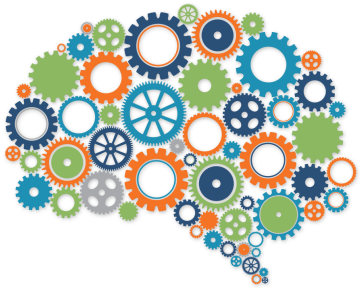

Critical Thinking & Innovation

Jake Chandler

Large Language Models, Pt 2

LA TROBE UNIVERSITY

Introduction

- Last week:

  What exactly can LLMs do in the first place?

- Benchmarking: performance assessment using standardised tests
- LLM responses can be open-ended: grading often by humans rather than automatic (e.g. Chatbot Arena)
- Similar to fine-tuning stage of training, without learning element
- Things change fast! Online leaderboards provide updated performance results:
  - Huggingface Open LLM Leaderboard
  - Chatbot Arena LLM Leaderboard

Introduction (ctd)

- LLMs show impressive capability across a broad range of tasks
  - Language understanding (MMLU, …)
  - Multi-turn dialogue (MT Bench, …)
  - Coding (HumanEval, …)
  - Basic maths (MATH, …)
- What about verbal reasoning tasks?
- Some tests:
  - HellaSwag: Tests common sense inference in various scenarios, asking for most plausible continuation of a given situation.
  - LogiQA: Tests logical reasoning skills with complex, multi-step logical inference problems.
  - FOLIO: Tests complex logical reasoning (+ translation into logic)

Introduction (ctd)

- Question:

  Human reasoning exhibits systematic shortcomings. How do LLMs compare?

Some studies

- Binz & Schulz (2023): GPT-3 vs 12 classic experiments
- GPT-3 performed as badly as humans on
  (1) Linda (Conjunction fallacy)
  (2) Items from Cognitive Reflection Test (CRT)
  (measures "ability …to reflect on a question and resist reporting the first response that comes to mind" Shane 2005)

CRT bat/ball question

"A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?"

(Correct answer: $0.05 ; common answer: $0.10)

- Poor performance on (2) corroborated in Hagendorff *et al* (2022)

Some studies (ctd)

- GPT-3 performed better than humans on
  (3) Cab (Base rate fallacy)
  (4) Cards (Conditional reasoning)

…*but* performance dropped with slight variations:

(3) asking for the prob. of outcome not mentioned in vignette
(4) changing order of presentation of cards

⇒ Initial good performance likely due to training on similar cases

Some studies (ctd)

- In humans, performance on card selection task is affected by non-formal aspects (Content Effect)
  Better performance when original abstract task (number and type of letter) replaced by non-abstract one (age and permission to drink)
- Lampinen *et al* (2024): similar situation in LLMs (incl. GPT-3.5)

Some studies (ctd)

- Content effects also seen in human assessments of validity
  - Validity less likely to be recognised with false conclusion
  - Invalidity less likely to be recognised with true conclusion

Content effects on validity judgments

Valid with False Conclusion	*Invalid with True Conclusion*
(1) All US presidents have been men. (2) Obama is a man. ——— (3) Obama was a US president.	(1) All cats are reptiles. (2) All lizards are cats. ——— (3) All lizards are reptiles.

- Lampinen *et al* (2024): similar situation in LLMs

Adding System 2 Thinking?

- Early LLMs have shortcomings associated with human System 1
- Maybe they reason similarly?
- Two suggestions:
  (1) add module for S2 thinking (Nye *et al* 2021)
  (2) emulate S2 thinking with different prompt style
- *Possible* candidate re (2):

  Chain of Thought (CoT) prompting (Wei *et al* 2022): forcing the LLM to report sequence of reasoning steps

CoT prompting

- Increases performance in general, esp. in larger models, for harder problems (see Open CoT Leaderboard)
- Increases performance wrt tasks difficult for System 1 in humans
  - Lampinen *et al* (2024): CoT prompting improves performance on card selection
  - Hagendorff *et al* (2022):
    - CoT prompting improves CRT performance in GPT-3 (from 5% to 28%; compare 38% in humans)
    - CRT performance improved dramatically in GPT-3.5 (59%) & GPT-4 (96%): with these models, CoT answers emerged without being requested

CoT prompting (ctd)

- Zero-shot CoT: request step by step reasoning in the prompt

Zero-shot CoT

```
Solve the equation 4z + 6 = 22. Let's approach this step by
step. / Let's think through this logically.
```


CoT prompting (ctd)

- Few-shot CoT: add worked examples involving the relevant steps

Few-shot CoT

```
Here are some examples of solving a linear equation:

    Example 1: Solving 2x + 5 = 13
    Here is how to solve the equation 2x + 5 = 13 step by
    step:
    Step 1: Subtract 5 from both sides 2x + 5 - 5 = 13 - 5
    2x = 8
    Step 2: Divide both sides by 2 2x ÷ 2 = 8 ÷ 2 x = 4
    Therefore, the solution is x = 4.

    Example 2: …

Now, solve the equation 4z + 6 = 22.
```

Resources

- Huggingface Open LLM Leaderboard: `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`
- Chatbot Arena LLM Leaderboard: `https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard`
- Huggingface Open CoT Leaderboard: `https://huggingface.co/spaces/logikon/open_cot_leaderboard`

References

- Bellini-Leite SC (2024). 'Dual Process Theory for Large Language Models: An overview of using Psychology to address hallucination and reliability issues'. *Adaptive Behavior* 32(4):329-343.
- Binz, M. & E. Schulz (2023). 'Using cognitive psychology to understand GPT-3'. *PNAS* 120(6).
- Evans J, *et al* (1983). 'On the conflict between logic and belief in syllogistic reasoning'. *Mem Cogn.* 11(3):295–306.
- Hagendorff, T. *et al* (2022). 'Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT'. *Nature Computational Science*, 3, 833 - 838.

References (ctd)

- Lampinen, A. *et al* (2024). 'Language models, like humans, show content effects on reasoning tasks'. *PNAS Nexus*, Volume 3, Issue 7. Available at `https://doi.org/10.1093/pnasnexus/pgae233`
- Nye, M. *et al* (2021). 'Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning'. *Neural Information Processing Systems*.
- Shane F. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19(4), pp. 25–42.
- Wei J, *et al* (2022). 'Chain-of-thought prompting elicits reasoning in large language models'. *Advances in neural information processing systems*. 35:24824–24837.