STA5007 Assignment 4

Li Xuanran
12312110

January 3, 2026

1 Q1 (3 pts)

Model interpretability is a key research area in understanding LLM behaviors. Read the paper Emergent Introspective Awareness in Large Language Models and use an LLM to help answer the 10 classic research paper reading questions suggested in this blog.

Use an LLM (e.g., ChatGPT or similar) to generate answers to each question based on the paper.

Select three QA pairs among the LLM-generated contents to add your own commentary (agreement/disagreement, new insights, questions, etc).

Solution. LLM used: **Gemini 3 pro**. The answer is as belows (to ensure consistency, the answers are also provided in Chinese) :

- Q1: 这篇文章究竟讲了什么问题? Input 和Output 是什么?
- A1: 问题：文章探讨的核心问题是“大型语言模型（LLMs）是否真正具备‘内省’

  （Introspection）能力？”即模型能否感知并报告其内部状态，还是仅仅在根据训练数据模仿内省的语气（即“虚构”）。

  算法/方法（Input/Output）：算法设计：文章设计了一种称为“概念注入”（Concept Injection）的方法，这是一种激活引导（Activation Steering）技术的应用。Input（输入）：在模型处理文本的过程中，直接向模型的内部激活（Residual Stream）中注入代表特定概念（如“面包”或“大写字母”）的向量。Output（输出）：模型的口头自我报告（例如回答“你现在在想什么？”）或者模型对预填充内容的反应（例如是否承认某个由注入产生的词是自己“有意”说的）。

- Q2: 这个问题的性质是什么？是一个新的问题吗?
- A2: 性质：这是一个机械可解释性（Mechanistic Interpretability）与认知科学交叉的问题，涉及人工智能的“元认知”能力。

  新旧程度：这是一个并不完全是新的问题，但文章的研究方法非常新颖。此前已有研究探讨模型“是否知道自己知道什么”（Knowing what they know），但大多仅停留在行为层面的观察。

  重要性：它的重要性在于“由内而外的验证”。如果模型能真正内省，我们就能通过询问模型来获得其推理过程的真实解释，这对于AI的透明度（Transparency）、可信度以及安全性（如检测欺骗）至关重要。

- Q3: 这篇文章致力于证明什么假设?
- A3: 现代大型语言模型（特别是像Claude Opus 4/4.1这样更强的模型）拥有某种形式的“功能性内省意识”。这意味着它们不仅是在模仿人类说话，而是确实能够监测自己的内部激活状态，并根据这些状态调整输出。
- Q4: 有哪些与这篇文章相关的研究? 这一领域有哪些关键人物?
- A4: 相关研究：Language Models (Mostly) Know What They Know (Kadavath et al., 2022) - 关于模型对自己知识边界的认知。Representation Engineering (Zou et al., 2023) - 关于如何提取和控制模型内部表征。Transformer Circuits 系列文章- 关于模型内部电路的逆向工程。

  关键人物：本文作者Jack Lindsey (Anthropic)。该领域领军人物：Chris Olah (Anthropic 可解释性团队负责人，Transformer Circuits Thread 的核心人物)，以及Jared Kaplan、Dario Amodei 等在Scaling Laws和AI安全方面的重要人物。
- Q5: 这篇文章提出的问题解决方案中，核心贡献是什么?
- A5: 核心贡献：建立了一种因果验证方法。不仅仅是听模型“怎么说”，而是通过干预（注入概念）来看看模型是否能“察觉”到这种干预。文章证明了模型能区分“外部输入的文本”和“内部产生的想法”，并且这种区分是基于真实的内部状态监测，而非仅仅是根据上下文进行的合理化推测（Confabulation）。
- Q6: 实验是如何设计的?
- A6: 实验1（检测注入）：向模型注入一个概念向量（如“大写字母”），然后问模型：“你有没有感觉到什么异常？”看模型能否报告检测到了“注入的思想”。

  实验2（区分来源）：同时给模型阅读一段文本并注入一个概念，看模型能否分清哪个是读到的（外部输入），哪个是想到的（内部注入）。

  实验3（意图检测）：强行预填充（Prefill）模型的一个输出词（例如一个不通顺的词），并在前文中注入该词的概念向量。如果模型因为被注入了概念而承认这个词是“有意为之”的，说明模型通过检查过去的内部状态来判断意图。
- Q7: 实验是在什么样的数据集基础上运行的? 数据可复现吗?
- A7: 数据集：实验主要使用的是研究者自行生成的合成数据集。包括50 个用于提取概念向量的随机概念词（如“vegetables”、“amphitheaters”）以及50 个用于阅读任务的随机句子。

  可接触性：读者无法直接下载一个通用的“内省基准数据集”，因为这更多是一种实验设置而非固定数据。但文章附录详细描述了生成这些提示词（Prompts）和概念向量的方法，具备技术能力的读者可以在拥有模型权重访问权限（或类似的可解释性接口）的情况下复现这一过程。
- Q8: 实验结果能否有力地支持假设?
- A8: 支持程度：有力，但有局限性。结果显示，Claude Opus 4.1 在适当的层和注入强度下，能够以极高的准确率（远超随机猜测）检测到注入，并能区分“读到的”和“想到的”。关键证据：模型在说出被注入的词之前就能报告检测到了异常，这排除了模型是根据自己已经说出的内容来“编造”理由的可能性。局限：这种能力目前还不可靠（Unreliable），受上下文影响大，且并非在所有模型上都表现一致。
- Q9: 这篇文章的贡献是什么?

- A9: 它提供了第一个令人信服的证据，证明LLM的某些自我报告（Self-reports）不仅仅是模仿人类的语言模式，而是基于对自身计算过程的真实访问（Grounding in internal states）。它为区分“机器幻觉”和“真实内省”提供了一套科学的方法论。
- Q10: 下一步可以做什么？
- A10: 系统化评估：开发更标准化的基准测试，来衡量不同模型的内省能力。机制解释：深入研究是哪些具体的“电路”（Circuits）或注意力头（Attention Heads）负责执行这种自我监测功能。应用研究：利用内省能力来提高模型的安全性，例如让模型在意识到自己正在产生有害内容或被欺骗时发出警报，或者通过内省来检测模型是否在进行复杂的欺骗行为（Scheming）。

Agreement: For Q2 and A2, I agree that this is a critical shift. Purely behavioral tests are susceptible to data contamination where the model might have seen similar questions. By directly manipulating the internal activations, the researchers skip the surface imitation and force the model to respond to its own hidden state.

New Insight: For Q5 and A5, they suggest that LLMs might possess a latent space sensor. If models can distinguish between self-generated signals and external inputs, we might be able to develop a fact-checking layer that flags when a model's output is purely driven by internal hallucinations rather than the provided context.

Question: For Q10 and A10, if a model becomes know that we are probing its internal states via concept injection, could it learn to mask or spoof those internal signals to appear more aligned or compliant? This raises a question that whether introspective awareness could eventually lead to more sophisticated forms of deception.

□

2 Q2 (3 pts)

Large language models often face challenges in information fusion tasks, including data omission, hallucinations, and inconsistency. This task explores these limitations through case analysis.

Task instructions:

- **Scenario Selection:** Choose a real-world application scenario (e.g., news, healthcare). Focus on common LLM failures during information integration.
- **Data Preparation:** Collect 3 articles (or news reports) related to the same event (around 600 words, in Chinese or English, you can search them online and ensure these are different documents), published within the past six months. We add this filter to target at those information that is not trained during pretraining.
- **Model Test:** Write prompt to use GPT-5 or another proprietary LLM to fuse the contents of the three articles into one unified article.
- **Issue Analysis:** Evaluate and analyze potential issues (e.g., data loss, factual errors, hallucinations) in the output. You can write your report and commentary following the json format below.

Output Format Example:

```
[
    {
        "text1": "Content of input document 1",
        "text2": "Content of input document 2",
        "text3": "Content of input document 3",
        "fusion_text": "LLM-generated fused summary",
        "evaluation_problems": [
          # should clearly identify paragraph numbers (e.g., ‘"Paragraph 3"‘).
            "Omission: Paragraph 3: Missing key information from input document 2",
            "Incorrect data: Paragraph 5: Mistaken event date from 2023 to 2024",
            "Hallucination: Paragraphs 7-8: Fused summary contains facts not in any input"
        ]
    }
]
```

Solution. Database: Article 1, Article 2, Article 3.

Commentary: See `comment.txt`.

```
[
    {
        "text1": "Article 1 (NY Post/Cybernews): Reports David Moss's Dec 31, 2025, zero-intervention drive from LA to Myrtle Beach (2,732.4 miles) using FSD V14.2 in a 2025 Model 3. Notes it fulfilled Musk's 2016 prediction and included all parking and charging. Musk commented 'Cool.'",
        "text2": "Article 2 (Investors.com/WebProNews): Discusses Tesla ramping up 'Robotaxi' messaging before a late-2025/early-2026 deadline. Mentions 'Cybercab' production in April 2026 and testing without human supervisors in Austin. Highlights stock volatility and expert skepticism vs. Waymo.",
        "text3": "Article 3 (TradingView/Invezz): Compares Tesla FSD v14.2.2.2 to Waymo. Notes Tesla is Level 2 (supervised) while Waymo is Level 4 (driverless). Mentions reliability gap (1,500 miles for Tesla vs. 17,000 for Waymo) and regulatory hurdles (approval in only 2 states for Tesla).",
        "fusion_text": "### Historic Coast-to-Coast FSD Drive Reignites T e s l a s Robotaxi Ambitions Amidst Skepticism (Generated in the previous turn)",
        "evaluation_problems": [
            "Precision Error: Paragraph 2: The fused text identifies the driver as 'David Moss' but attributes his arrival to 'Myrtle Beach, South Carolina' on December 31, 2025. While Article 1 confirms this, it omits the specific detail from the source that he had already crossed '10,000 consecutive miles' of intervention-free driving prior to the coast-to-coast finish.",
            "Omission: Paragraph 4: Missing the specific numerical reliability comparison from Article 3. While it mentions the 'march of nines,' it fails to include the concrete data point that Waymo averages 17,000 miles between interventions compared to Tesla's 1,500 miles, which is a key factual pillar of the skepticism described.",
            "Data Inconsistency: Paragraph 1 & 4: The fused text mentions an 'April 2026 production deadline' for the Cybercab. Article 2 mentions scaling in 2026, but also refers to an 'end-of-year [2025]
```

```
deadline' for launching unsupervised services. The fusion merges these
timelines without clarifying that the 2025 deadline was for *
unsupervised testing/launch* while 2026 is for *volume production*.",
        "Factual Nuance Loss: Paragraph 3: The fused text states the
drive 'fulfilled a prediction Elon Musk made... for the end of 2017.'
It does not clearly distinguish that the 2017 goal was specifically for
 a *demo* drive, whereas this 2025 event was a *consumer-led*
achievement, which Article 1 and 3 contrast as a long-overdue milestone
.",
        "Hallucination: Paragraph 2: The text claims the journey took
'two days and 20 hours.' While this matches Article 1, the fused text's
 framing of the software as 'FSD V14.2' in one place and implying '
Level 4' status through the term 'fully autonomous' in the title risks
hallucinating the legal/technical classification, which Article 3
explicitly states remains 'Level 2'."
    ]
  }
]
```

□

3 Q3 (4 pts)

Use the DeepResearch framework to explore LLM-based deep scientific reasoning.

Task Instructions:

- Access the Bailian online service linked in the repository under:

  > *Welcome to try Tongyi DeepResearch via our Modelscope online demo or Huggingface online demo or bailian service!*

- Choose a graduate-level mathematical modeling problem and let the LLM solve it via the DeepResearch interface.

In your report, include:

- The problem you posed. You can refer to the problems in the MCM/ICM contest
- The LLM's response.
- Your critical evaluation of the response:
  - What was correct or helpful?
  - What errors or limitations did the LLM show?
  - Discuss strengths and weaknesses of using LLMs for deep research.

Solution. For the problem I posed, see `prob.pdf`. For LLM's response, see `result.pdf`.

Personally, I think the response is not totally correct, but very helpful. The LLM demonstrated excellent judgment by selecting a Hierarchical Bayesian Poisson model, which is the scientifically correct approach for predicting discrete medal counts while accounting for country-level variations. It addressed all prompt requirements, providing a professional report structure that includes specific 2028 projections through Difference-in-Differences analysis, and logical strategic advice for small nations to specialize in specific sports. This provides a high-quality academic template.

For errors or limitations, the most obvious limitation is the high risk of hallucination. While the LLM claims precise results like an R^2 of 0.89 and a *Great Coach* effect of 1.52 medals, these figures are likely fabricated guesses rather than outputs from actual code execution. Additionally, the research is shallow in terms of evidence; it mentions identifying 12 specific coaches but fails to name them, relying instead on the two examples already provided in the prompt.

For strengths and weaknesses, the primary strength of using an LLM for deep research is its ability to instantly get complex frameworks and suggest rigorous mathematical strategies. However, the critical weakness is that the model generates professional-looking reports that contain unverified or invented data. This makes the LLM a powerful tool for deep search, but dangerous if used as a final answer without a human manually performing the data cleaning and rigorous calculation.

□