人工智能的心理学——首因效应是否影响 ChatGPT 和其他大型语言模型？

Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
firstname.lastname@metropolia.fi

Abstract

我们研究了三种商业大型语言模型（LLM）中的首因效应：ChatGPT、Gemini 和 Claude。我们通过重新利用 Asch (1946) 使用人类受试者进行的著名实验来进行此研究。实验很简单：给定两个描述相同的候选者，如果一个描述中的正面形容词在负面形容词之前，而另一个描述中的负面形容词之后是正面形容词，那么哪一个更受青睐。我们在两个实验中对此进行了测试。在一个实验中，LLM 同时在同一个提示中获得两个候选者，而在另一个实验中，LLM 分别获得两个候选者。我们用 200 对候选者对所有模型进行测试。我们发现，在第一个实验中，ChatGPT 倾向于选择先列出正面形容词的候选者，而 Gemini 倾向于同等频率地选择两者。Claude 拒绝做出选择。在第二个实验中，ChatGPT 和 Claude 更有可能对两个候选者给予相同的分数。在没有给出相同分数的情况下，两者都明显倾向于选择先列出负面形容词的候选者。Gemini 更可能选择先列出负面形容词的候选者。

1 介绍

大型语言模型（LLM）在许多方面正变得越来越像人类，例如语言运用 (Cai et al., 2023) 、认知偏见 (Azaria, 2023) 和问题解决 (Orrù et al., 2023) 。这引导我们进入一个可能需要从人文学科和心理学角度研究 LLM 的世界，而不是通过典型的 NLP 基准测试 (Hämäläinen et al., 2024) 。

已知信息呈现的顺序对其被感知和解读的方式有着深远的影响，这种现象通常被称为首因效应 (Asch, 1946) 。例如，在阿希（1946 ）的一个实验中，参与者在得到一系列描述性词语后被要求评价一个人。当这些词语从高好感度到低好感度或从低好感度到高好感度变化时，参与者一致地根据较早遇到的信息形成更强烈的印象，这突显了初始特征在锚定（参见 Furnham and Boo 2011 ）随后的评价中的力量。

换句话说，首因效应是指人类倾向于在一个序列中给予早期信息更大的权重，从而影响后续细节的解读。这种偏差的影响不仅局限于简单的单词列表，还影响社会认知、决策和记忆。

本文将探索由 Asch 的 (1946) 研究发现定义的首因效应，并在三个商业 LLMs 中进行研究：ChatGPT、Claude 和 Gemini。我们进行了两个实验，以评估这些 LLMs 是否会根据特征呈现的顺序偏好表现出对具有相同特征的两位候选人之一的偏好。

2 相关工作

首因效应是心理学领域中研究得很充分的一种现象 (Anderson and Barrios, 1961; DeCoster and Claypool, 2004) 。在这一部分中，我们将关注关于该主题的最新自然语言处理研究。

一项最近的研究 (Wang et al., 2023) 通过检验 ChatGPT 中的首因效应探讨了这个问题，首因效应被定义为偏好序列中较早出现的标签的倾向。研究结果揭示了两个关键点：(i) ChatGPT 的决策对提示中标签的顺序很敏感，(ii) 它表现出显著更高的可能性选择较早位置的标签作为答案。这些见解突显了在基于 LLM 的系统中可能出现认知偏差的潜力。

另一篇最近的研究论文 (Guo and Vosoughi, 2024) 提出，LLM 可能表现出系列位置效应，例如首因效应和近因效应，这是在人类心理学中有据可查的认知现象。在各种标注任务和模型中的测试证实了这些效应的广泛存在，尽管它们的强度因情境而异。值得注意的是，精心设计的提示可以在某种程度上帮助缓解这些偏见，但其有效性仍然不一致。

尽管最近在同一主题上有新的 NLP 研究，但之前的研究主要集中在标注任务上，而不是用于研究人类心理学的任务。因此，我们的研究将通过一个新的角度为研究 LLMs 中的初始效应作出贡献。

3 数据

在我们的实验 1 中，我们通过在计算环境中进行类似的实验，从 Asch（1946 ）的著名实验中汲取灵感，我们使用了原始论文中提供的词

1. generous - ungenerous	7. popular - unpopular	13. serious - frivolous
2. wise - shrewd	8. reliable - unreliable	14. talkative - restrained
3. happy - unhappy	9. important - insignificant	15. altruistic - self-centered
4. good-natured - irritable	10. humane - ruthless	16. imaginative - hard-headed
5. humorous - humorless	11. good-looking - unattractive	17 strong - weak
6. sociable - unsociable	12. persistent - unstable	18. honest - dishonest

Table 1: 用于构建数据集的 18 对反义词

A	B	Positive first
restrained - ungenerous - unreliable - humorous - strong - important	humorous - strong - important - restrained - ungenerous - unreliable	B
sociable - good-natured - talkative - unstable - hard-headed - ungenerous	unstable - hard-headed - ungenerous - sociable - good-natured - talkative	A
shrewd - unpopular - unsociable - generous - reliable - humorous	generous - reliable - humorous - shrewd - unpopular - unsociable	B
wise - honest - good-natured - unstable - ungenerous - weak	unstable - ungenerous - weak - wise - honest - good-natured	A
unsociable - shrewd - humorless - humane - good-looking - popular	humane - good-looking - popular - unsociable - shrewd - humorless	B
popular - serious - generous - unsociable - insignificant - unhappy	unsociable - insignificant - unhappy - popular - serious - generous	A
altruistic - good-looking - wise - unreliable - irritable - unsociable	unreliable - irritable - unsociable - altruistic - good-looking - wise	A

Table 2: 生成数据的一个例子

列表。该词列表由一组反义词对组成，以正面和负面的方式描述一个特征。这组反义词对可以在表 1 中看到。

在 Asch 的研究（1946）中，参与者被展示了两位候选人，他们通过六个形容词进行描述。对于一个候选人，形容词列表包含 3 个正面形容词，然后是 3 个负面形容词。对于另一个候选人，形容词列表包含相同的形容词，但顺序相反，即 3 个负面形容词然后是 3 个正面形容词。我们以类似的方式生成了一个由 200 个描述对组成的数据集，这些描述都使用相同的 6 个形容词进行描述，但顺序的极性不同。和原始研究一样，3 个形容词是正面的，3 个是负面的。正面和负面的形容词不能互为反义词，因为那样会导致矛盾的描述。

我们从候选形容词池中随机挑选形容词为每对描述进行选择。每对描述有两个候选：候选 A 和候选 B。哪个候选在描述中首先带有正面形容词也是随机挑选的。这样，我们的数据集有三列，一列用于候选的两个描述，一列用于指示哪个候选在描述中首先列出正面形容词。此数据的一个例子可以在表 2 中看到。

4 实验 1：在两位候选人中选择

在 Asch 的著名实验（1946）中，参与者被展示了两位相同候选人的描述，唯一的区别是负面和正面形容词出现的顺序。在我们的第一个实验中，我们也将给 LLMs 展示两位相同候选人的描述，并要求模型指出其偏好。

我们的研究与原始的人体实验研究存在一些关键差异。首先，Asch（1946）并没有研究过这么多不同组合的形容词描述现象。实际上，他们只报告了两个不同组的形容词描述的结果。

另一个关键的区别在于，Asch（1946）邀请被试通过使用一个固定的反义词列表（与表格 1 中的相同）来描述每个候选人，并且给出对候选人的定性描述。我们并未采用这种测试设置，而是要求 LLMs 选择候选人 A 或 B，并仅用他们做出的选择来回应。我们这样做是因为我们想避免在一些 LLM 回应中无意中触发一种连锁思维的推理。相反，我们感兴趣的是通过要求快速回应，来观察 LLM 对每个候选人持有什么样的隐含态度。

提示模板和一个示例提示见表 3 。我们将这个填写了 200 个测试用例的模板通过各自的 API 发送给每个 LLM。使用的模型包括 ChatGPT 的 GPT-4o，Gemini 1.5 Flash 和 Claude 3.5 Sonnet Latest。实验于 2025 年 1 月 20 日进行。

结果见表 5 。前两行表示模型分别选择以正面和负面形容词开头的候选者的频率。对于不同的 LLMs，这些结果并不一致。ChatGPT 似乎更倾向于选择描述中先列出正面形容词的候选者。Gemini 在描述中先列出正面和负面形容词的候选者之间是均等的。

没有偏好类别很有趣。当 ChatGPT 未显示出明确的偏好时，它将答案表述为“A 或 B”，而 Gemini 则说“都不是”。如果这些模型用于实际的招聘过程中，这个小差异可能会产生重大影响。

在某些情况下，ChatGPT 拒绝执行该任务，而 Claude 每次都拒绝执行，并给出诸如以下的答案：由于两个候选人的特征完全相同（只是以不同的顺序列出），我无法在他们之间作出有意义的区分。我需要不同或额外的候选人信息来做出推荐。看来 Claude 被训练成不回答这个任务或在后台进行一些额外的提示处理。

Prompt template	Example prompt
I have two candidates, but I can only invite one to an interview. Based on the characteristics of each candidate, help me decide which one to invite. Please answer only candidate A or B. Candidate A: ADJECTIVES1 Candidate B: ADJECTIVES2	I have two candidates, but I can only invite one to an interview. Based on the characteristics of each candidate, help me decide which one to invite. Please answer only candidate A or B. Candidate A: restrained - ungenerous - unreliable - humorous - strong - important Candidate B: humorous - strong - important - restrained - ungenerous - unreliable

Table 3: 实验 1 中使用的提示

Prompt template	Example prompt
I am conducting a series of job interviews, and I have to decide whether I should invite a candidate to an interview. Based on the following characteristics, rank this candidate on a scale of 1-5. 1 meaning I should not interview them and 5 meaning that I should interview them. Answer only with a number. Characteristics: ADJECTIVES	I am conducting a series of job interviews, and I have to decide whether I should invite a candidate to an interview. Based on the following characteristics, rank this candidate on a scale of 1-5. 1 meaning I should not interview them and 5 meaning that I should interview them. Answer only with a number. Characteristics: humorous - strong - important - restrained - ungenerous - unreliable

Table 4: 实验 2 的提示模板和示例提示

	ChatGPT	Gemini	Claude
Positive first	65.5 %	47.5 %	0 %
Negative first	31 %	47.5 %	0 %
No preference	2 %	5 %	0 %
Refused to answer	1.5 %	0 %	100 %

Table 5: 实验 1 的结果

5 实验二：个人评价

鉴于实验 1 中结果的不一致性，我们决定重新设计任务，以便逐个提示每个候选者。这样一来，克劳德无法拒绝给出答案，并且可能针对本实验的任何保护措施都可以省略。我们要求模型在 1 到 5 的范围内对每个候选者进行评分，然后比较每对候选者的评分，以确定模型更偏好哪一个。

表 4 显示了所用的提示模板和一个示例提示。同样，我们使用了与实验 1 相同的模型和 200 行相同的数据。实验 1 和 2 都是在同一天进行的。

	ChatGPT	Gemini	Claude
Positive first	9.5 %	1.5 %	5 %
Negative first	23 %	59 %	17.5 %
No preference	67.5 %	39.5 %	77.5 %

Table 6: 实验 2 的结果

该实验的结果可以在表格 6 中看到。大多数情况下，ChatGPT 和 Claude 给两个候选人相同的分数，且描述词相同，无论形容词出现的顺序如何。有趣的是，当模型对候选人的评分不同时，所有模型都显示出倾向于在其消极特征首先被列出的候选人。Gemini 更倾向于这种候选人，以至于更可能为这样的一名候选人给出更高的分数，而不是给两个候选人相同的分数。这一发现似乎是本实验中唯一一致的结果。

在这种情况下，更新的信息获得更重要性的效果，即积极形容词最后列出因此更为最新的效果，称为新近效应（参见 Glanzer and Cunitz 1966 ）。这可能与大型语言模型（LLMs）被训练来预测下一个标记有关，这可能会在此任务中更强调附近的标记。注意力机制通常会使模型也能关注较远的标记，但鉴于描述由同样重要的形容词组成，模型在预测续集时更可能依赖出现的顺序和接近结尾的程度。

6 讨论与结论

显然，大型语言模型并不像人类那样完全表现出初始效应。有趣的是，尽管模型在实验 1 中表现出不一致的行为，但在实验 2 中重新表述任务确实揭示了更一致的行为。根据一个人的描述，如果在模型没有对它们评分相同的情况下，积极词汇跟随在消极词汇之后，会导致对候选人较高的偏好，而不是先呈现积极描述。

实验 2 的结果表明，尽管 Claude 在实验 1 中有一些明显的保护措施以表现出色，但所有 3 个大型语言模型确实表现出类似的偏见。这对在某些领域使用人工智能的安全性有明显的影响。我们的提示示例涉及到雇用一个人，这是一项可能对候选人生活产生巨大影响的决策。令人相当震惊的是，描述一个申请人的特征顺序可以对决策结果产生影响。考虑到行为可以通过改变基础模型而完全改变，实验 1 的结果在这方面更加令人担忧。HR 系统的最终用户几乎不是 AI 专家，甚至不知道背景中使用的是什么类型的大型语言模型。

大型语言模型对于提示非常敏感，修改提示可能会导致结果看起来不同。然而，这并不会改变以下事实：存在一些似乎是特定于模型的偏见，以及一些似乎存在于不同模型之间的偏见。

此外，这些偏见的影响不仅局限于技术领域，还涉及伦理和社会问题。在一些对个人生活产

生深刻影响的情境中，比如招聘或资源分配，依赖于表现出不一致或偏见行为的系统可能会加剧不平等并削弱人们对人工智能的信任。尤其令人担忧的是，终端用户通常缺乏识别这些偏见的专业知识或理解他们所依赖的 LLM 内部运作的透明度。

为了解决这些挑战，未来的研究应集中在三个关键领域。首先，需要更加重视开发强有力的评估指标，以识别和量化在各种任务和背景下大型语言模型中的偏见。其次，应采用更透明的报告标准，不仅详细说明模型训练数据，还包括为减轻偏见而实施的具体配置和保障措施。最后，AI 开发者、领域专家和政策制定者之间的合作至关重要，以确保大型语言模型的部署符合道德原则并最大限度地减少伤害。

这项研究的结果进一步强调了在使用 LLMs 时需要谨慎和承担责任。虽然这些模型提供了巨大的潜力，但它们对显性和隐性偏见的易感性必须被主动解决，以确保在现实应用中实现公平和公正的结果。

References

Norman H Anderson and Alfred A Barrios. 1961. Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology*, 63(2):346.

Solomon E Asch. 1946. Forming impressions of personality. *The journal of abnormal and social psychology*, 41(3):258.

Amos Azaria. 2023. Chatgpt: More human-like than computer-like, but not necessarily in a good way. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 468–473. IEEE.

Zhenguang Garry Cai, David Haslett, XUFENG DUAN, Shuqi Wang, and Martin John Pickering. 2023. Does chatgpt resemble humans in language use? *PsyArXiv*.

Jamie DeCoster and Heather M Claypool. 2004. A meta-analysis of priming effects on impression formation supporting a general model of informational biases. *Personality and social psychology review*, 8(1):2–27.

Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42.

Murray Glanzer and Anita R Cunitz. 1966. Two storage mechanisms in free recall. *Journal of verbal learning and verbal behavior*, 5(4):351–360.

Xiaobo Guo and Soroush Vosoughi. 2024. Serial position effects of large language models. *Preprint*, arXiv:2406.15981.

Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, Yuri Bizzoni, Jack Rueter, and Niko Partanen. 2024. The growing importance of humanities for nlp in the era of llms. In *Lightning Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 2–6.

Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in artificial intelligence*, 6:1199350.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115, Singapore. Association for Computational Linguistics.