LLMs as assistants or tests subjects?

Matthijs Westera
Leiden University Centre for Linguistics

Takeaways

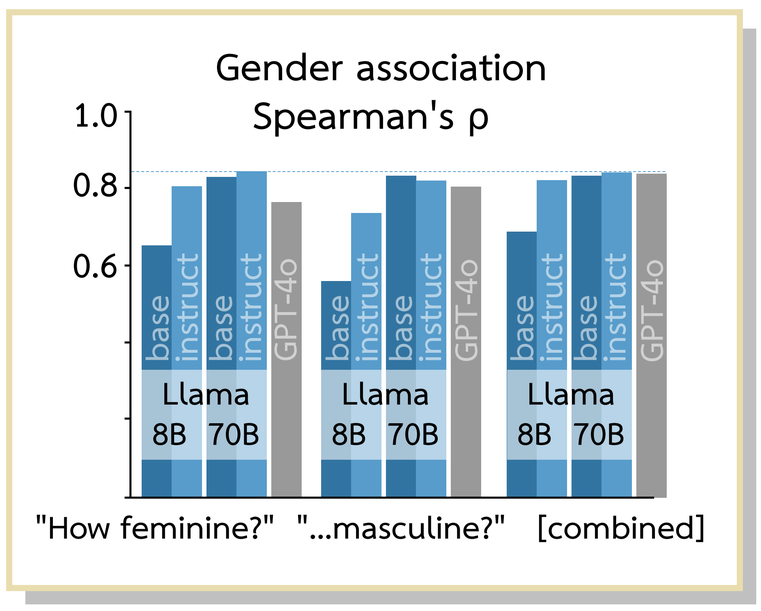
- Two distinct conceptions of LLMs.
- Be mindful of issues known from experiment/questionnaire design.

Method

- Compare LLMs to human data:
  - *Gender association* word norms.[1]
  - *Word relatedness* scores.[2]
- Compare different models, and different ways of prompting.
- Reusable tool: *ChoiceLLM* (on PyPI)

Framing

Potential benefit of aggregating 'masculine' + 'feminine' framing



References

1. Scott et al. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior research methods.*
2. Westera et al. (2021). Distributional models of category concepts based on names of category members. *Cognitive Science.*
3. Kauf et al. (2024). Comparing plausibility estimates in base and instruction-tuned large language models. *ALTA.*

Issues in experiment design

```
How masculine is... ?
  dress    1  2  3  4  5
  truck    1  2  3  4  5
  gun      1  2  3  4  5
   ...
```

- Nominal/ordinal/interval?
- Reference level?
- Ordering effects.
- Framing effects:

```
How feminine is... ?
```

- 'Good participant bias'

Base vs. instruct-tuned models?

- Prompt: plain continuation vs. chat
- Base models may better 'fit' humans.[3]

Scalar rating vs. comparison

- Compare against (e.g.) 200×3 others:

```
Which is more ... (masculine, related)?
```

- Use aggregated 'win ratio' as rating.[2]

- And ρ rapidly increases with # comparisons: