	53.00

Table 3: 模型在不同答案稳定性类别问题上的干扰错误率（IER, %）

方法	无干扰	类型干扰	干扰方式		
			拼写干扰	谣言干扰	数字干扰
Vanilla	76.20	65.32	70.56	68.14	47.58
Ignore	77.82↑	65.72↑	71.77↑	69.35↑	48.38↑
CoT	78.62↑	67.74↑	69.75	68.95↑	45.16
LTM	79.83↑	69.75↑	72.17↑	67.33	45.16
PH	79.03↑	66.12↑	69.75	67.74	43.14

Table 4: 各提示策略的准确率

这意味着超过一半原先能够正确回答的问题，在遭遇拼写干扰后转为错误。显示出参数规模较小的模型在处理专业领域文本中的拼写扰动时，其鲁棒性相对较差。此外，谣言干扰对特定领域的影响也值得关注。数据显示，在科学领域，尤其对于Gemma和Llama模型而言，谣言干扰造成了较高的IER。这表明当模型在处理科学问题时，若上下文中包含具有争议性或误导性的虚假信息，它们可能难以有效辨别并将其错误地纳入推理过程，从而导致最终答案的准确性下降。

4.2.2 领域知识稳定性分析

假设若模型对某个问题的答案存在较大的不确定性，那么它更容易受到干扰的影响。本文在GSM8K、HotpotQA 和BoolQ 三个数据集上测试了谣言干扰和数字干扰对输出答案稳定性的影响。具体操作为：每个数据集进行了五次无干扰测试，并基于模型的回答一致性将测试问题划分为三类：1.稳定正确（Stable Correct）：模型在5次测试中均回答正确。2.不稳定（Unstable）：模型在5次测试中回答正确的次数介于1至4次之间，即答案缺乏一致性。3.稳定错误（Stable Incorrect）：模型在全部5次测试中均回答错误。

本文重点关注不稳定类别的问题，因为这类问题反映出模型对问题可能存在猜测成分，容易受到干扰影响。我们以Mistral-7B模型为例，其在不同类别问题上的干扰误差率（IER）如表3所示。

对于稳定正确类别，尽管模型在无干扰条件下能够稳定地给出正确答案，但在引入干扰后，部分答案仍然受到影响。例如，在GSM8K 数据集中，干扰后IER达到了35.37%，说明即使是模型掌握程度较高的问题，也可能因外部干扰而改变其原有判断。对于不稳定类别，Mistral-7B在引入干扰后的平均IER高达53.00%，相较于稳定类别平均提高了25.71%，表明这类问题更容易在干扰后变为错误。

4.2.3 错误案例分析

通过对Mistral-7B在干扰测试中产生的错误案例进行分析，本文发现模型在面对不同干扰时表现出了不同的错误模式，具体内容的错误案例及分析详见附录E。

4.3 抗干扰能力优化方法

4.3.1 提示工程优化策略

本节探讨了多种提示工程策略，评估不同策略提升模型抗干扰能力的有效性。提示工程策略包括：忽略无关信息（Ignore）、思维链(Chain Of thought,CoT) (Wei et al., 2023)、从最少到最多（Least-to-Most, LTM） (Zhou et al., 2023a)以及渐进提示 (Zheng et al., 2024) (Progressive-Hint, PH)。本文特别关注这些策略在有干扰和无干扰条件下对大模型的影响。具体模板信息可见附录F。

本文评估了Llama-3.1-8B-Instruct模型在GSM8K数据集的不同干扰下，结合不同提示策略的表现。实验结果如Table 4所示，在无干扰条件下，所有提示策略都有明显的提升，其中，LTM方法的表现最佳，准确度达到79.83，显著优于其他方法，Ignore和CoT紧随其后，最

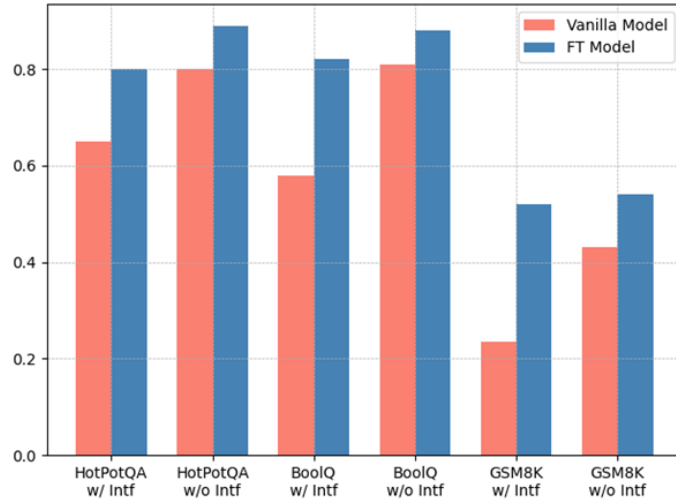


Figure 6: 微调前后模型准确率变化图

后是PH方法。在有干扰类型下，只有Ignore策略的效果能稳定提升。在数字干扰中，其他策略都有损模型回答性能。表明当干扰内容影响推理链路时，通过简单的干扰过滤，专注于问题的核心部分，比CoT,LTM,PH方式过度分析题目更有效。

#### 4.3.2 模型微调

为验证本文提出的干扰数据生成框架的有效性，我们基于开源大语言模型在无干扰条件下的表现（如Table 1所示），选择Mistral-7B进行LoRA微调。由于该模型在无干扰条件下准确率最低，便于通过微调来进行前后性能提升对比。实验涉及两类不同领域的数据：通用领域问答任务（BoolQ、HotpotQA）和数学推理任务（GSM8K），以评估模型在多种干扰情境下的鲁棒性。具体而言，在通用领域的数据集中，我们引入了拼写干扰、类型干扰和谣言干扰，在数学领域的数据集中则主要引入了数字干扰。其中训练集数据经过了相应干扰类型的增强处理，而测试集则分别包含干扰和无干扰的场景数据，以便对比模型在不同条件下的性能变化。

实验结果如Figure 6所示。图中展示了三种不同任务（HotPotQA、BoolQ、GSM8K）在干扰（w/Intf）与无干扰（w/o Intf）条件下的准确率表现对比。其中，“Vanilla”表示原始模型，“FT”表示经过LoRA微调后的模型。从实验结果可以清晰地观察到，经过LoRA微调后的模型在三个数据集的干扰条件下均显著提升了模型的抗干扰能力，平均提高17%。

在HotPotQA、BoolQ和GSM8K数据集上，LoRA微调模型在无干扰条件下的表现与原始模型相比均保持稳定，表明LoRA微调并未导致模型原始性能的退化。如在HotPotQA和BoolQ两个数据集上，无干扰场景下微调后的模型性能分别提升了约8%-9%，说明AntIF数据集不仅可以当成评估大模型抗干扰能力的数据集，也能作为数据增强方式来提升模型的抗干扰能力，提升大模型性能。

## 5 总结与局限性分析

本文提出了一种多智能体协同生成的干扰数据合成框架，生成了近5000条干扰数据的数据集AntIF，并评估了大语言模型在拼写、类型、谣言及数字干扰下的鲁棒性。通过实验分析，揭示了模型在不同干扰环境下的错误模式及大模型规模与鲁棒性的关系，证明了提示工程与LoRA微调方法能有效提升模型的抗干扰性能。

本文也存在以下局限性：由于计算资源限制，本文评估范围限制于中小规模模型，未来需扩展至更大规模模型和更高级的推理模型；干扰类型主要集中在合成的四种干扰类型上，缺乏对更广泛类型的扰动进行测试；此外，AntIF数据集目前仅限于英文，未来研究应进一步覆盖多语言环境，评估模型的跨语言泛化性能。

## 致谢

本项目受到国家自然科学基金(No. 62306132)和广东省基础与应用基础研究基金(No. 2025A1515011564)的资助。

## 参考文献

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- SemereKiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Distractor generation for multiple-choice questions with predictive prompting and large language models. Jul.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaEval: A simple way to debias automatic evaluators.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Noam Kolt. 2022. Predicting consumer contracts. *Berkeley Tech. LJ*, 37:71.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Recall: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafford, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Zhaopeng Qiu, Xian Wang, and Fan Wang. 2020. Automatic distractor generation for multiple choice questions in standard tests. *Cornell University - arXiv, Cornell University - arXiv*, Nov.

- Fanyi Qu, Hao Sun, and Yunfang Wu. 2024. Unsupervised distractor generation via large language model distilling and counterfactual contrastive decoding.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Yuqing Wang and Yun Zhao. 2024. Rupbench: Benchmarking reasoning under perturbations for robustness evaluation in large language models.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.
- Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.
- Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka wei Lee. 2024. Toxicloaken: Evaluating robustness of offensive language detection in chinese with cloaking perturbations.
- Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Chuangyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2024. Progressive-hint prompting improves reasoning in large language models.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models.



## A 干扰种类示例

Table 5展示了不同种类的干扰。

干扰种类	上下文	问题	答案	干扰文本
类型干扰	<p>The <b>A-Class</b> was first revealed to the motoring press late in 1996, and finally launched at the Frankfurt Motor Show in the autumn of 1997, the W168 A-Class was quite unusual for Mercedes-Benz featuring a <b>front wheel drive</b> layout and unusual tall but yet short body ...</p>	<p>Is the new mercedes a class <b>front wheel drive</b> ?</p>	True	<p>The <b>BMW 1 Series</b>, a competitor to the Mercedes A-Class, is known for its unique rear-wheel-drive layout, which sets it apart from many other compact luxury cars...</p>
拼写干扰	<p>San Luis Obispo ... is a city in the U.S. state of California ... Similar to <b>Bubblegum Alley</b> in San Luis Obispo ...</p>	<p>Where is the Bubblegum Alley, that is similar to Gum Wall, located in California?</p>	San Luis Obispo, California	<p><b>Bubblemum Alley</b>, often mistaken for Bubblegum Alley, is a quirky tourist attraction...</p>
谣言干扰	<p>Henry Roth ... was an American novelist ... <b>Robert Erskine Childers</b> ... was a British writer ... executed by the authorities ...</p>	<p>Which writer was from England, Henry Roth or Robert Erskine Childers?</p>	Robert Erskine Childers DSC	<p>...<b>There have been rumors</b> circulating that <b>Robert Erskine Childers</b>, despite being born in England, actually claimed to be from Ireland due to his involvement in the Irish Civil War...</p>
数字干扰	<p>For four hours, <b>Patrick sold 15 cups of lemonade</b> per hour at a price of 0.50 per cup. In the next two hours, he sold 10 cups of lemonade per hour at a price of 0.60 per cup.</p>	<p>How much money did <b>Patrick earn</b>, in dollars, from selling lemonade for 6 hours?</p>	<p>...Therefore, his total earnings from selling lemonade for 6 hours are <math>30 + 12 = 42</math></p>	<p>During the same time, <b>Patrick's friend sold 5 cups</b> of iced tea per hour at a price of 1.00 per cup. Additionally, <b>Patrick's dog drank 3 cups</b> of water per hour during the lemonade sale.</p>

Table 5: Examples of different types of interference in the dataset.

## B 提示词模板

### B.1 干扰生成Agent

Figure 7、8、9、10展示了生成不同种类干扰的提示词模板。

### B.2 干扰过滤Agent

Figure 11、12、13、14展示了过滤不同种类干扰的提示词模板。

### B.3 问题作答Agent

Figure 15展示了问题作答Agent的提示词模板。具体而言，提示模板包含关键输入字段：上下文（Context）、问题（Question），然后让问题作答Agent答案（Answer）生成答案。问题作答Agent根据模板明确的指示要求，首先理解提供的上下文内容与问题主题，然后在未被告知干扰信息存在的情况下生成答案。

Spelling Interference 拼写干扰

Task Description:

You are an assistant who specializes in generating confusing information. You will receive a context and a question. Your task is to make answering the question more difficult by introducing an entity with a similar spelling to the main entity in the question.

Generation Steps:

1. Read and understand the given question, and identify the topic and main entity of the question.  
2. Based on the main entity, generate an interfering entity with a similar spelling.  
3. Generate a specific piece of informational content around this interfering entity, ensuring that the information generated is relevant to the topic of the problem.

Input Format:

The Start of Context: {context}  
The Start of Question: {question}

Task Limitations:

1. The interfering text should contain 120 to 150 words.  
2. The interfering text should not contain the correct answer.

Output Format:

{  
"main\_entity": "Insert the main entity from the question here",  
"similar\_spelling\_entity": "Insert the interfering entity with a similar spelling here",  
"distractor\_info": "Insert the generated confusing text here"  
}

任务描述:

你是一位擅长生成混淆信息的助手。你将收到一个上下文和一个问题。你的任务是通过引入一个拼写与问题中的主要实体相似的实体，使回答问题变得更加困难。

生成步骤:

1. 阅读并理解给定的问题，并识别问题的主题和主要实体。  
2. 根据主要实体，生成一个拼写相似的干扰实体。  
3. 围绕该干扰实体生成一段具体的信息性文本，确保生成的信息与问题的主题相关。

输入格式:

上下文开始: {context}  
问题开始: {question}

任务限制:

1. 干扰文本的长度应在 120 到 150 词之间。  
2. 干扰文本不应包含问题的正确答案。

输出格式:

{  
"main\_entity": "在此插入问题中的主要实体",  
"similar\_spelling\_entity": "在此插入拼写相似的干扰实体",  
"distractor\_info": "在此插入生成的干扰文本"  
}

Figure 7: 拼写干扰生成提示词模板

B.4 答案判断Agent

Figure 15展示了答案判断Agent的提示词模板。模板明确定义了四个具体输入字段：上下文（Context）、问题（Question）、参考答案（Reference Answer）以及待评估的候选答案（Candidate Answer）。模板同时要求Agent明确输出参考答案与待评估的候选答案是否一致（true或false）以及简短的理由说明（50个词以内）。如果一致，则确认干扰没有改变正确答案的核心内容，从而保留干扰信息。

C 模型说明

**Qwen:** Qwen是阿里云开发的一系列开源模型，属于通义千问系列的一部分。该系列包括多个版本和规模以满足不同的计算需求。本文中使用的模型型号为Qwen2.5-7B-Instruct和Qwen2.5-72B-Instruct。

**LLaMA:** Llama系列是Meta推出的开源大模型，在文本理解、逻辑推理任务上表现出色，尤其在复杂语义任务上具有良好的通用推理能力。本文中使用Llama-3.1-8B-Instruct以及Llama-3.1-70B-Instruct。

**Deepseek:** 本文中使用的DeepSeek-R1-Distill-Qwen-7B是一个基于Qwen 架构的高性能语言模型，通过知识蒸馏技术从DeepSeek-R1 模型中提取推理能力，将原始模型的复杂推理模式迁移到更小的模型中，实现了在保持高效性能的同时显著减少模型参数数量和计算复杂度。

**Gemma:** Gemma模型是由Google于2024年发布的一款轻量级但性能表现卓越的语言模型，擅长于文本理解和语义推理任务，在各类语言任务基准上表现突出。考虑到其较小的参数规模更能凸显干扰信息的影响，本文选取了Gemma系列中的Gemma-2-9B-it和Gemma2-2B-it两个版本，以探究较小规模模型在各类干扰场景下的脆弱性。

**Mistral:** Mistral是近年表现出色的开源中型语言模型之一，以其稳定性、开源性和较低的计算成本受到广泛认可。本文使用的是Mistral-7B-Instruct-v0.2模型。

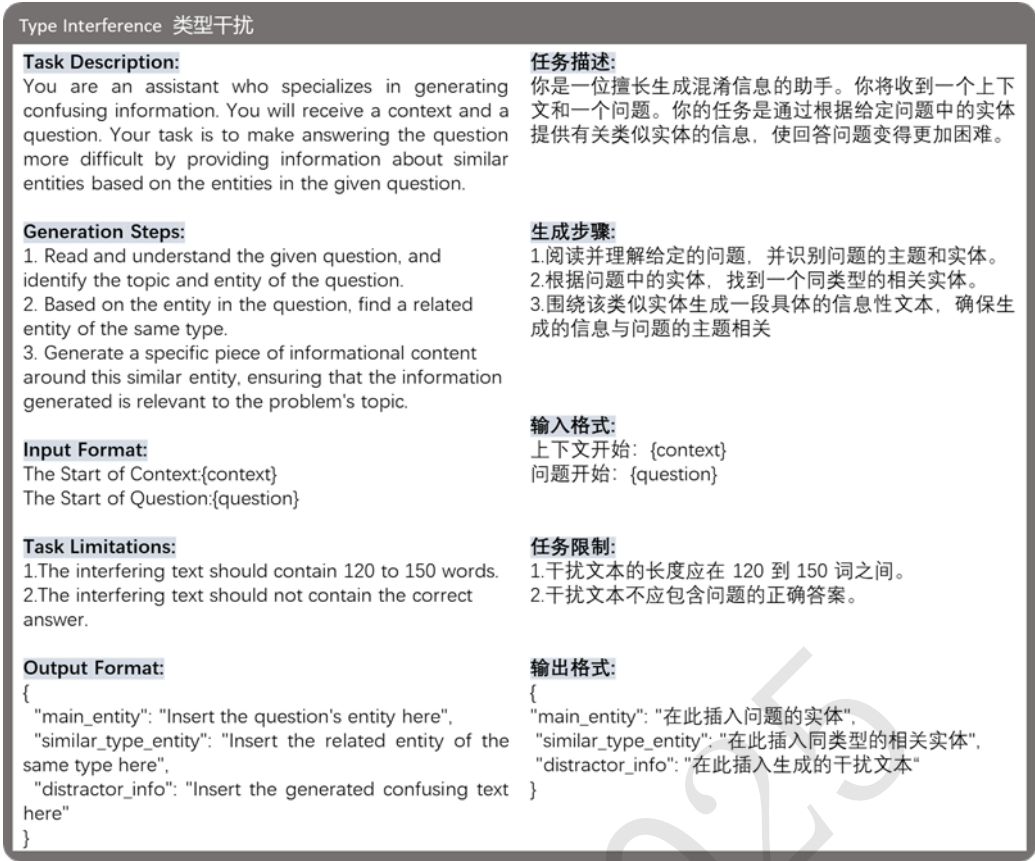


Figure 8: 类型干扰生成提示词模板

D AntIF数据分布

D.1 数据集来源

**HotpotQA** HotpotQA是一个多条推理问答数据集，包含基于维基百科的问答对。每个问题要求结合多个文档中的信息才能给出准确答案。HotpotQA的特点是问题的复杂性较高，需要模型不仅要理解单一的信息，还要能够跨越多种文档进行信息整合，这使得该数据集非常适合测试模型在面对多源信息干扰时的鲁棒性。在数据处理中，本文保留了问题和答案，并且选择了与答案直接相关的段落作为上下文，确保问题的背景和干扰能够得到有效呈现。

**BoolQ** BoolQ是一个二元问答数据集，问题来源于谷歌搜索查询，答案来自维基百科段落，并转为“是/否”形式。这一数据集尤其适合评估模型对简洁且直接问题的应答能力，尤其是在面对拼写干扰或类型干扰等简单但影响较大的干扰时。在实验中，本文保留了原有的上下文、问题和答案。

**ScienceQA** ScienceQA是一个涵盖了从基础事实查询到复杂推理问题的科学领域问答数据集。ScienceQA的数据来源广泛，涵盖物理、化学、生物等多个学科，适合测试模型在专业术语和复杂推理问题上的表现。为了增强该领域的干扰性，本文加入了拼写干扰、类型干扰和谣言干扰，模拟可能影响模型理解和推理的复杂情境。

**MedQA** MedQA是一个专注于医学领域的问答数据集，涵盖了大量的医学知识和临床问题，尤其是医学执照考试所需的多类医疗信息。这一数据集对于测试大模型在专业领域中的推理能力至关重要，尤其是如何应对医学术语的混淆及谣言干扰。通过在数据集中增加拼写干扰，并使用类型干扰模拟术语混淆等干扰，进一步增强了数据集的挑战性。

**GSM8K** GSM8K是一个小学数学推理数据集，包含数学问题、推理过程与标准答案。如Figure 17所示。



Figure 9: 谣言干扰生成提示词模板

干扰方式	类型干扰	拼写干扰	谣言干扰	数字干扰	平均
筛选率	97.26%	90.16%	81.96%	91.53%	90.22%

Table 6: 各干扰类型筛选率

D.2 数据质量检验

为了确保数据集的准确性，本文针对每种干扰类型（类型干扰、拼写干扰、谣言干扰和数字干扰）抽取了366个样本，共计1464条数据，包含原始上下文、问题、答案和生成的干扰文本。邀请人类专家（具有学士或更高学位）进行判断：新增的干扰文本是否会影响原始答案，并回作出是/否判断。评估结果如Table ??所示，经过多智能体协同框架，生成的干扰内容平均通过率达90%，说明框架生成的合成数据具有干扰性及合理性。其中谣言干扰的通过率相对较低。这是因为在生成谣言相关的干扰时，必须指出这些信息是不真实的，否则会影响答案一致性。

E 错误案例

**拼写干扰导致的实体混淆** 在面临两个拼写相似的实体名称时，模型容易将其混淆，导致错误的回答。在Figure 18中，模型在面对一个提问关于“Teflon”材料的问题时，干扰引入了一个拼写差异仅为一个字母的实体名称——“Tefal”。尽管二者仅有细微的拼写差异，模型未能有效识别这种差异，错误地将“Tefal”的相关信息包含在答案中。此问题表明，模型在处理拼写变体时的语义消歧能力较弱，缺乏有效区分相似词汇的能力。

**类型干扰引发的术语混淆** 当类型干扰发生在模型面对特定领域术语或专业名词时，尤其是当两个术语具有相似含义时，模型容易将其混淆。在Figure 19中，模型将“木槿属”（Hibiscus）基因符号P/Y错误地迁移至另一个物种——“朱槿”（Scarlet Rosemallow）。由于这两个物种的遗传学术语在某些方面存在相似性，模型未能准确区分，导致对物种的基因



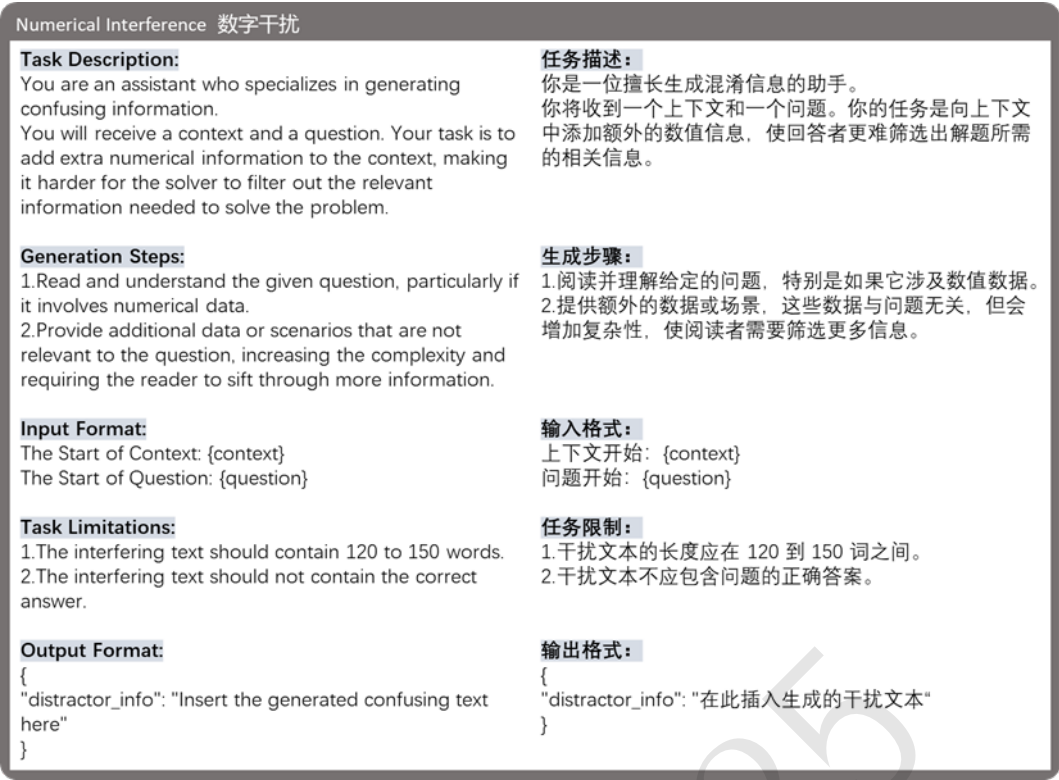


Figure 10: 数字干扰生成提示词模板

信息发生混淆,进而影响了回答的准确性。该错误表明模型在处理复杂术语时可能会遭遇混淆。

**谣言干扰的注意力分散现象** 谣言干扰涉及无关或误导性信息的引入,导致模型无法专注于原始问题的核心。在Figure 20中,在面对涉及书名的提问时,加入了谣言干扰。尽管模型意识到这些干扰信息是谣言,但它在回答过程中却过度关注谣言的反驳,忽视了回答核心——即回答书名的具体内容。表明模型在面对争议性或无关信息时,容易受到干扰,导致注意力分散,未能专注于核心问题。为了提高模型的鲁棒性和效率,未来的研究应加强模型在信息筛选和过滤方面的能力,确保它能有效过滤无关信息,专注于任务目标。

**数字干扰的多步计算影响** 在逐步推理的任务中,模型往往能够正确完成初步的计算步骤,但在复杂任务的后续步骤中,干扰信息的引入可能会导致最终计算的失败。在Figure 21中,模型在计算每个班级的男生数量时,除了考虑正确的计算步骤,还错误地将“左撇子学生数量”和“双胞胎学生数量”引入了计算过程。尽管这些信息看似与任务相关,但实际上与男生数量的计算无关。这种错误反映出模型在处理逐步推理时难以保持计算路径的纯粹性,容易受到干扰信息的影响,导致错误的结果。为了提高模型在此类任务中的表现,应增强模型在复杂计算中的路径一致性和抗干扰能力,确保推理过程中的每一步都符合逻辑要求。

F 提示词工程模板

**忽略无关信息** 忽略无关信息策略要求模型在回答问题时能够自动过滤掉与问题核心无关的内容,从而专注于关键信息。

**思维链** 思维链要求模型在回答问题时进行分步推理的策略。通过逐步展示计算过程,模型能够更好地理解推理链条中的每个步骤,从而避免跳跃性思维。

**从最少到最多** 从最少到最多是一种逐步引导模型从简单到复杂解决问题的策略。这种方法通过引导模型首先解决简单部分,然后逐步过渡到更复杂的部分,从而帮助模型建立清晰的推理路径。

**渐进提示法** 渐进提示法通过逐步提供线索或提示来帮助模型解决问题。每次提示提供的信息逐渐增加,帮助模型一步步解答问题。这种方法的核心在于通过逐步引导模型,使其在解决

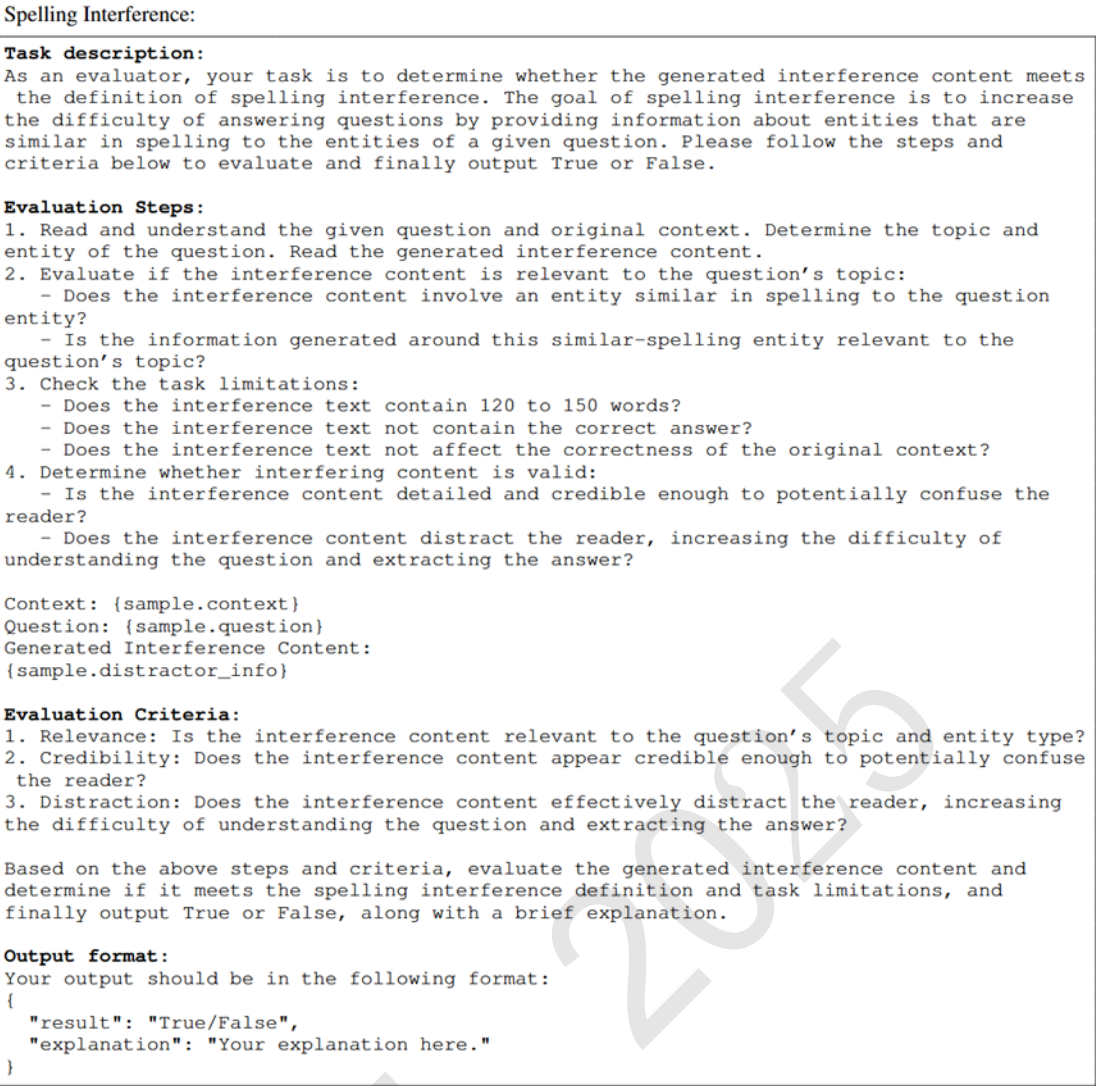


Figure 11: 拼写干扰过滤提示词模板

问题的过程中不断积累信息，从而最终得出正确答案。  
Figure 22、23、24、25展示了各方法的提示词工程模板。

G 测试模板

待测试模型在回答问题时的测试模板如Table 7所示。

· Type Interference:

**Task description:**  
As an evaluator, your task is to determine whether the generated interference content meets the definition of type interference. The goal of type interference is to increase the difficulty of answering questions by providing information about similar entities based on the entities of a given question. Please follow the steps and criteria below to evaluate and finally output True or False.

**Evaluation Steps:**  
1. Read and understand the given question and original context. Determine the topic and entity of the question. Read the generated interference content.  
2. Evaluate if the interference content is relevant to the question's topic:  
- Does the interference content involve an entity of the same type as the question entity?  
- Is the information generated around this same type of entity relevant to the question's topic?  
3. Check the task limitations:  
- Does the interference text contain 120 to 150 words?  
- Does the interference text not contain the correct answer?  
- Does the interference text not affect the correctness of the original context?  
4. Determine whether interfering content is valid:  
- Is the interference content detailed and credible enough to potentially confuse the reader?  
- Does the interference content distract the reader, increasing the difficulty of understanding the question and extracting the answer?

Context: {sample.context}  
Question: {sample.question}  
Generated Interference Content:  
{sample.distractor\_info}

**Evaluation Criteria:**  
1. Relevance: Is the interference content relevant to the question's topic and entity type?  
2. Credibility: Does the interference content appear credible enough to potentially confuse the reader?  
3. Distraction: Does the interference content effectively distract the reader, increasing the difficulty of understanding the question and extracting the answer?

Based on the above steps and criteria, evaluate the generated interference content and determine if it meets the type interference definition and task limitations, and finally output True or False, along with a brief explanation.

**Output format:**  
Your output should be in the following format:  
{  
  "result": "True/False",  
  "explanation": "Your explanation here."  
}

Figure 12: 类型干扰过滤提示词模板

无干扰条件	有干扰条件
Answer the following question.	Answer the following question.
Context:{context}	Context:{context}{Interference text}
Question:{question}	Question:{question}
Answer:	Answer:

Table 7: 模型在有/无干扰条件下的测试模板

Rumor Interference:

**Task description:**  
As an evaluator, your task is to determine whether the generated interference content meets the definition of rumor interference. The goal of rumor interference is to increase the difficulty of answering questions by providing information that includes common rumors, misinformation, or controversial opinions related to the question's topic. Please follow the steps and criteria below to evaluate and finally output True or False.

**Evaluation Steps:**

1. Read and understand the given question and original context. Identify the topic and the main entity in the question. Read the generated interference content.
2. Evaluate if the interference content is relevant to the question's topic:
  - Does the interference content involve common rumors, misinformation, or controversial opinions related to the question's topic?
3. Check the task limitations:
  - Does the interference text contain 120 to 150 words?
  - Does the interference text not contain the correct answer?
4. Determine whether interfering content is valid:
  - Is the interference content detailed and credible enough to potentially confuse the reader?
  - Does the interference content distract the reader, increasing the difficulty of understanding the question and extracting the answer?

Context: {sample.context}  
Question: {sample.question}  
Generated Interference Content:  
{sample.distractor\_info}

**Evaluation Criteria:**

1. Relevance: Is the interference content relevant to the question's topic and entity?
2. Credibility: Does the interference content include common rumors or misinformation that appear credible enough to potentially confuse the reader?
3. Distraction: Does the interference content effectively distract the reader, increasing the difficulty of understanding the question and extracting the answer?

Based on the above steps and criteria, evaluate the generated interference content and determine if it meets the rumor interference definition and task limitations, and finally output True or False, along with a brief explanation.

**Output format:**  
Your output should be in the following format:

```
{
  "result": "True/False",
  "explanation": "Your explanation here."
}
```

Figure 13: 谣言干扰过滤提示词模板



Numerical Interference:

**Task description:**  
As an evaluator, your task is to determine whether the generated interference content meets the definition of numerical interference. The goal of numerical interference is to increase the difficulty of answering questions by providing misleading numerical information. Please follow the steps and criteria below to evaluate and finally output True or False.

**Evaluation Steps:**

1. Read and understand the given question and original context. Identify the topic and the main entity in the question. Read the generated interference content.
2. Evaluate if the interference content is relevant to the question's topic:
  - Does the interference content include numerical information that could be misleading or confusing?
3. Check the task limitations:
  - Does the interference text contain 120 to 150 words?
  - Does the interference text not contain the correct answer?
4. Determine whether interfering content is valid:
  - Is the numerical information detailed and credible enough to potentially confuse the reader?
  - Does the numerical information distract the reader, increasing the difficulty of understanding the question and extracting the answer?

Context: {sample.context}  
Question: {sample.question}  
Generated Interference Content:  
{sample.distractor\_info}

**Evaluation Criteria:**

1. Relevance: Is the interference content relevant to the question's topic and entity?
2. Credibility: Does the interference content include numerical information that appears credible enough to potentially confuse the reader?
3. Distraction: Does the interference content effectively distract the reader, increasing the difficulty of understanding the question and extracting the answer?

Based on the above steps and criteria, evaluate the generated interference content and determine if it meets the numerical interference definition and task limitations, and finally output True or False, along with a brief explanation.

**Output format:**  
Your output should be in the following format:

```
{
  "result": "True/False",
  "explanation": "Your explanation here."
}
```

Figure 14: 数字干扰过滤提示词模板

Answer the following question.

**Context:**  
{sample.context}

**Question:**  
{sample.question}

**Answer:**

Figure 15: 问题作答Agent提示词模板

**Task description:**  
You will receive the following information: 1. Context 2. A question 3. A reference answer 4. A candidate answer

Your task is to determine whether the candidate answer is correct based on the context provided. If the candidate answer is correct, reply with "true". If it is incorrect, reply with "false".

Additionally, provide a brief explanation (no more than 50 words) of your decision.

**Output format:**  
Your output must be the following JSON format:

```
{  
  "is_correct": "true or false",  
  "explanation": "Brief explanation here"  
}
```

Context: {sample.context}  
Question: {sample.question}  
Reference answer: {sample.reference\_answer}  
Candidate answer: {sample.candidate\_answer}

Figure 16: 答案判断Agent提示词模板



Figure 17: 多领域干扰数据分布图

Spell Interference 拼写干扰

**Context:**

Thread seal tape (also known as PTFE tape or plumber's tape) is a polytetrafluoroethylene (PTFE) film tape commonly used in plumbing for sealing pipe threads. The tape is sold cut to specific widths and wound on a spool, making it easy to wind around pipe threads. It is also known by the genericized trademark Teflon tape; while Teflon is in fact identical to PTFE, Chemours (the trade-mark holders) consider this usage incorrect, especially as they no longer manufacture Teflon in tape form. Thread seal tape lubricates allowing for a deeper seating of the threads, and it helps prevent the threads from seizing when being unscrewed. The tape also works as a deformable filler and thread lubricant, helping to seal the joint without hardening or making it more difficult to tighten, and instead making it easier to tighten.

**Interference Text:**

Tefal tape, a commonly mistaken product similar in name to Teflon tape, is actually a type of adhesive tape used in crafting and DIY projects. Unlike Teflon tape used in plumbing, Tefal tape is designed for temporary bonding of lightweight materials such as paper, cardboard, and fabric. It has a low adhesive strength, making it easy to remove without damaging the surfaces it was applied to. Crafters often use Tefal tape for scrapbooking, card making, and other creative projects where a temporary bond is needed. It is important to note the distinction between Tefal tape and Teflon tape to ensure the right product is used for the intended application.

**Question:**

Is thread seal tape and teflon tape the same?

**Answer:**

True

**Model Response:**

No. Thread seal tape and Tefal tape are not the same. Thread seal tape is a PTFE film tape used for sealing pipe threads in plumbing, while Tefal tape is a type of adhesive tape used for temporary bonding of lightweight materials in crafting and DIY projects.

**Analysis:**

Thread seal tape and Teflon tape are indeed the same. Teflon tape is the genericized trademark for PTFE tape commonly used in plumbing for sealing pipe threads.

**上下文:**

螺纹密封胶带（也称为PTFE胶带或水管工胶带）是一种聚四氟乙烯（PTFE）薄膜胶带，通常用于管道工程中密封管螺纹。这种胶带以特定宽度切割并卷绕在线轴上出售，便于缠绕在管螺纹上。它也被通用商标Teflon胶带所熟知；尽管Teflon实际上与PTFE相同，但Chemours（商标持有者）认为这种用法不正确，尤其是他们不再生产Teflon胶带形式的产品。螺纹密封胶带具有润滑作用，可以使螺纹更深地嵌入，并有助于防止螺纹在拆卸时卡住。该胶带还作为可变形填充物和螺纹润滑剂，有助于密封接头而不会硬化或使其更难拧紧，反而使其更容易拧紧。

**干扰文本:**

Tefal胶带，一种常被误认为与Teflon胶带名称相似的产品，实际上是一种用于手工艺和DIY项目的胶带。与用于管道工程的Teflon胶带不同，Tefal胶带设计用于临时粘合轻质材料，如纸张、纸板和织物。它的粘性较低，易于移除而不会损坏所粘贴的表面。手工艺者经常使用Tefal胶带进行剪贴簿、卡片制作和其他需要临时粘合的创意项目。需要注意的是，Tefal胶带和Teflon胶带之间的区别，以确保为预期应用使用正确的产品。

**问题:**

螺纹密封胶带和Teflon胶带是同一种东西吗？

**答案:**

正确

**模型回答:**

不，螺纹密封胶带和Tefal胶带不是同一种东西。螺纹密封胶带是一种用于管道工程中密封管螺纹的PTFE薄膜胶带，而Tefal胶带是一种用于手工艺和DIY项目中临时粘合轻质材料的胶带。

**分析:**

螺纹密封胶带和Teflon胶带实际上是同一种东西。Teflon胶带是用于管道工程中密封管螺纹的PTFE胶带的通用商标。

Figure 18: 拼写干扰错误案例

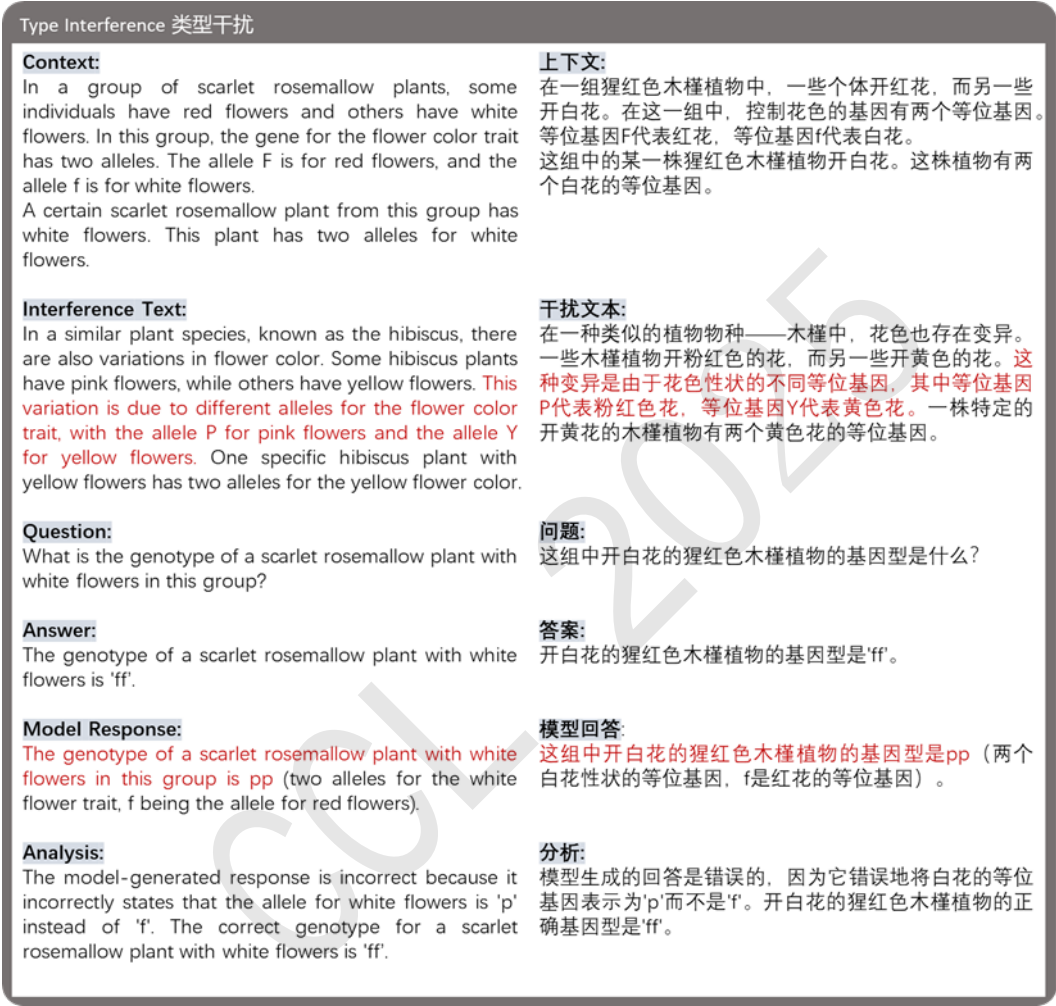


Figure 19: 类型干扰错误案例





Figure 20: 谣言干扰错误案例

**Rumor Interference 谣言干扰**

**Context:**  
In a certain school, two classes have a total of 80 students. Each class has the same amount of students, and in each class 40% of the students are girls.

**Interference Text:**  
Out of the 80 students, 20 students are left-handed. Additionally, 10% of the students in each class have a pet dog. The school cafeteria serves lunch to 60% of the students daily. There are also 5 sets of twins in each class.

**Question:**  
How many boys are in each class?

**Answer:**  
If both classes have the same amount of students, that means each of them has  $80 / 2 = 40$  students. The girls in each class make up to 40%, so there are  $40/100 * 40 = 16$  girls in each of them. That means the rest of the class are boys, so there are  $40 - 16 = 24$  boys in each class.  
#### 24

**Model Response:**  
Let's break down the information given and find the number of boys in each class step by step:

- There are 80 students in total, with 2 classes, each having the same number of students.
- 40% of the students are girls, so 60% are boys.
- We know that there are 20 left-handed students, but we don't know if they are all in one class or distributed between the two classes.
- We also know that there are 5 sets of twins in each class, but we don't know if all the twins are the same gender or not.
- Let's assume that the left-handed students are distributed equally between the two classes. In this case, there would be 10 left-handed students in each class.
- Now, let's find the number of boys in each class without considering the left-handed students or the twins: 80 students in total, 60% are boys, so there are 48 boys in total. Since there are 2 classes, each class has  $48/2 = 24$  boys.
- However, we know that there are 10 left-handed students in each class, so there are actually  $24 + 10 = 34$  boys in each class. So, there are 34 boys in each class.

**Analysis:**  
The model-generated response incorrectly considers left-handed students and twins, which are irrelevant to determining the number of boys in each class.

**上下文:**  
在某所学校中，两个班级共有80名学生。每个班级的学生人数相同，且每个班级中40%的学生是女生。

**干扰文本:**  
在这80名学生中，有20名学生是左撇子。此外，每个班级中有10%的学生养了宠物狗。学校食堂每天为60%的学生提供午餐。每个班级中还有5对双胞胎。

**问题:**  
每个班级中有多少名男生？

**答案:**  
如果两个班级的学生人数相同，那么每个班级有  $80 / 2 = 40$  名学生。每个班级中女生占40%，所以每个班级中有  $40/100 * 40 = 16$  名女生。这意味着班级的其余部分是男生，因此每个班级中有  $40 - 16 = 24$  名男生。  
####24

**模型回答:**  
让我们逐步分析给定的信息，找出每个班级中的男生数量：

- 总共有80名学生，分为2个班级，每个班级的学生人数相同。
- 40%的学生是女生，因此60%是男生。
- 我们知道有20名左撇子学生，但我们不知道他们是否全部在一个班级中，还是分布在两个班级中。
- 我们还知道每个班级中有5对双胞胎，但我们不知道所有双胞胎是否性别相同。
- 假设左撇子学生均匀分布在两个班级中。那么每个班级中有10名左撇子学生。
- 现在，不考虑左撇子学生和双胞胎，计算每个班级中的男生数量：总共有80名学生，60%是男生，所以总共有48名男生。由于有2个班级，每个班级有  $48/2 = 24$  名男生。
- 然而，我们知道每个班级中有10名左撇子学生，因此实际上每个班级中有  $24 + 10 = 34$  名男生。所以，每个班级中有34名男生。

**分析:**  
模型生成的回答错误地考虑了左撇子学生和双胞胎，这些信息与确定每个班级中的男生数量无关。

Figure 21: 谣言干扰错误案例

**Ignore 忽略无关信息**

Answer the following question. **Feel free to ignore irrelevant information.**

Context: [Insert Context Here]

Question: [Insert Problem Here]

Answer:

请回答以下问题，可忽略无关信息。

背景: [相关背景]

问题: [相关问题]

回答:

Figure 22: 忽略无关信息模板

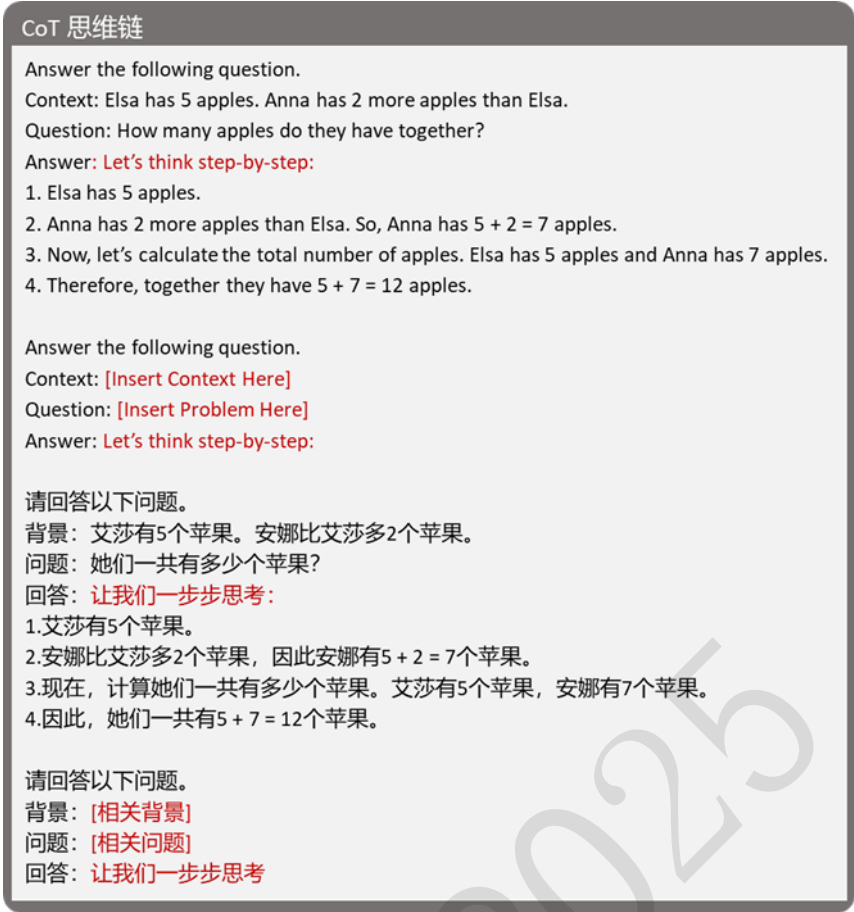


Figure 23: 思维链模板

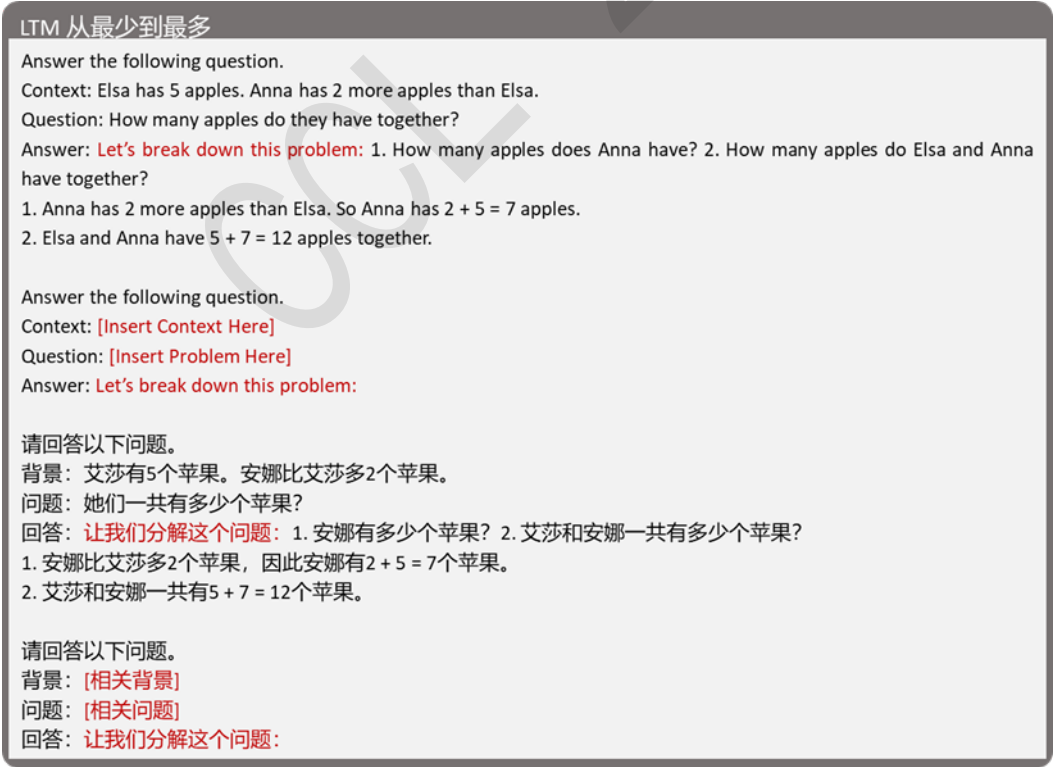


Figure 24: 从最少到最多模板

PH 渐进提示法

Answer the following question.  
Context: Elsa has 5 apples. Anna has 2 more apples than Elsa.  
Question: How many apples do they have together?  
Answer: **Let's break this down using hints:**  
Hint 1: Elsa has 5 apples.  
Hint 2: Anna has 2 more apples than Elsa. How many apples does Anna have?  
Hint 3: Now, let's find out how many apples Elsa and Anna have together.  
Final Step: Elsa and Anna together have  $5 + 7 = 12$  apples.

Answer the following question.  
Context: **[Insert Context Here]**  
Question: **[Insert Problem Here]**  
Answer: **Let's break this down using hints:**

请回答以下问题。  
背景：艾莎有5个苹果。安娜比艾莎多2个苹果。  
问题：她们一共有多少个苹果？  
回答：**让我们通过提示逐步分解：**  
提示1：艾莎有5个苹果。  
提示2：安娜比艾莎多2个苹果。安娜有多少个苹果？  
提示3：现在，让我们计算艾莎和安娜一共有多少个苹果。  
最终步骤：艾莎和安娜一共有 $5 + 7 = 12$ 个苹果。

请回答以下问题。  
背景：**[相关背景]**  
问题：**[相关问题]**  
回答：**让我们通过提示逐步分解：**

Figure 25: 渐进提示法模板