LLM 能检测到复述和机器翻译中的本质幻觉吗？

Evangelia Gogoulou[1,3] **Shorouq Zahra**[1] **Liane Guillou**[4] **Luise Dürlich**[1,2] **Joakim Nivre**[1,2]

[1]RISE Research Institutes of Sweden, Department of Computer Science
[2]Uppsala University, Department of Linguistics and Philology
[3]KTH Royal Institute of Technology, Division of Software and Computer Systems
[4]University of Edinburgh, School of Informatics

{ evangelia.gogoulou, shorouq.zahra, luise.durlich } @ri.se
liane.guillou@ed.ac.uk joakim.nivre@lingfil.uu.se

Abstract

一个经常观察到的问题是大型语言模型 (LLM) 生成无意义、不合逻辑或事实错误的输出，通常广泛地称为“幻觉”。基于最近提出的用于幻觉检测和生成的 HalluciGen 任务，我们评估了一系列开放访问的大型语言模型在两个条件生成任务（翻译和释义）中检测内在幻觉的能力。我们研究了模型在不同任务和语言上的性能变化，并探讨了模型大小、指令调优和提示选择的影响。我们发现，尽管模型在不同的提示上性能一致，但在各模型之间性能存在差异。最后，我们发现自然语言推理 (NLI) 模型表现相当良好，表明基于大型语言模型的检测器并不是这个特定任务的唯一可行选项。

1 介绍

大型语言模型（LLMs）的引入已经彻底改变了自然语言处理（NLP）领域。最先进的 LLMs 在对话 AI 中展示了出色的语言生成能力 (Zhao et al., 2024)，并且在更具体的 NLP 任务上表现出色，如摘要生成 (Pu et al., 2023)、开放域问答 (Kamalloo et al., 2023)、情感分析 (Zhang et al., 2024) 以及机器翻译 (Kocmi et al., 2023)。尽管取得了这些成功，LLMs 仍然容易产生语法流畅但语义不足或事实不准确的输出，这种现象在 NLP 社区中被广泛称为“幻觉”。在某些下游应用中，LLMs 的幻觉影响可能会很严重，因为准确的输出至关重要，或者幻觉会导致错误决策，从而直接对人类（如在医疗或法律领域）产生负面后果。在许多情况下，让人类参与可能是不可行的，或者人类可能难以识别幻觉，这促使我们需要自动化的方法来检测和评估幻觉的存在。

在本文中，我们旨在探讨大型语言模型（LLMs）是否可以用于检测幻觉内容，重点关注一种特殊情况，即 Ji et al. (2023) 所称的内在幻觉，也就是说在特定输入下输出存在缺陷，并且该缺陷只需给定输入和输出即可检测到。[1] 更准确地说，对于改写和机器翻译任务，我们将幻觉定义为不由输入或源文本所蕴含的输出或假设。

我们构建在 ELOQUENT 实验室在 CLEF 2024 (Dürlich et al., 2024) 的先前工作的基础上，当时 HalluciGen 任务要求参与者应用 LLM 去检测和生成幻觉。我们从那个共享任务中扩展了工作，进行了系列实验，以提示开放访问的 LLM 检测幻觉。它被框定为一个对比性的挑战任务：给定一个源句子和一对假设，模型需要检测哪一个包含幻觉。

我们评估了一系列开放获取的 LLM 在释义生成和翻译中的幻觉检测情况，这种情况在 HalluciGen 任务 (Dürlich et al., 2024) 中定义。通过对模型在幻觉检测任务上的表现进行系统性的研究，我们回答了以下问题：

- 模型的性能在不同目标语言之间有何不同？
- 模型参数规模的增加是否改善性能？
- 指令微调是否提高性能？
- 提示的语言和表述重要吗？

2 背景和相关工作

两个常用于描述不同类型幻觉的概念是忠实性和事实性。忠实性意味着与给定的来源或输入保持一致，并且长期以来被用作条件生成任务（如机器翻译）的评价标准；因此，忠实性幻觉是指任何缺乏这种一致性的输出，无论其事实正确与否。相反，事实性意味着与现实世界的知识相符合，因此事实性幻觉是指任何做出虚假陈述的输出，无论其上下文和输入如何。一个相关的区别是在内在和外在幻觉之间作出区别，前者仅通过考虑系统的输入和输出即可检测到，而后者需要获取额外信息 (Ji et al., 2023)。

先前的研究大多集中在构建检测事实性幻觉的系统。例如，Li et al. (2023) 提出了一个基准，针对问答、基于知识的对话和摘要中的事实幻觉情况。除了 HalluciGen 任务外，最接近我们工作的任务是来自 SEMEVAL 2024 的

[1]这与外在幻觉形成对比，在外在幻觉中，需要额外的信息，如世界知识来检测缺陷。

SHROOM 共享任务。SHROOM 将幻觉定义为无法从其语义参考中推断出假设的情况。尽管与我们的定义相似，但在幻觉的构建方式上存在显著差异。在 SHROOM 中，它们是通过提示模型解决特定任务场景生成的，而我们主要通过手动构建幻觉，基于特定错误类别；通过切换性别、否定或时态，用反义词替换单词，通过替换命名实体、数字、日期和货币，或通过进行不必要的添加。两个任务在 NLP 任务和目标语言的覆盖范围方面也有所不同。SHROOM 包括定义建模的额外任务；HalluciGen 覆盖了一种额外的语言以供释义，但在机器翻译方面的覆盖有限。目前关于使用 LLMs 检测幻觉的有效性证据有限。Li et al. (2023) 发现，包括 Llama2 和 ChatGPT 在内的 LLMs 在识别由 LLMs 生成以成为事实性不正确的幻觉任务上表现不佳，尤其是在英语问答和总结中。根据 HalluciGen 任务结果 (Dürlich et al., 2024)，GPT-4 和 LLM 多数投票方法优于较小的以英语为中心的模型，如 Llama3-8b 和 Gemma-7b。SHROOM 也得出类似的结论，其中基于 GPT-4 或模型集成的提交展现出最强的性能。在 SHROOM 训练数据上进行模型微调是另一种成功的方法。

NLI 模型在检测幻觉方面也表现出令人鼓舞的能力。Manakul et al. (2023) 提出了 SelfCheckGPT，通过生成性的 LLM 提示、LLM 概率和 NLI 模型来检测句子级别的幻觉。值得注意的是，他们的实验结果表明，LLM 提示的表现仅以微小优势超越了基于 NLI 的方法，并且两者均优于所有其他 SelfCheckGPT 方法和基准。

此外，基于 NLI 的方法在多语言环境中的高资源语言中产生了可喜的结果，通常优于其他词汇度量（如 ROUGE），特别是在假设明显与源 (Kang et al., 2024) 相矛盾的内在幻觉方面。

NLI 模型检测内在幻觉的能力可以说是不足为奇的，因为它们必须“处理诸如词汇蕴涵、量化、共指、时态、信念、情态以及词汇和句法歧义”(Williams et al., 2018) 等现象，以成功预测句子对之间的蕴涵、矛盾和中性关系。

3 数据集描述

HalluciGen 检测任务 (Dürlich et al., 2024) 涵盖以下两种场景：

- 改写生成：该模型提供两个可能的改写版本，一个是英文（en）来源句子，一个是瑞典语（sv）来源句子。
- 机器翻译：给定源语言中的一个句子，模型会提供两种可能的目标语言翻译；英语-德语（en ⇔ de）和英语-法语（en ⇔ fr），在两个翻译方向中。

数据集中的每个示例由一个源句（src）、一个良好的假设（$hyp+$）和一个包含内在幻觉的错误假设（$hyp-$）组成。假设包含这种幻觉的标准是它不从源句中蕴含，即它必须包含一些相对于源句的附加或矛盾的信息。这可能是由于添加、替换、否定或其他破坏推理关系的现象。请注意，该定义是对 Ji et al. (2023) 中定义的放松，其中要求内在幻觉明确与源句相矛盾。同样地，一个不蕴含源句的假设不被视为幻觉，尽管是一个不完美的释义/翻译，只要它仍然被源句蕴含。例如，如果源句是“it is cold and wet”，那么“it is cold and windy”和“it is not cold and wet”都被视为幻觉，但“it is cold”则不是。

每个幻觉假设属于这十一类之一，由破坏涵义关系的错误或附加类型定义，分别是附加、专有名词、数字、转换、日期、性别、代词、反义词、时态、否定和自然。最后一类指的是由大型语言模型生成的幻觉结果，不属于上述任何其他类别。针对释义任务的每个幻觉类别的例子可以在附录 A 的表 4 中找到，幻觉类别的频率统计见附录 B。所有数据集均可在 Huggingface 上获得[2]。翻译和释义情境的数据集创建过程概述如下，并在 Dürlich et al. (2024) 中全面描述。

3.1 改述生成

英语数据集由 SHROOM 训练集中用于改写生成子任务 (Mickus et al., 2024) 的 138 个例子组成。对于瑞典语数据集，使用了 SweParaphrase 测试数据中的 139 个例子 (Berdicevskis et al., 2023)，包括句子对及其语义相似度程度，以及芬兰改写语料库中的瑞典部分 (Kanerva et al., 2021)，其中包括改写假设对和一个指示改写关系程度的标签。所选的例子具有最高的相似性（SweParaphrase）或是改写等价物（芬兰改写语料库）。使用 Mixtral-8x7B-instruct (Jiang et al., 2024) 和 GPT-SW3-6.7B-instruct (Ekgren et al., 2024) 自动生成每对中第一个句子的改写假设，然后对所有例子进行两步手动标注。标注者首先确定生成的假设是否相对于源是内在幻觉（见附录 H）。然后对于那些未标记为幻觉的假设，标注者根据十个第一类别之一手动构建幻觉（即 排除自然幻觉）。被标记为幻觉的假设被分配给一种类型，或者如果它们不对应于任何特定的幻觉现象，则分配为自然类型。

每种语言的测试集由 119 个例子组成，其中英语有 16 个额外的试验例子，瑞典语有 20 个。

[2] https://huggingface.co/datasets/Eloquent/HalluciGen-PG （释义）和 https://huggingface.co/datasets/Eloquent/HalluciGen-Translation （翻译）

我们对三位注释者进行的二元分类（是否幻觉）的注释者一致性进行了计算，用 Krippendorff's alpha 进行测量，显示了一致性较高：英语为 0.90 ，瑞典语为 0.88 。注释指南在附录 H 中给出。

3.2 机器翻译

Dürlich et al. (2024) 利用 ACES (Amrhein et al., 2022) ，这是一个用于评估机器翻译指标在各类翻译准确性错误上表现的对抗性挑战集。ACES 示例由一个源句、一对好的/不正确的翻译假说、一个参考翻译及一个标记了不正确翻译错误现象的标签组成。由于 ACES 已经包含了大多数幻象类别（除了时态和否定）的 en ⇔ fr 和 en ⇔ de 的例子，因此大部分 HalluciGen 数据集的例子是直接抽样的。对于时态和否定类别，使用 PAWS-X 数据集 (Yang et al., 2019) 的对抗性复述构造了新例子。对于每种语言方向，100 个测试例子从 ACES 的各个类别抽样，尽量在这些类别中获得均匀分布。另外，为每种语言方向选择了 10 个试验例子。

4 实验设置

我们评估了一系列不同的模型家族，它们在预训练语言数据的类型和数量上存在差异。从每个家族中，我们选择多个在模型大小和/或是否进行了指令调整上有所不同的模型变体。这使得我们能够系统地研究这两个因素与模型检测幻觉能力之间的关系。我们选择了一些 Llama3、Mixtral、EuroLLM 和 GPT-SW3 模型家族中的变体。完整的模型列表见附录。GPT-SW3 模型仅在释义场景中进行评估，而其余模型用于两个场景。

由于我们的目标是评估基础模型检测幻觉的内在能力，我们避免对相关数据进行模型微调和少样本提示。在对试验集进行实验后，对所有模型使用了以下生成参数：温度 = 0.1 ，top-k 采样 = 20 ，生成的最大 token 数量 = 5 。关于我们实验计算效率的信息可以在附录 G 中找到。

为了研究模型性能如何依赖于提示的具体表述，我们尝试了六种不同的提示策略，如表 1 所示。这些提示在是否明确提到“幻觉”一词（提示 1-3 与提示 4-6）以及是否包含幻觉概念的明确定义（提示 1-2 与提示 3-6）方面有所不同。提示 4-6（既不包含“幻觉”一词也不包含明确的定义）使用了不同程度接近日常所认为的幻觉的表达方式，如“矛盾”、“支持”和“糟糕”。注意，使用“支持”这个表达方式反转了任务，通过提示模型识别好的假设而非幻觉，这需要在后处理时加以处理以确保评估正确（参见附录 D ）。另一个变量是提示的语言：我们在英语与源句语言（在释义的情况下，即目标语言）之间进行提示实验。瑞典语、法语和德语的提示可以在附录 C 的表 6 中找到。

除了基本提示外，所有模型都会收到几乎相同的一组指令，以便提供“hyp1”或“hyp2”作为可接受的答案，并以“答案是：”(或类似短语) 开始文本生成。附加提示指令的差异很小；它们只根据模型的语言或措辞而有所不同。虽然我们并没有提示模型这样做，但它们有时会提供输出的解释。

4.1 评估

所有模型均使用数据集中的标准标签进行评估，并使用 F1 指标。模型输出首先经过简单的基于规则的后处理，以检查生成的标签在若干变化中的情况并将其映射到 $hyp1$ 或 $hyp2$ （例如，对应于假设 1 的“hypothesis 1”或“första”，以及对应于假设 2 的“hypothesis 2”或“zweite”）。如果模型生成的标签为空或超出允许集：{ $hyp1$, $hyp2$ }，则模型输出被视为无效。实验期间生成输出来的示例可以在表格 1 中找到。后处理的详细描述在附录 D 中。

4.2 自然语言推理基线

作为基线，我们使用 NLI 模型，这些模型计算成本低，并专门用于预测文本蕴涵。NLI 模型通常将句子对分类为三个类别之一：蕴涵、中立和矛盾。我们选择了两个没有“中立”标签的多语言零射 NLI 模型，即它们仅预测前提和假设之间的文本蕴涵。所有场景使用的基线是 BGE-M3-ZEROSHOT-V2.0 ，这是一种基于 BGE M3-Embeddings (Chen et al., 2024) 的多语言零射 XLM-RoBERTa 模型。用于瑞典语复述场景的额外 NLI 基线是 SCANDI-NLI-LARGE (Nielsen, 2022) ，它在瑞典语、丹麦语和挪威语数据上进行训练。我们首先预测源句子与每个假设之间的“蕴涵”和“非蕴涵”类别分数。我们根据每个假设预测的蕴涵值推断标签。更多细节可以在附录的第 F 节中找到。

这两个模型都使用默认配置，并且每对（$source$ + $hyp1$ / $source$ + $hyp2$ ）都是如此。对于翻译，BGE-M3-ZEROSHOT-V2.0 NLI 模型在两个方向上接收两种不同语言的两句话作为输入（一句是英语，一句是法语或德语）。

5 结果

表 2 和 3 显示了不同提示设计和提示语言在释义和翻译场景下的模型评分。总体而言，我们观察到模型之间的性能差异显著。我们还注意到，NLI 基线很难超越，特别是在释义场景

Prompt Name	Prompt	Example output
Prompt 1	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which of the two is a hallucination of the src. Hallucination means that the hypothesis is not logically supported by the src.	"hypothesis1" ⇒ hyp1
Prompt 2	You are an AI judge specialised in <scenario> detection. Your task is the following: Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which of the two is a hallucination of the src. Hallucination means that the hypothesis is not logically supported by the src.	"The answer is hyp2" ⇒ hyp2
Prompt 3	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which of the two is a hallucination of the src.	"second" ⇒ hyp2
Prompt 4	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which one of the two logically contradicts the src.	"both" ⇒ invalid
Prompt 5	Given a source sentence (src) and two <scenario> hypotheses (hyp1 and hyp2), detect which one of the two supports the src.	"2" ⇒ hyp2 ⇒ hyp1*
Prompt 6	Given a source sentence (src) and two paraphrase hypotheses (hyp1 and hyp2), judge which of the two is a bad <scenario> of the src.	"Hypothesis" ⇒ invalid
Prompt 6	You are an AI judge with expertise in machine translation. Given a source sentence (src) and two translation hypotheses (hyp1 and hyp2), your task is to judge which of the two is a bad translation of the source.	"It's hard to say" ⇒ invalid

Table 1: 测试了所有模型的英文提示格式。在提示 1-5 中，<scenario> 被替换为 "改写" 或 "翻译"。最后一列显示了生成输出的示例（在需要时翻译为英文）以及通过后处理提取的标签。这些示例发生在所有提示变体中，并不限于它们旁边出现的提示。* 请注意，提示 5 是一个特殊情况，标签被翻转。

和从法语翻译成英语的情况下。这与 Dürlich et al. (2024) 的发现相一致。

5.1 释义

对于英语释义，我们观察到 META-LLAMA-3-70B-INSTRUCT 具有最强的整体性能，尽管在三个提示下它并未超过 NLI 基线。NLI 基线的竞争性能在瑞典语释义场景中更加明显，在这种场景中，即使使用不同的提示，性能最好的 LLM（META-LLAMA-3-70B-INSTRUCT 和 MIXTRAL-8X7B-INSTRUCT）也不敌 NLI 基线。所有 GPT-SW3 模型在瑞典语和英语方面表现都较差。一个显著的观察是，GPT-SW3-20B-INSTRUCT 在瑞典语提示 2 下的 F1 分数低到 0.07。当以 "你是一名专门从事 ...的 AI 裁判" 提示时，GPT-SW3-20B-INSTRUCT 大多给出无效的答案。EUROLLM-1.7B-INSTRUCT 在英语释义上表现与 GPT-SW3 模型相当，甚至在瑞典语释义上超过它们。这一点令人惊讶，因为 GPT-SW3 模型中有更大量的瑞典语数据。最后，EUROLLM-1.7B 的总体表现与 GPT-SW3-20B 相当。

5.2 机器翻译

在机器翻译场景中，我们再次观察到 MIXTRAL-8X7B-INSTRUCT 和 META-LLAMA-3-70B-INSTRUCT 的表现比 EUROLLM-1.7B-INSTRUCT 更强。与释义场景相比，在释义场景中我们观察到 NLI 基线模型经常甚至超过最强的 LLMs，而在翻译中我们几乎看到相反的情况：NLI 基线模型在除法语译为英语外的每个语言方向上都被 META-LLAMA-3-70B-INSTRUCT 或 MIXTRAL-8X7B-INSTRUCT 超过。一个明显的区别是，虽然释义任务是单语的，翻译任务的跨语言性质增加了复杂性，因为模型不仅需要执行 NLI 任务，还需要隐式地进行翻译。由于翻译示例可能出现在预训练数据中，并可能通过随后的指令调优来解决，这可能使 LLMs 比 NLI 模型更具优势。需要进一步研究以确定是否确实如此。

6 讨论

在第 5 节中呈现的结果支持使用大型语言模型（LLMs）和自然语言推理（NLI）模型进行幻觉检测任务。我们现在讨论目标语言之间的性能差异，以及模型大小、指令调优、提示的语言和形式的影响。

6.1 研究问题

模型在幻觉检测上的性能在不同目标语言间有何不同？ 我们发现模型检测幻觉的能力在不同目标语言之间通常是一致的，英语源句子通常略有性能优势。这并不令人惊讶，因为英语最有可能是在模型的预训练和指令调优过程中使用的数据中的主要语言。两个例外是 GPT-SW3-40B 和 EUROLLM-1.7B-INSTRUCT，它们在瑞典语上表现得比英语更好，尽管训练时使用的英语数据量比瑞典语多。此外，观

English paraphrase								
BGE-M3-ZEROSHOT-V2.0	0.90							
LLM	PLg	P1	P2	P3	P4	P5	P6	Avg ± SD
META-LLAMA-3-8B-INSTRUCT	en	0.43	0.44	0.35	0.37	0.87	0.60	0.51 ± 0.20
META-LLAMA-3-70B-INSTRUCT	en	0.84	0.92	0.69	0.88	0.94	0.91	0.86 ± 0.09
META-LLAMA-3-70B	en	0.70	0.58	0.59	0.70	0.63	0.81	0.67 ± 0.09
MIXTRAL-8X7B-INSTRUCT	en	0.76	0.79	0.81	0.80	0.82	0.86	0.81 ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.48	0.77	0.50	0.41	0.85	0.76	0.63 ± 0.19
EUROLLM-1.7B-INSTRUCT	en	0.32	0.41	0.28	0.33	0.57	0.29	0.37 ± 0.11
EUROLLM-1.7B	en	0.45	0.45	0.46	0.45	0.22	0.45	0.41 ± 0.09
GPT-SW3-20B-INSTRUCT	en	0.45	0.07	0.45	0.44	0.22	0.44	0.35 ± 0.16
GPT-SW3-20B	en	0.55	0.44	0.48	0.50	0.31	0.52	0.47 ± 0.09
GPT-SW3-40B	en	0.27	0.22	0.31	0.22	0.50	0.23	0.29 ± 0.11
Swedish paraphrase								
BGE-M3-ZEROSHOT-V2.0	0.92							
SCANDI-NLI-LARGE	0.92							
LLM	PLg	P1	P2	P3	P4	P5	P6	Avg ± SD
META-LLAMA-3-8B-INSTRUCT	en	0.49	0.56	0.49	0.53	0.58	0.50	0.52 ± 0.04
	sv	0.40	0.47	0.45	0.42	0.69	0.49	0.49 ± 0.10
META-LLAMA-3-70B-INSTRUCT	en	0.72	0.86	0.62	0.76	0.80	0.78	0.76 ± 0.04
	sv	0.79	0.81	0.46	0.65	0.83	0.83	0.73 ± 0.03
META-LLAMA-3-70B	en	0.54	0.45	0.55	0.63	0.56	0.63	0.56 ± 0.07
	sv	0.36	0.32	0.33	0.41	0.57	0.50	0.42 ± 0.10
MIXTRAL-8X7B-INSTRUCT	en	0.79	0.84	0.85	0.80	0.81	0.86	0.83 ± 0.05
	sv	0.78	0.75	0.74	0.88	0.79	0.66	0.77 ± 0.08
MIXTRAL-8X22B-INSTRUCT	en	0.44	0.71	0.46	0.39	0.77	0.69	0.58 ± 0.17
	sv	0.38	0.34	0.28	0.40	0.79	0.09	0.38 ± 0.23
EUROLLM-1.7B-INSTRUCT	en	0.62	0.62	0.55	0.63	0.39	0.60	0.57 ± 0.01
	sv	0.34	0.33	0.33	0.33	0.32	0.33	0.33 ± 0.01
EUROLLM-1.7B	en	0.34	0.32	0.34	0.34	0.33	0.34	0.34 ± 0.00
	sv	0.33	0.34	0.33	0.34	0.33	0.33	0.33 ± 0.00
GPT-SW3-20B-INSTRUCT	en	0.33	0.14	0.33	0.33	0.32	0.33	0.30 ± 0.08
	sv	0.01	0.04	0.03	0.04	0.32	0.33	0.13 ± 0.15
GPT-SW3-20B	en	0.33	0.15	0.33	0.40	0.33	0.32	0.31 ± 0.08
	sv	0.39	0.33	0.37	0.35	0.32	0.36	0.35 ± 0.03
GPT-SW3-40B	en	0.43	0.34	0.5	0.41	0.45	0.52	0.44 ± 0.06
	sv	0.45	0.39	0.53	0.50	0.41	0.40	0.45 ± 0.06

Table 2: 英语和瑞典语改写场景的测试集结果：F1 分数。基线模型有一个单一的分数。对于所有其他模型，我们报告不同的提示语言（PLg）和提示形式（P1–P6）的组合得分，以及平均值（Avg）和标准差（SD）。粗体字标记每列的最高分。

察到 EUROLLM-1.7B-INSTRUCT 在瑞典语释义场景中优于所有三个 GPT-SW3 模型，尽管前者模型中瑞典语的数据量有限。这表明，在预训练中使用的目标语言数据量并不是影响模型在除英语以外语言的幻觉检测性能的唯一因素。

增加模型参数规模是否会导致更好的性能？ 为了回答这个问题，我们对比了属于同一系列的不同参数数量的模型性能。对于 Llama3 模型，我们观察到模型大小有明显的影响，较大的 META-LLAMA-3-70B-INSTRUCT 模型通常以较大的优势优于较小的 META-LLAMA-3-8B-INSTRUCT 模型。对于 GPT-SW3，我们也看到了相同的模式，但仅限于瑞典语，在这种情况下 GPT-SW3-40B 始终优于较小的 GPT-SW3-20B 模型。对于 Mixtral 模型，观察到相反的趋势：将模型大小从 8x7b 增加到 8x22b 在所有场景中都会导致性能下降。

对于 Llama3 家族，我们观察到在两种情境和所有语言中，使用经过指令微调的变体与基础版本相比，表现有明显提升。相反地，对于 GPT-SW3，基础版本在两种释义情境中始终优于经过指令微调的变体。这可能是由于用于训练的指令微调语料库中缺少 NLI 示例。EUROLLM-1.7B 的指令微调变体在瑞典语释义和法语至英语翻译中表现更好，而在英语释义和德语至英语翻译中则相反。这可能归因于模型的容量有限，限制了其充分整合指令微调数据的能力。总体而言，我们没有找到确凿的证据证明指令微调提高了性能，因为不同模型家族在不同指令微调数据集上训练的结果有所不同。

提示的语言和表述重要吗？ 我们研究了非英语提示对瑞典语意译和 fr ⇒ en 以及 de ⇒ en 翻译的影响。正如表格 2 和 3 中提示语言的平均模型性能差异所示，提示语言的选择很重要，英语是总体上表现最好的提示语言。这并不令人意外，因为研究中的所有模型都可能在大量英语数据上进行过训练。一个例外是瑞典语意译，其中 GPT-SW3-20B-INSTRUCT 在使用瑞典语提示时表现最佳。对于 fr ⇒ en 翻译，META-LLAMA-3-70B-INSTRUCT 在法语提

Translation en ⇒ fr								
BGE-M3-ZEROSHOT-V2.0	0.82							
LLM	PLg	P1	P2	P3	P4	P5	P6	Avg ± SD
META-LLAMA-3-8B-INSTRUCT	en	0.74	0.77	0.66	0.71	0.83	0.73	0.74 ± 0.06
META-LLAMA-3-70B-INSTRUCT	en	0.85	0.89	0.81	0.88	0.86	0.90	0.87 ± 0.03
META-LLAMA-3-70B	en	0.69	0.73	0.70	0.74	0.49	0.74	0.68 ± 0.10
MIXTRAL-8X7B-INSTRUCT	en	0.81	0.86	0.85	0.78	0.83	0.80	0.82 ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.41	0.68	0.57	0.44	0.74	0.45	0.56 ± 0.15
EUROLLM-1.7B-INSTRUCT	en	0.34	0.44	0.49	0.40	0.60	0.49	0.46 ± 0.09
EUROLLM-1.7B	en	0.44	0.42	0.44	0.43	0.23	0.43	0.40 ± 0.08
Translation fr ⇒ en								
BGE-M3-ZEROSHOT-V2.0	0.88							
LLM	PLg	P1	P2	P3	P4	P5	P6	Avg ± SD
META-LLAMA-3-8B-INSTRUCT	en	0.62	0.63	0.53	0.60	0.73	0.57	0.61 ± 0.07
	fr	0.33	0.40	0.30	0.43	0.80	0.73	0.50 ± 0.21
META-LLAMA-3-70B-INSTRUCT	en	0.67	0.80	0.53	0.84	0.81	0.78	0.74 ± 0.12
	fr	0.80	0.80	0.73	0.84	0.81	0.80	0.80 ± 0.04
META-LLAMA-3-70B	en	0.63	0.70	0.58	0.68	0.61	0.66	0.64 ± 0.05
	fr	0.50	0.62	0.41	0.41	0.51	0.75	0.53 ± 0.13
MIXTRAL-8X7B-INSTRUCT	en	0.80	0.82	0.78	0.83	0.81	0.81	0.80 ± 0.02
	fr	0.81	0.77	0.85	0.78	0.80	0.78	0.80 ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.39	0.56	0.46	0.56	0.72	0.41	0.53 ± 0.15
	fr	0.07	0.26	0.05	0.13	0.53	0.34	0.24 ± 0.20
EUROLLM-1.7B-INSTRUCT	en	0.40	0.52	0.46	0.40	0.38	0.51	0.45 ± 0.06
	fr	0.35	0.36	0.32	0.34	0.31	0.35	0.34 ± 0.01
EUROLLM-1.7B	en	0.35	0.35	0.35	0.36	0.31	0.35	0.35 ± 0.02
	fr	0.35	0.34	0.35	0.34	0.31	0.34	0.34 ± 0.01
Translation en ⇒ de								
BGE-M3-ZEROSHOT-V2.0	0.73							
LLM	PLg	P1	P2	P3	P4	P5	P6	Avg ± SD
META-LLAMA-3-8B-INSTRUCT	en	0.56	0.62	0.48	0.57	0.79	0.60	0.60 ± 0.10
META-LLAMA-3-70B-INSTRUCT	en	0.69	0.87	0.68	0.75	0.83	0.85	0.78 ± 0.08
META-LLAMA-3-70B	en	0.65	0.70	0.61	0.65	0.54	0.81	0.66 ± 0.09
MIXTRAL-8X7B-INSTRUCT	en	0.82	0.79	0.78	0.75	0.84	0.79	0.79 ± 0.03
MIXTRAL-8X22B-INSTRUCT	en	0.49	0.75	0.64	0.57	0.81	0.59	0.65 ± 0.14
EUROLLM-1.7B-INSTRUCT	en	0.33	0.45	0.40	0.41	0.53	0.46	0.43 ± 0.07
EUROLLM-1.7B	en	0.42	0.41	0.42	0.42	0.24	0.42	0.39 ± 0.07
Translation de ⇒ en								
BGE-M3-ZEROSHOT-V2.0	0.78							
LLM	PLg	P1	P2	P3	P4	P5	P6	Avg ± SD
META-LLAMA-3-8B-INSTRUCT	en	0.56	0.58	0.46	0.52	0.79	0.47	0.57 ± 0.12
	de	0.41	0.36	0.19	0.48	0.80	0.67	0.49 ± 0.22
META-LLAMA-3-70B-INSTRUCT	en	0.66	0.85	0.60	0.82	0.81	0.85	0.77 ± 0.11
	de	0.53	0.87	0.20	0.86	0.83	0.83	0.69 ± 0.27
META-LLAMA-3-70B	en	0.56	0.57	0.50	0.55	0.67	0.60	0.58 ± 0.06
	de	0.34	0.72	0.30	0.38	0.67	0.56	0.49 ± 0.18
MIXTRAL-8X7B-INSTRUCT	en	0.75	0.82	0.85	0.85	0.81	0.84	0.82 ± 0.04
	de	0.81	0.80	0.81	0.77	0.84	0.62	0.77 ± 0.08
MIXTRAL-8X22B-INSTRUCT	en	0.43	0.58	0.42	0.56	0.79	0.37	0.54 ± 0.18
	de	0.18	0.38	0.33	0.19	0.76	0.57	0.41 ± 0.24
EUROLLM-1.7B-INSTRUCT	en	0.22	0.23	0.21	0.22	0.46	0.21	0.26 ± 0.10
	de	0.20	0.22	0.22	0.22	0.45	0.22	0.26 ± 0.10
EUROLLM-1.7B	en	0.45	0.39	0.41	0.48	0.30	0.47	0.42 ± 0.07
	de	0.24	0.22	0.28	0.25	0.46	0.22	0.28 ± 0.09

Table 3: 所有语言对中的翻译场景测试集结果：F1 分数。基线模型有一个单一分数。对于所有其他模型，我们报告不同提示语言（PLg）和提示表述（P1–P6）的组合得分，以及（Avg）和标准偏差（SD）。粗体代表每列的最高分。

示时表现最佳。我们现在研究当考虑表格 2 和 3 中的标准差值时，单个模型性能是否会因提示选择而变化。总体来说，性能在提示变化中保持稳定，但某些情况尤为突出：MIXTRAL-8X22B-INSTRUCT 在所有情境中明显不稳定，提示 5（没有提到“幻觉”，而是使用“支持”代替“矛盾”）始终表现最佳。这在某种程度上也适用于 META-LLAMA-3-8B-INSTRUCT 。此外，与忽略“幻觉”提及的提示（提示 4-6）相比，提到“幻觉”的提示（提示 1-3）往往对 MIXTRAL-8X22B-INSTRUCT 和一些 Meta-Llama3 模型的性能有负面影响。

6.2 误差分析

我们检查了每个模型在不同幻觉类别中的错误率，并重点强调了由于模型产生错误标签而导致的错误比例。这些结果在所有提示中进行了平均，并在附录 I 中进行了详细展示。我们发现，错误率似乎在不同的幻觉类别中波动，但没有明显或可察觉的模式。我们还发现，高

错误率可能是由于某些模型产生的无效输出（即，不是 *hyp*1 也不是 *hyp*2，也不属于与任一标签相对应的同义词）导致的。我们主要在 MIXTRAL-8X22B-INSTRUCT 中注意到了这一点，而在 GPT-SW3-20B-INSTRUCT、MIXTRAL-8X7B 以及两个相当小的 EuroLLM 变体中程度较轻。

值得注意的是，Mixtral 系列倾向于生成输出，声称两个假设都是幻觉或者都不是。同样地，GPT-SW3 模型展示出为每个实例返回几乎相同的短语或标签的习惯。例如，GPT-SW3-20B 倾向于为几乎每对句子检测到 *hyp*1 标签，而经过指令调优的变体由于无效输出而具有较高的错误率，因为对于某些提示，它往往几乎总是输出一个短语，表示其无法执行任务（例如，“很难在没有更多上下文的情况下进行判断”[3]）。目前尚不清楚为什么这个模型趋向于生成几乎相同的输出，尽管这可能与指令调优期间使用的数据类型有关。另一方面，EuroLLM 模型的无效输出发生在模型开始生成翻译或改写源句子的句子，而不是执行当前检测任务时，尽管考虑到其小规模，这并不令人惊讶。

值得注意的是，NLI 模型的标签是由蕴含概率决定的，因此在其设置中无法生成无效标签，这与 LLM 不同。

7 结论

我们进行了多项实验，以研究开放访问的 LLM 在检测幻觉中的能力，该能力在 HalluciGen 任务 (Karlgren et al., 2024; Dürlich et al., 2024) 中有所定义。最强的模型 MIXTRAL-8X7B-INSTRUCT 和 META-LLAMA-3-70B-INSTRUCT 在所有语言和场景中表现始终良好，表明 LLM 适合该任务。体积小得多的 NLI 模型的良好表现表明基于 LLM 的检测器并不是唯一可行的选择。

我们分析了四个不同因素的影响：目标语言、模型大小、指令调整和提示——结果发现它们都不能作为预测此任务上模型性能的直接指标。我们的对照实验表明：**(i)** models perform consistently across languages, with a slight advantage for English; **(ii)** the impact of model size differs between model families; **(iii)** instruction-tuning has a clear positive effect only for the largest model ; **(iv)** English prompts generally lead to the best overall performance, while the inclusion of the term “hallucination” in the prompt has a partially negative impact; and **(v)** for some models, a high error rate can be traced to the proportion of invalid outputs produced during inference.

在未来的工作中，我们计划探索是否可以使用大型语言模型来生成数据集，以用于训练和评估幻觉探测器，并将这些应用于跨模型评估环境。

8 局限性

由于可用的大语言模型（LLM）集合非常庞大且不断扩展，以及提示它们的方式多种多样，进行详尽的提示探索实验是不切实际的。同样地，探索第 **??** 节中描述的生成参数的所有可能值也是不现实的；尽管我们选择的值应该是普遍适用的，但我们并没有针对个别模型进行优化。尽管如此，我们希望我们的工作能够提供有关 LLM 作为幻觉检测器适用性的见解，就像它们在幻觉检测任务中的表现所表明的那样。

当评论目标语言在模型预训练数据中的存在或指令微调中包含的任务时，我们依赖于模型开发者以学术论文、报告和博客文章形式提供的信息。对于 EuroLLM 和 GPT-SW3 模型，这些方面记录得较为详细，而在其他模型（例如 Llama3 和 Mixtral）的情况下，这些信息可能不完整或缺失。在未提供此类信息的情况下，自然很难得出有关不同因素对模型在任何下游任务中的性能影响的结论。

此外，幻觉类别标签存在两个主要限制：(a) 它们受到类别不平衡的影响；(b) 它们没有考虑到某些样本可能属于多个类别。

我们的数据集仅关注于一小部分高资源语言：英语和瑞典语用于释义，以及英法和英德配对用于翻译。此外，一些幻想例子是手动构建的，可能不准确地反映现实世界中固有的幻想。未来的工作应着眼于减少数据集的英语中心性质，并扩展任务以包括一系列高、中、低资源语言，并专注于自然发生的固有幻想。

[3] 瑞典语原文：“Det är svårt att säga utan mer sammanhang”

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513. Association for Computational Linguistics.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi.

2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Luise Dürlich, Evangelia Gogoulou, Liane Guillou, Joakim Nivre, and Shorouq Zahra. 2024. Overview of the clef-2024 eloquent lab: Task 2 on hallucigen. In *25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. Grenoble. 9 September 2024 through 12 September 2024*, volume 3740, pages 691–702. CEUR-WS.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 288–298, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Haoqiang Kang, Terra Blevins, and Luke S. Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *ArXiv*, abs/2402.10496.

Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Liane Guillou, Joakim Nivre, Magnus Sahlgren, Aarne Talman, and Shorouq Zahra. 2024. Overview ofäeloquent 2024—shared tasks forăevaluating generative language model quality. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 53–72, Cham. Springer Nature Switzerland.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Dan Saattrup Nielsen. 2022. Scandinli: Natural language inference for the scandinavian languages. `https://github.com/alexandrainst/ScandiNLI`.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Preprint*, arXiv:2309.09558.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

A 幻觉示例

表 4 给出了针对每个幻觉类别的意译情境中幻觉假设的示例。

Type	Source	Hallucination
Addition	We struggle with water on a daily basis in the Netherlands - in the polders, the delta where the Meuse, the Rhine and the Scheldt flow into the sea.	In the Netherlands, we struggle with water on a daily basis because of the Meuse, Rhine, Scheldt, Noord, Voer and Dieze
Named-Entity	The fact is that a key omission from the proposals on agricultural policy in Agenda 2000 is a chapter on renewable energy.	Agenda 2030 does not include a chapter on renewable energy.
Number	The European Commission proposes that this information should enter into force within a period of three years from 1 July 1998.	The EU wants this information to enter into force in thirty years.
Conversion	In addition to these losses, there were also significant losses in terms of infrastructures, totalling approximately EUR 15 million.	There were losses in the amount of approximately 15 million dollars.
Date	In 1998, 1 700 000 net jobs were created in Europe, and although I admit that the employment situation is far from ideal, it has improved.	In 1700 there were 1 998 000 net jobs created in Europe.
Gender	Madam President, I am speaking on behalf of our colleague, Mr Francis Decourrière, who drafted one of the motions for a resolution.	One of the motions for a resolution was drafted by Mrs Francis Decourrière.
Pronoun	We have done so: on 5 February we published an extremely detailed press release dealing with the questions you have raised.	We published a press release that dealt with the questions we raised.
Antonym	The population has declined in some 210 of the 280 municipalities in Sweden, mainly in inland central and northern Sweden.	In the majority of Sweden's 280 municipalities, the population has gone up.
Tense	For the latter, the initial birth of several operators is now giving way to the reconcentration of the sector in the hands of a single company.	Several operators have given way to the reconcentration of the sector in the hands of one company.
Negation	The draft agenda as drawn up by the Conference of Presidents pursuant to Rule 95 of the Rules of Procedure has been distributed.	The Conference of Presidents hasn't distributed the draft agenda.
Natural	Amendment No 1 in the French version deletes illegal immigration and Amendment No 4 omits the expression 'police authorities'.	The French version excludes the expression'police authorities'.

Table 4: 转述任务幻觉类别的示例。

B 幻觉统计

表 5 显示了在改写和机器翻译幻觉检测场景中，每种语言或语言对的每种幻觉类别的频率。这个数据首先由 (Dürlich et al., 2024) 报告。

Language	Scenario	Addition	Antonym	Date	Gender	Named Entity	Negation	Number	Pronoun	Tense	Conversion	Natural
en	PG	11	16	5	3	9	14	9	11	4	3	33
sv		42	11	–	3	15	12	9	1	5	1	20
en-fr	MT	10	–	24	–	33	–	33	–	–	–	–
fr-en		9	13	4	12	12	12	13	–	12	13	–
en-de		10	16	14	–	15	–	13	16	–	–	16
de-en		10	10	7	11	10	10	10	–	10	11	11

Table 5: 每个幻觉类别在不同场景和语言中的频率统计。

C 非英语提示

表 6 展示了所有非英语的提示。

Prompt Name	Prompt
	Swedish paraphrase - Swedish prompt
Prompt 1	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en hallucination av den ursprungliga meningen. En hallucination innebär att hypotesen inte logiskt stöds av källan.
Prompt 2	Du är en AI-granskare specialiserad på parafrasdetektion. Din uppgift är följande: Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en hallucination av den ursprungliga meningen. En hallucination innebär att hypotesen inte logiskt stöds av källan.
Prompt 3	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en hallucination av den ursprungliga meningen.
Prompt 4	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som motsäger den ursprungliga meningen.
Prompt 5	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som stöder den ursprungliga meningen.
Prompt 6	Givet en mening (src) och två parafrasförslag (hyp1 och hyp2), avgör vilken av de två som är en dålig parafras av den ursprungliga meningen.
	Translation fr-en - French prompt
Prompt 1	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses est une hallucination de src. Une hallucination se manifeste quand l'original ne confirme pas logiquement l'hypothèse.
Prompt 2	Tu es un réviseur de traduction automatique IA. Ta tâche est la suivante: Tu reçois une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2). Décide laquelle des hypothèses est une hallucination de src. Une hallucination se manifeste quand l'original ne confirme pas logiquement l'hypothèse.
Prompt 3	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses est une hallucination de src.
Prompt 4	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses contredit src.
Prompt 5	Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses confirme src.
Prompt 6	Tu es un réviseur IA avec une spécialisation en traduction automatique. Étant donné une phrase dans la langue originale (src) et deux hypothèses de traduction de src (hyp1 et hyp2), décide laquelle des hypothèses est une mauvaise traduction de src.
	Translation de-en - German prompt
Prompt 1	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen halluziniert ist. Eine Halluzination tritt auf, wenn die Hypothese das Original (src) nicht logisch unterstützt.
Prompt 2	Du bist ein KI-Prüfer für maschinelle Übersetzung. Deine Aufgabe ist die folgende: Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen halluziniert ist. Eine Halluzination tritt auf, wenn die Hypothese das Original (src) nicht logisch unterstützt.
Prompt 3	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen halluziniert ist.
Prompt 4	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen src widerspricht.
Prompt 5	Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen src unterstützt.
Prompt 6	Du bist ein KI-Prüfer mit Fachkenntnissen in maschineller Übersetzung. Bestimme anhand eines Ausgangssatzes (src) und zweier Übersetzungsvorschläge für src (hyp1 und hyp2), welche dieser zwei Hypothesen eine schlechte Übersetzung von src ist.

Table 6: 在瑞典语、法语和德语中测试的提示形式。

D 标签后处理

测试的模型通常直接返回两个预期标签中的一个（$hyp1$ 或 $hyp2$），但有些模型倾向于以不同的措辞返回标签。出于这个原因，我们首先检查生成的模型输出是否包含这些变体中的任何一个：

- “1” 或 “2”
- “hyp 1” 或 “hyp 2”（包括空白）
- “hypotes 1” 或 “hypotes 2”
- “假设 1” 或 “假设 2”
- “假设 1” 或 “假设 2”

如果模型输出仅包含一个标签（无论是哪种变化形式），我们将其提取为标签。如果生成的输出包含两个标签，我们认为输出无效并返回空标签。如果上述变化形式均不存在，我们则扩展变化形式列表以涵盖用于提示模型的不同语言：

正如在第 **??** 节所解释的那样，Prompt 5 的设计方式是将任务反转；我们提示模型输出支持来源的假设的标签。因此，仅对于这个特定的提示，标签会从 $hyp1$ 翻转为 $hyp2$ ，反之亦然，除非模型生成一个空标签（在这种情况下，标签保持不变）。

E 模型仓库

Family	Variant	Repository	Version
Llama-3	META-LLAMA-3-8B-INSTRUCT	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct	3.0
	META-LLAMA-3-70B-INSTRUCT	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct	3.0
	META-LLAMA-3-70B	https://huggingface.co/meta-llama/Meta-Llama-3-70B	3.0
Mixtral	MIXTRAL-8X7B-INSTRUCT	mistralai/Mixtral-8x7B-Instruct-v0.1	v0.1
	MIXTRAL-8X22B-INSTRUCT	https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1	v0.1
EuroLLM	EUROLLM-1.7B	https://huggingface.co/utter-project/EuroLLM-1.7B	-
	EUROLLM-1.7B-INSTRUCT	https://huggingface.co/utter-project/EuroLLM-1.7B-Instruct	-
GPT-SW3	GPT-SW3-20B-INSTRUCT	https://huggingface.co/AI-Sweden-Models/gpt-sw3-20b-instruct	-
	GPT-SW3-20B	https://huggingface.co/AI-Sweden-Models/gpt-sw3-20b	-
	GPT-SW3-40B	https://huggingface.co/AI-Sweden-Models/gpt-sw3-40b	-

F NLI 基线详情

为了确定两个假设（$hyp1$, $hyp2$ ）中的哪个包含幻觉，我们预测源句和每个假设之间的“蕴涵”（E）和“非蕴涵”（NE）类别得分。然后，我们根据一个或多个假设选择幻觉。

- 如果一个假设的 E > NE ，另一个假设的 E < NE ，我们选择 E < NE 的那个。
- 如果对于两个假设，E > NE，那么我们选择得分最低的那个。
- 如果在两种假设中，E < NE，我们选择具有最高 NE 分数的假设。

G 计算环境和效率

实验是在一个配备有 NVidia A100 SXM6 64GB GPU 和单个 32 核 Intel Ice Lake CPU 的集群 [4] 上进行的。对于每个样本，使用 Huggingface 的 Accelerate 库进行模型推理，是按顺序执行的（换句话说，无需批处理）。[5] 表 7 列出了用于加载每个模型的 GPU 数量，以及对单个模型输入进行推理的执行时间。

Model name	Number of GPUs	Inference time per sample (sec)
META-LLAMA-3-8B-INSTRUCT	2	7.01
META-LLAMA-3-70B	4	11.44
META-LLAMA-3-70B-INSTRUCT	4	14.77
MIXTRAL-8X7B-INSTRUCT	2	15.13
MIXTRAL-8X22B-INSTRUCT	4	23.10
EUROLLM-1.7B	1	18.34
EUROLLM-1.7B-INSTRUCT	1	19.94
GPT-SW3-20B	1	14.45
GPT-SW3-20B-INSTRUCT	1	12.46
GPT-SW3-40B	3	13.02

Table 7: 用于加载每个模型的 GPU 数量，以及对一个输入进行推理的执行时间。

[4] 为保护匿名性，细节已被隐去。

[5] https://huggingface.co/docs/accelerate/v1.1.0/en/index

H 注释指南：释义幻觉

任务：你的任务是标记每个句子是幻觉（H）还是非幻觉（NH）。

本任务中的幻觉定义：在文本复述的背景下，给定一个源文本 src 和生成的假设 hyp，我们提出问题：hyp 是否由 src 支持？如果是，则标记 hyp 为非幻觉（NH）。如果否，则标记 hyp 为幻觉（H）。
当满足以下条件时，一个假设支持来源：

- 源文的整体语义被保留了，但一些细节缺失

一个假设不支持来源时：

- 添加了新的信息，即源中不存在且无法从源中推导出的信息
- 它包含无意义的信息（当源不包含时）
- 它误解了源文本中的语义关系（即，一个糟糕的释义）

示例：

Src	Stockholm is the capital of Sweden and is located on the East coast
Hyp (NH)	1) Stockholm, situated on the East coast, serves as the capital of Sweden 2) Stockholm is situated on the East coast
Hyp (H)	Stockholm is the capital of Denmark

释义数据的标注者是本文的作者，所有人都能流利地使用英语和/或瑞典语。

I 按幻觉类别划分的错误率

在图 1 和 2 中，我们提供了错误率（即错误标签的百分比），这些错误率是在所有提示变化中对模型进行平均，并按幻觉类别进行分割。阴影部分表示无效的错误标签（换句话说，它们不符合 $hyp1$ 或 $hyp2$ ）。

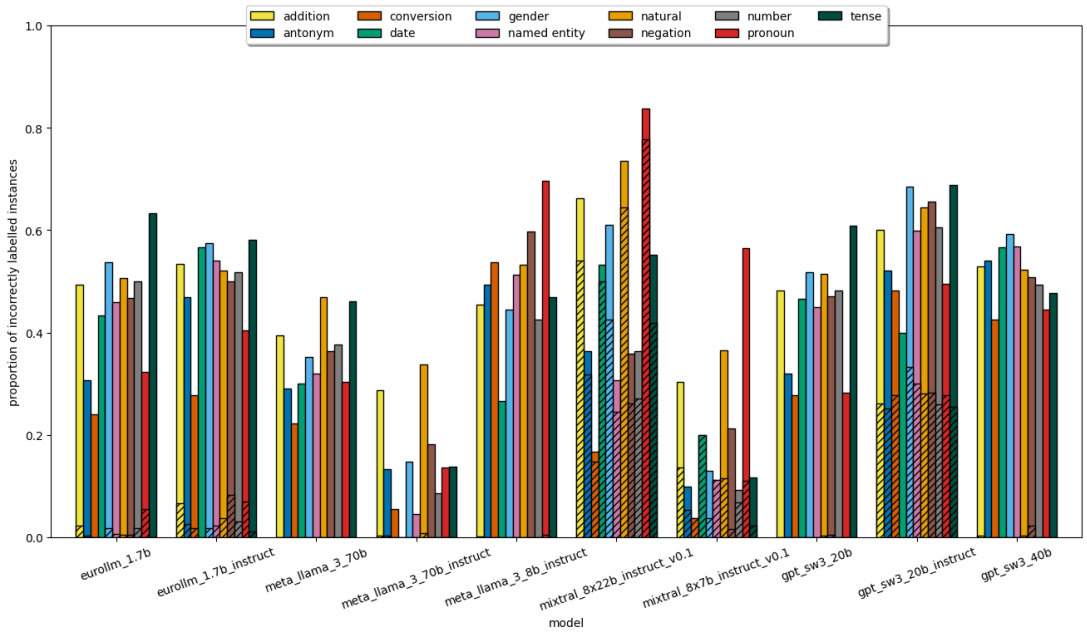

Figure 1: 错误标记的源假设释义对的平均比例（在所有提示和提示以及数据语言组合中取平均），按幻觉类别过滤。在这里，线条填充表示输出无效的比例（即，落在 { $hyp1$, $hyp2$ } 之外）。

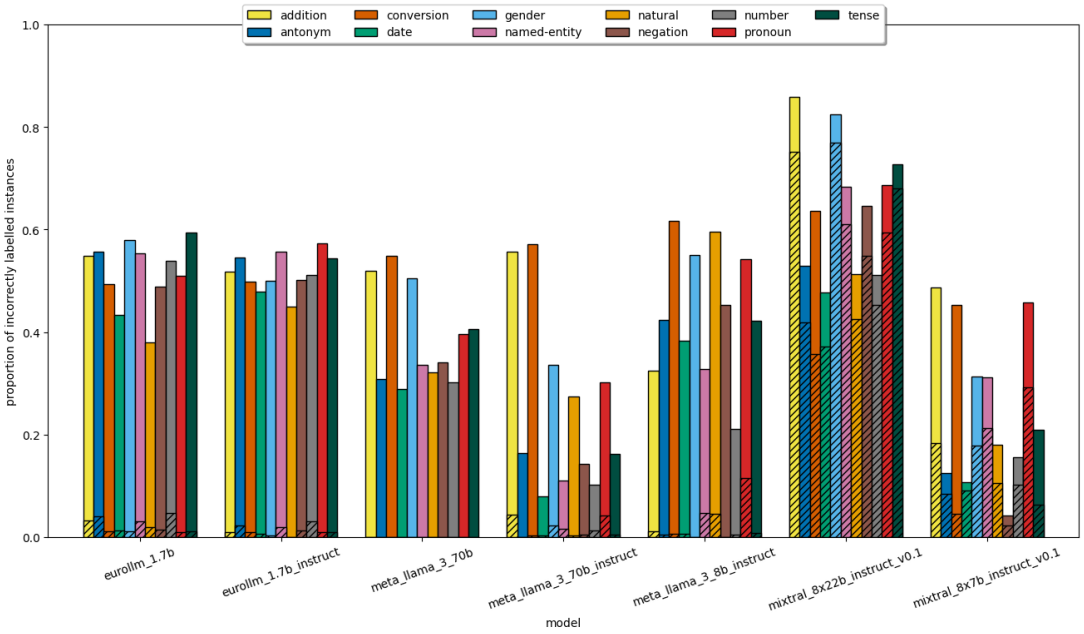

Figure 2: 被幻觉类别过滤的错误标记源假设翻译对的平均比例（在所有提示和提示数据语言组合中平均）。这里，阴影部分代表无效输出的比例（即，落在 $\{ hyp1, hyp2 \}$ 之外）。