

大模型 “自问自答” 研究



汇报人：肖力铭



汇报时间：20250403



目录

01

文献阅读报告

02

空间语义异常文本的生成与回答

03

SpaCE2025相关讨论

大模型 “自问自答”

Write a two paragraph story about 3 people building a house

[...] Finally, after years of saving money and sketching designs, they decided to build their dream house. Bob with his architectural expertise, sketched a breathtaking design of a two storied country house.

How many total designs were sketched in the story?

Expected answer: Many

The story only mentions one design being sketched [...]

Reverse Question Answering (RQA)

Generate a one-sentence question with the answer: "488".
The only possible answer to the question must be "488".
The question should not contain the text "488".

Question: What is the sum of the first eight prime numbers?

Question Answering (QA)

Generate the answer to the question: "What is the sum of the first eight prime numbers?". Give just the answer and no explanation.

Answer: 77

前八个质数的和

$$2+3+5+7+11+13+17+19 = 77$$

研究动机

意义：大模型的一致性（Consistency）检测

探讨：

1.Generation vs Comprehension

生成意味着理解吗？

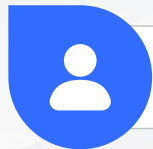
2.大模型的能力稳定吗？

...

THE GENERATIVE AI PARADOX:

“What It Can Create, It May Not Understand”

华盛顿大学，2023



汇报人：肖力铭



汇报时间：20250403



观点

生成式AI悖论：在生成类任务上表现出超越人类的能力，例如快速生成高质量的文本或图像；但在理解方面却不及人类。

核心观点：生成式AI的生成能力可能不依赖于理解能力，与人类智能形成对比。

实验一：不同形式

形式一：生成式任务
人类评分

形式二：判别式任务
自动评分

Austin had a rough day at work
and decided to go to the bar.
Austin had a drink that night.

Why did Austin do this?

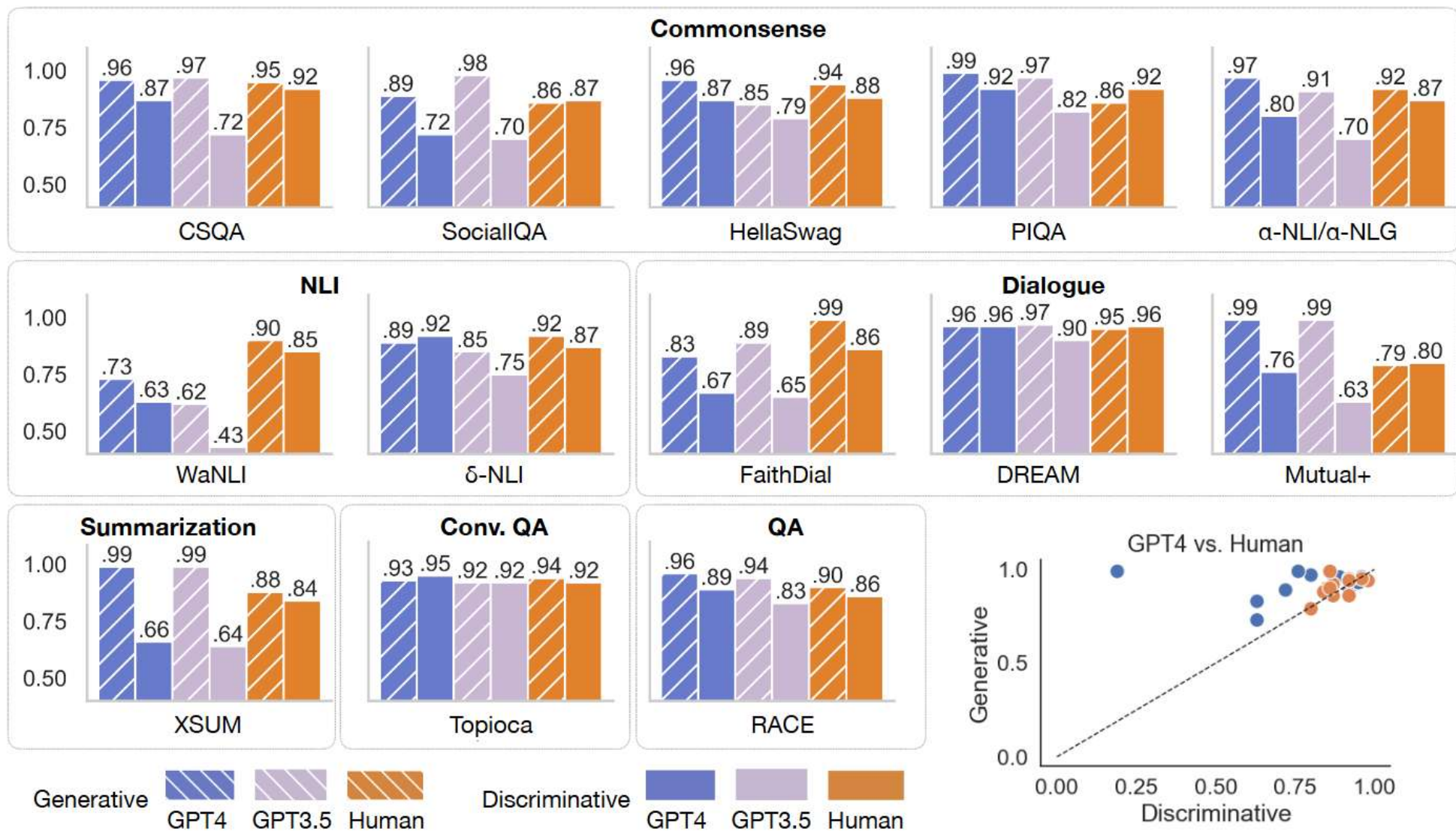
Austin likely went to the bar and had
a drink because he had a rough day
at work and wanted to *relax, unwind,*
or temporarily escape from the stress
and frustrations he experienced
during the day.

Select the best answer from the
choices provided:

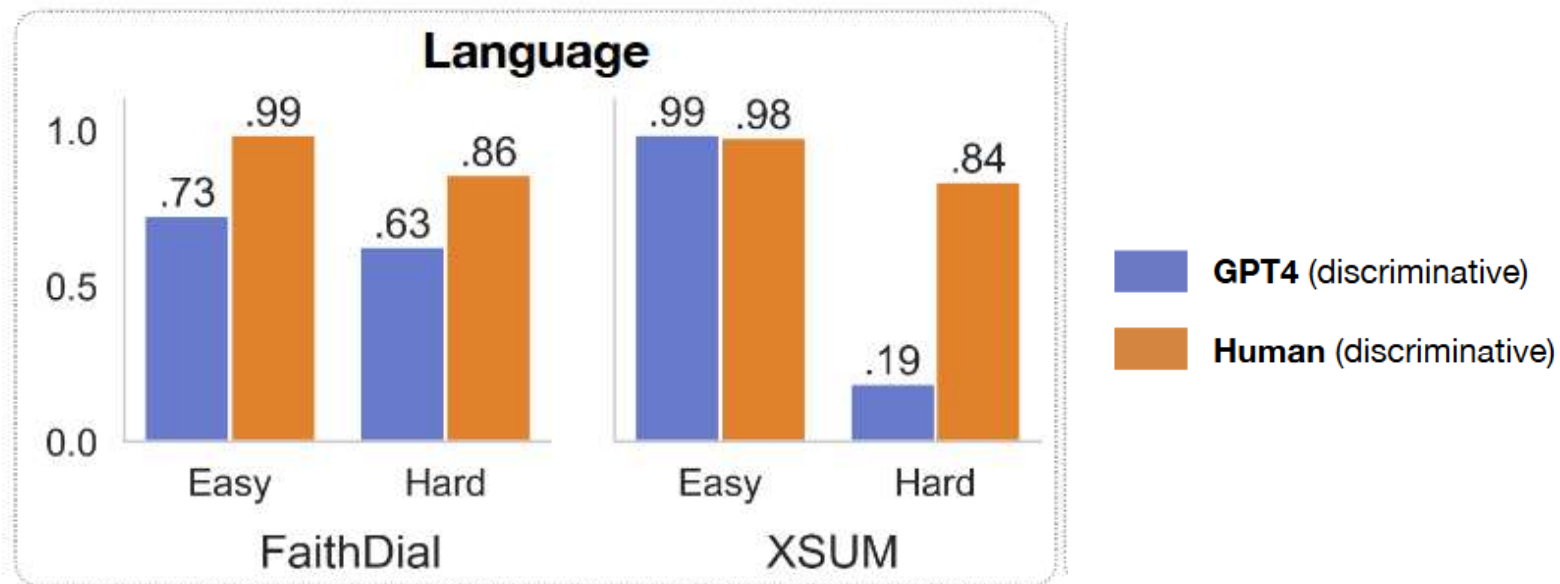
- A. Settle down** ★
- B. Go to the bar
- C. Order a drink

B. Go to the bar

实验一：结果

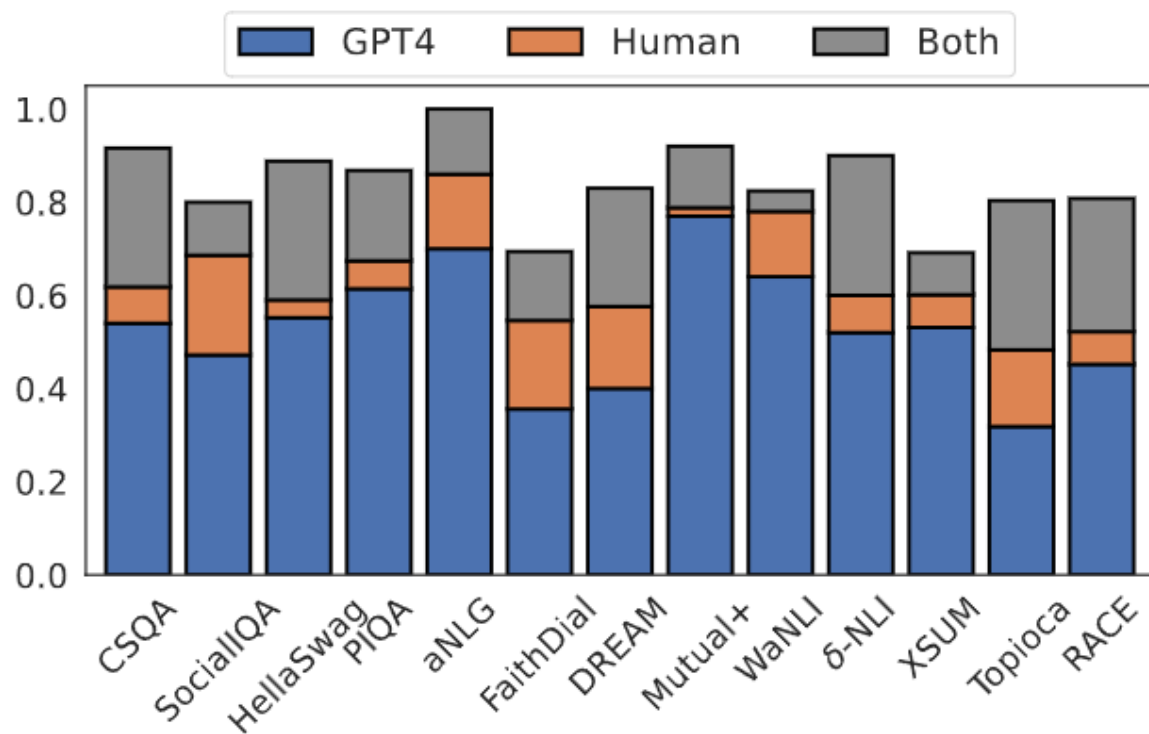


实验一：判别式任务的选项干扰结果



大模型在判别式任务上比人类更易受选项干扰。

实验一：机器和人类在生成式任务上的比较



由于大模型生成的文本更长、行文更正式、信息量更大，人类觉得大模型生成的文本更好。

实验二：自生自答

第一步：生成内容
人类判定合规与否

第二步：基于文本
做选择题
自动评分

Write a two paragraph story
about 3 people building a house

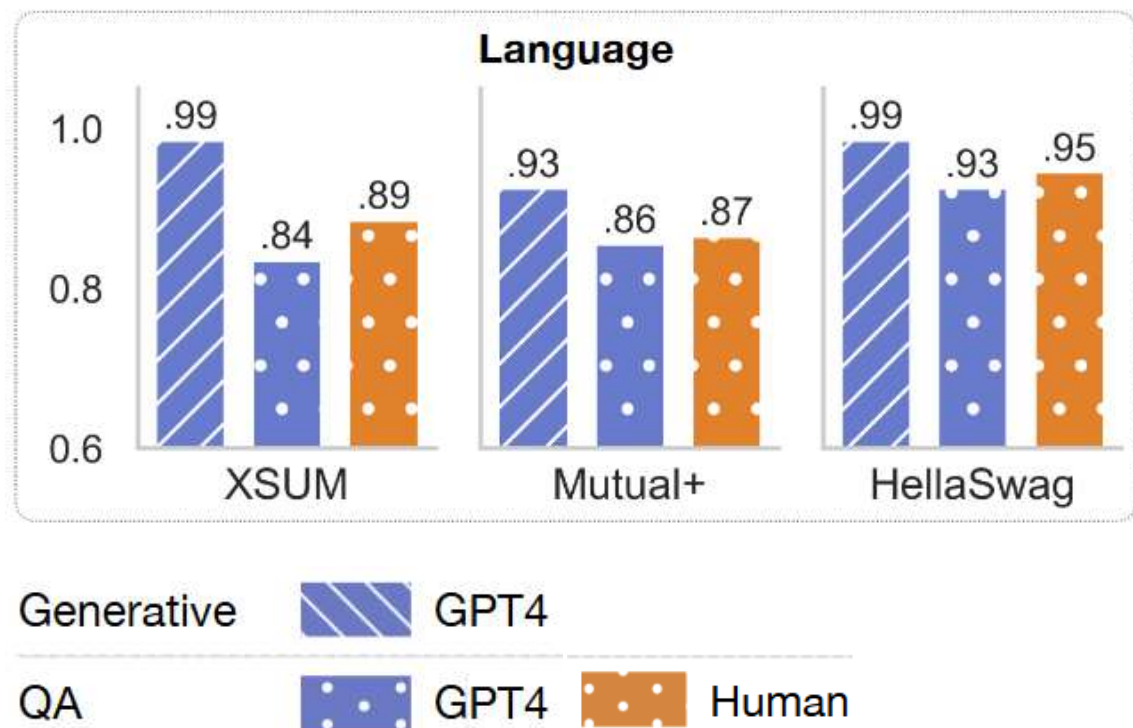
[...] Finally, *after years of saving money and sketching designs*, they decided to build their dream house. Bob with his architectural expertise, *sketched a breathtaking design* of a two storied country house.

How many total designs were
sketched in the story?

Expected answer: **Many**

The story only mentions
one design being sketched [...]

实验二：结果



XSUM (摘要)

- 1.要求模型基于摘要生成文本
- 2.要求模型基于文本选择摘要

Mutual+ (对话)

- 1.要求模型基于对话生成记录
- 2.要求模型基于记录选择最佳后文

HellaSwag (常识)

- 1.要求模型基于目标句生成前文
- 2.要求模型基于前文选择最佳后文

结论

相比GPT-4，人类在判别式任务与生成式任务的表现联系更紧密。

相比人类，模型的生成能力要强于理解能力。

我们需要谨慎地将AI的能力与人类智能进行类比，因为AI的“智能”可能遵循完全不同的逻辑。

生成式AI悖论提醒我们，研究AI不应仅仅是为了模仿人类智能，而是为了探索一种全新的智能形式。

比较

1.训练目标的差异

生成式AI：学习模式，模仿人类。

人类：理解细节，追求深度。

2.数据规模的差异

生成式AI：多且广。

人类：少且深。

3.驱动力的差异

生成式AI：生成能力是市场更关注的需求。

人类：注重理解和推理在进化过程中对生存和繁衍至关重要。

...

Reverse Question Answering:

Can an LLM Write a Question so Hard (or Bad)
that it Can't Answer ?

马里兰大学，2025



汇报人：肖力铭



汇报时间：20250403



QA & RQA

QA: For an input question q , QA deduces the correct answer a .

RQA: Given an answer a , generate a valid question q to which a is the answer.

任务思路

1.RQA 溯因推理

Reverse Question Answering (RQA)

Generate a one-sentence question with the answer: "488".
The only possible answer to the question must be "488".
The question should not contain the text "488".

Question: What is the sum of the first eight prime numbers?

前八个质数的和

$$2+3+5+7+11+13+17+19 = 77$$

2.QA 演绎推理

Question Answering (QA)

Generate the answer to the question: "What is the sum of the first eight prime numbers?". Give just the answer and no explanation.

Answer: 77

实验一：独立测试

Answer Type	Description	Example Question	Example Answer	Count
(1) Number	Integers in [100, 1000)	What is 26 times 4?	104	900
(2) Number+Text	Integers with a text entity	When did Pope Hormisdas die?	523 AD	743
(3) Easy Fact	Well-known factual entity	Who is the artist that painted Starry Night?	Vincent van Gogh	900
(4) Hard Fact	Obscure factual entity	What is the final painting by Paolo Uccello?	The Hunt in the Forest	900

1.构建QA数据集

Number：规则生成

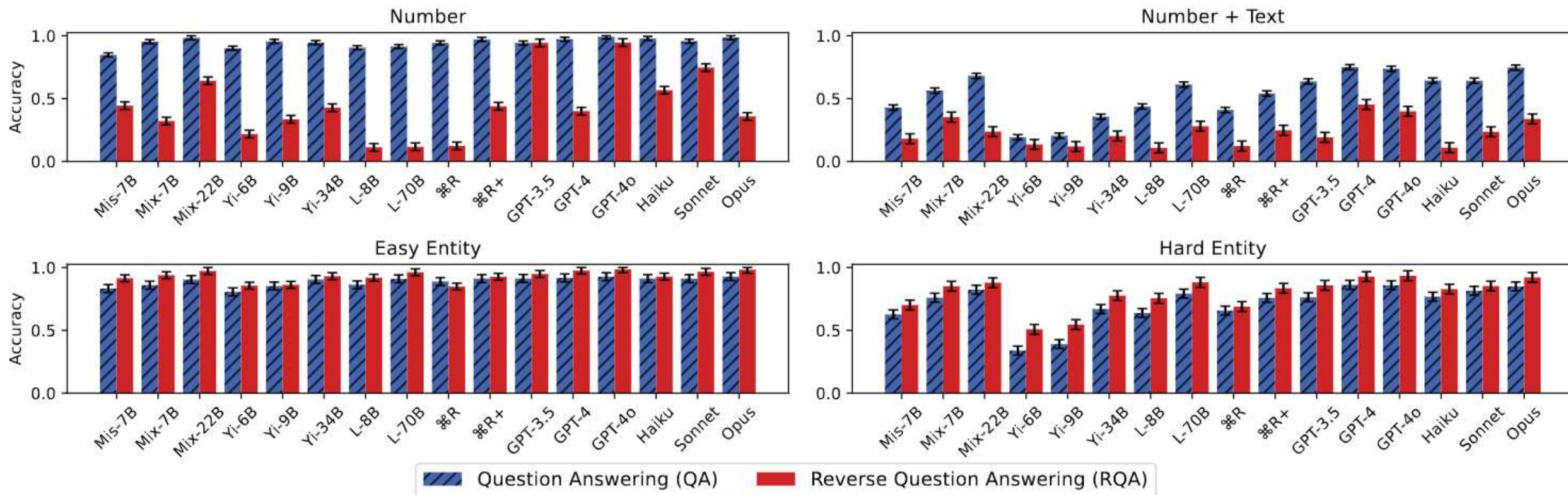
其他：提取自QANTA数据集

2.模型基于QA数据集的答案生成题目，得到RQA

This country defeated the invasion fleet of Liu Yan by inserting iron-tipped poles into a riverbed at the Battle of the Bach Dang River. Ma Yuan defeated an independence movement in this country led by the Trung sisters. This country's imperial capital was Hue [hway], and its last emperor was Bao Dai ["bough die"]. For 10 points--what country was reconstituted from former "north" and "south" portions in 1975?

Answer: Vietnam

实验一：独立测试



- 1.大模型在数字领域与事实领域的推理能力表现不一致。
- 2.大模型在数字领域的溯因推理能力远弱于演绎推理能力。

实验二：联合测试

Reverse Question Answering (RQA)

Generate a one-sentence question with the answer: "488".
The only possible answer to the question must be "488".
The question should not contain the text "488".

Question: What is the sum of the first eight prime numbers?

Question Answering (QA)

Generate the answer to the question: "What is the sum of the first eight prime numbers?". Give just the answer and no explanation.

Answer: 77

1. 输入 a , 输出 q'

a : 答案。RQA的输入

q' : 问题。RQA的输出

2. 输入 q' , 输出 a'

q' : 问题。QA的输入

a' : 回答。QA的输出

实验二：联合测试

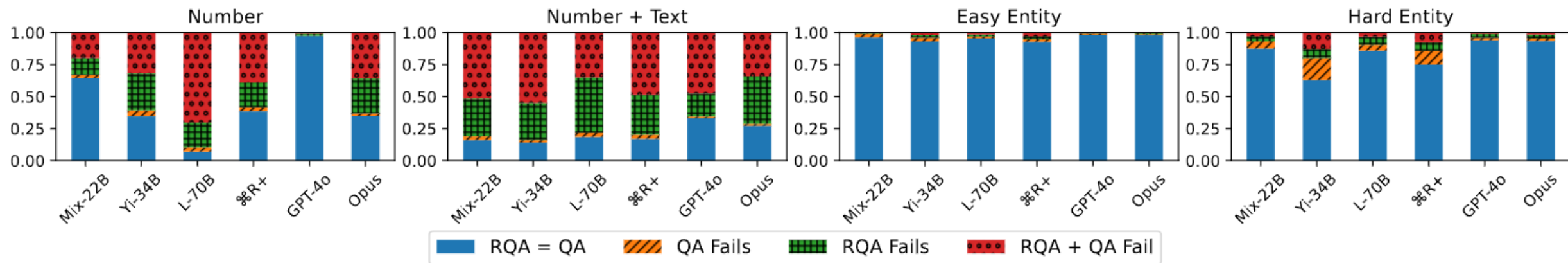
问题1：RQA是否成功？即 a 是否能回答 q'

问题2：QA是否成功？即 a' 是否能回答 q'

问题3： a 和 a' 是否等价？

问题1	问题2	问题3	结果
Y	Y	Y	RQA=QA
N	Y	N	Just RQA fails
Y	N	N	Just QA fails
N	N	N	RQA and QA fails
Y	Y	N	Ambiguous Question
N	N	Y	Lucky same mistakes
N	Y	Y	Metric Error
Y	N	Y	Metric Error

实验二：联合测试



除了GPT-4o，其他模型在数字领域的逻辑不一致性高。在数字领域，QA很少独自出错。文章将RQA错而QA对的现象归于大模型的幻觉问题。

实验二：联合测试

RQA更擅长出简单数学运算题
和考察事实的题目

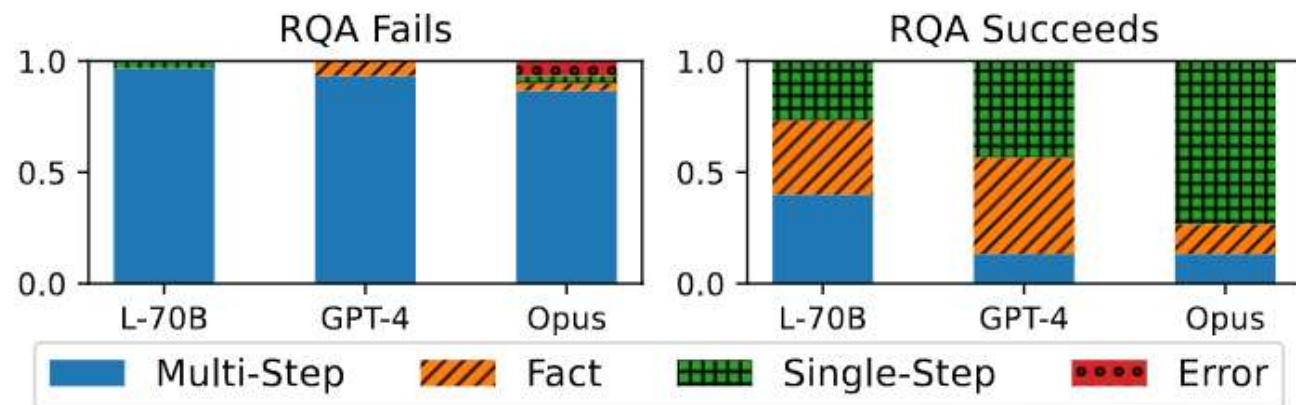
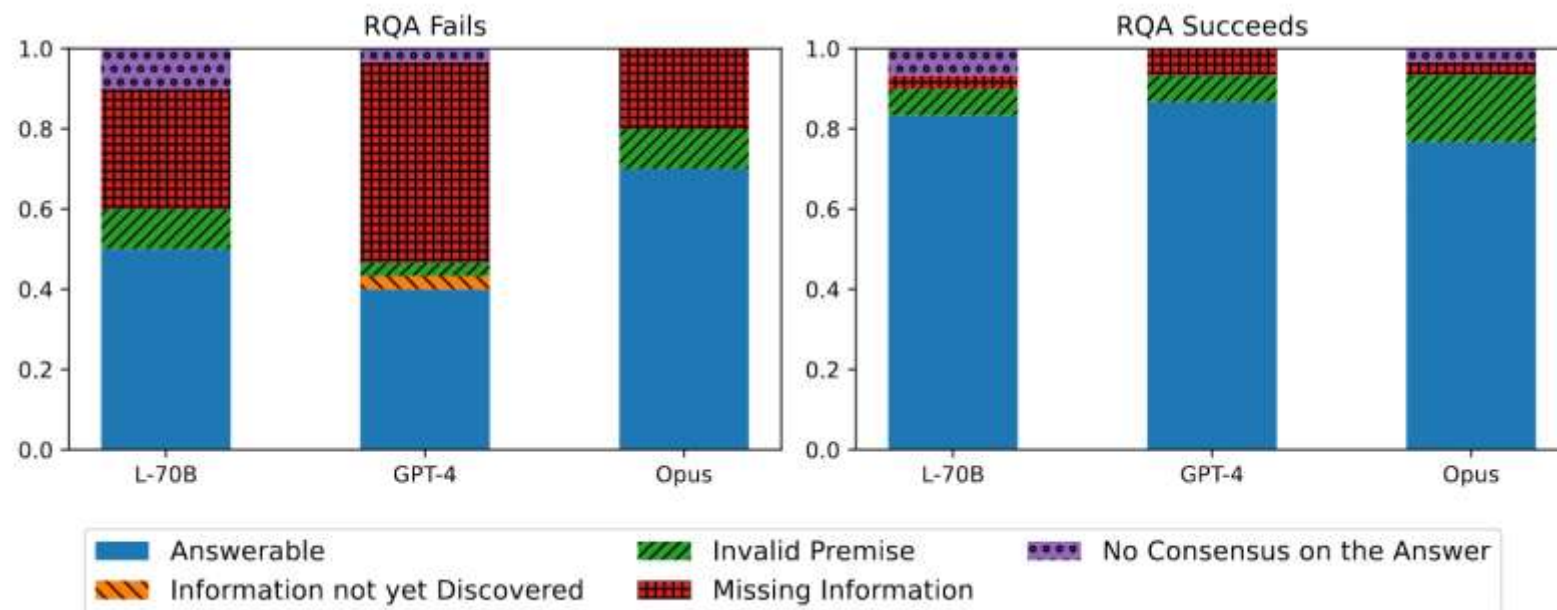


Figure 5: Analysis of Number RQA errors. RQA often fails when the LLM tries to give a complex, multi-step question.



模型出的题容易出现信息
不足而无法作答的情况

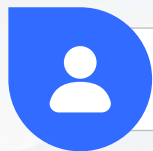
GPT-4根据输入“178人”输出：
How many individuals attended the
annual community festival last year
according to the final headcount?

实验二：联合测试

其他发现：

- 1.即使更换答案，大模型也会输出重复题。LLaMA-3 70b在数字领域输出了379道重复题（42.11%）。
- 2.同系列大模型之间的题目重复比例更高，可能因为共享训练集。
- 3.对于同一个答案，大模型出的题目与从数据集抽取的题目最高有40%的重复率（GPT-4）。

空间语义异常文本的生成 与正误判定



汇报人：肖力铭



汇报时间：20250403



实验准备

模型：deepseek-chat-V3

Prompt：3-shot

词汇：260个空间词

你是一个空间语义异常生成器。用户给定目标词汇；异常类型；文本长度。你生成满足要求的异常文本和异常解析。各种异常类型的示例如下：\n

- 词义冲突。目标词汇与相邻词汇的意思冲突。 \n

[输入示例]\n

目标词汇：上（方位词）\n

异常类型：词义冲突\n

文本长度：30字以内\n

[输出示例]\n

异常文本：海边有块礁石，涨潮时没入水上，退潮后露出海面。 \n

异常解析：“没入水上”存在词义冲突。“没入”指在水面之下，即水中，与“水上”的词义冲突。 \n\n

- 句义冲突。包含目标词汇的句子与上下文句子的意思冲突。

[输入示例]\n

目标词汇：出去\n

异常类型：句义冲突\n

文本长度：30字以上\n

[输出示例]\n

异常文本：这一天，他在街上闲逛，经过一家花店，从玻璃窗里望出去，隔着重重叠叠的花山，看见霓喜在里面买花。 \n

异常解析：根据上下文可知，主人公经过花店，在花店外面。但“从玻璃窗里望出去”表示视线从窗内望向窗外，说明主人公在花店内部。与上下文信息冲突。 \n\n

- 常识错误。包含目标词汇的句子违反常识。 \n

[输入示例]\n

目标词汇：前面\n

异常类型：常识错误\n

文本长度：50字以上\n

[输出示例]\n

异常文本：演出团34人分乘3辆小面包车行至亚当斯县境内的公路上时，第二辆车首先失去控制，造成翻车。前面一辆车的司机从后视镜中看到前面的车子出事，因紧急刹车也翻了车。 \n

异常解析：“从后视镜中看到前面的车子”存在常识错误。根据常识可知，后视镜是司机获取汽车后方和侧方信息的工具。对于正常行驶的车辆，司机无法通过后视镜看到前面的情况。

实验准备

人类对大模型生成的文本进行空间语义异常程度评价

异常程度	数量
异常	79
正常	121
不好判断	60

他站在悬崖边，望着脚下的深渊，突然向上坠落。 39

清晨的阳光洒在操场上，学生们整齐地排成队列，朝着主席台的方向行注目礼。校长站在主席台上，手持话筒，声音洪亮地宣布运动会正式开始。这时，站在队伍最后面的小明突然发现自己的鞋带松了，他蹲下身去系鞋带，却听到校长继续说道：“请全体同学向后转，朝着操场入口处行进。”小明慌忙站起身，却发现周围的同学都朝着主席台的方向走去，只有他一个人愣在原地，不知所措。 8

清晨，太阳从西边升起，金色的阳光洒满了整个山谷，鸟儿们开始欢快地歌唱。

32

实验准备

人类对大模型生成的文本进行空间语义异常程度评价

异常程度	数量
异常	79
正常	121
不好判断	60

空间语义理解能力不足：

他独自走在深夜的街道上，路灯将他的影子拉得很长。身后传来一阵急促的脚步声，他回头望去，却只看到空荡荡的马路。他继续向前走，心里有些不安，总觉得有人在跟踪他。突然，身后传来一声轻笑，他猛地转身，发现一个黑影正站在他面前，手里握着一把明晃晃的刀。

解析：“身后传来一声轻笑，他猛地转身，发现一个黑影正站在他面前”存在句义冲突。根据前文描述，“身后”指的是他背后的方向，而“站在他面前”则意味着黑影位于他的正前方。这两个方向是相反的，因此句子前后矛盾，无法同时成立。

实验准备

人类对大模型生成的文本进行空间语义异常程度评价

异常程度	数量
异常	79
正常	121
不好判断	60

异常强制扭转（转折或疑惑）：

他站在马路中央，看着两边的车辆川流不息。左边是呼啸而过的卡车，右边是缓慢行驶的自行车。突然，他意识到自己走错了方向，于是转身往回走。然而，两边的风景却丝毫未变，依旧是熟悉的建筑和人群。他感到困惑，明明已经改变了方向，为什么两边的景象还是完全一致？

解析：“两边的风景却丝毫未变”与上下文存在句义冲突。根据前文描述，主人公“转身往回走”，意味着他的行进方向已经改变。正常情况下，方向改变后，左右两侧的景物应随之变化（如原本的“左边卡车”应变为“右边卡车”）。但文中“两边的景象完全一致”与方向改变的动作矛盾，导致逻辑冲突。

实验准备

人类对大模型生成的文本进行空间语义异常程度评价

异常程度	数量
异常	79
正常	121
不好判断	60

文本信息不足：

电梯坏了，他只好从楼梯上去地下三层。

解析：“从楼梯上去地下三层”存在词义冲突。“上去”表示从低处到高处移动，而“地下三层”是更低的位置，与“上去”的词义冲突。

实验准备

人类对大模型生成的文本进行空间语义异常程度评价

异常程度	数量
异常	79
正常	121
不好判断	60

任务意图传递不到位：

他站在电梯门口，犹豫了一会儿，终于下定决心走了进去。电梯门缓缓关闭，他感到一阵轻微的失重感。透过电梯的玻璃墙，他看到外面的景色逐渐远去。突然，他意识到自己忘记带钥匙了，于是赶紧按下开门按钮，电梯门再次打开，他快步走出电梯，回到走廊上。

解析：“他站在电梯门口，犹豫了一会儿，终于下定决心走了进去”与后文“他快步走出电梯，回到走廊上”存在句义冲突。前文描述他“走进”电梯，并感受到电梯运行时的失重感，说明电梯已经启动上升或下降。然而后文却提到他“走出”电梯，回到走廊，仿佛电梯并未移动。这两部分信息矛盾，无法同时成立。

实验准备

人类对大模型生成的文本进行空间语义异常程度评价

异常程度	数量
异常	79
正常	121
不好判断	60

非空间语义异常：

那天晚上，我独自在家看书，突然听到外面传来一阵急促的敲门声。我放下书，走到门口，透过猫眼往外看，发现外面空无一人。正当我疑惑时，敲门声又响了起来，这次更加清晰。我打开门，外面站着一个浑身湿透的陌生人，他焦急地说：“我的车抛锚了，能借您的电话用一下吗？”我点点头，让他进来，却发现外面的雨已经停了，地上一点水渍都没有。

解析：“外面的雨已经停了，地上一点水渍都没有”与上下文存在句义冲突。前文提到陌生人“浑身湿透”，且敲门声急促，暗示外面正在下雨或刚下过雨。但后文却说“雨已经停了，地上一点水渍都没有”，这与“浑身湿透”的细节矛盾，因为如果雨停了且地面干燥，陌生人不可能被淋湿。

自问自答

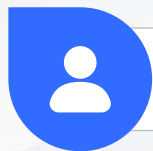
大模型对自己生成的文本进行空间语义异常判断

标签源	Accuracy
大模型自己	0.6769
人类	0.4615

YOUR LOGO

...

SpaCE2025讨论会



汇报人:



汇报时间:

