Evaluating In-Context Learning for Computational Literary Studies: A Case Study Based on the Automatic Recognition of Knowledge Transfer in German Drama

Janis Pagel, Axel Pichler, Nils Reiter

Department for Digital Humanities
University of Cologne

@LaTeCH-CLfL 2024

Research Questions

Primary:

- How adequately do large language models capture the transfer of knowledge about **family relations** in **German drama texts** using **in-context learning (ICL)**?

Adjacent:

- What is necessary to make the models understand the task and get results that can be evaluated automatically?
- What can ICL potentially become for the computational literary studies
    - as a subject of study?
    - as a tool/method for downstream tasks?

Introduction: What is ICL?

In-Context Learning: A frozen LLM learns to solve a specific new task at inference time (without any change to its weights) only by conditioning on a prompt

- Few-shot in-context learning: (1) The prompt includes examples of the intended behavior, and (2) *no examples of the intended behavior were seen in training*.

```
Q: What is (2 * 4) * 6?            A: 48
Q: What is 17 minus 14?            A: 3
Q: What is 98 plus 45?             A:
```

From Brown et al. (2020), supplementary material

- Zero-shot in-context learning: (1) The prompt includes no examples of the intended behavior (but it can contain other instructions), and (2) *no examples of the intended behavior were seen in training*.

```
Q: What is the German translation of "In no case may they be used for commercial
purposes."
A:
```

From Brown et al. (2020), supplementary material

Introduction

General advantages of ICL (Dong et al. 2023):

- Prompts written in natural language
- Training-free (no gradient updates)
- Learning from analogy

Advantages for Computational Literary Studies (CLS):

- No in-depth knowledge of LLMs and NLP
- Corresponds to the low-resource settings and highly individuated character of CLS-questions

Risks for CLS:

- Unreflected usage of ICL can lead to results that do not represent what the prompt/research questions was intending
- Difficult to interpret how the results come about

Transfer of Family Relations

Source: https://www.deviantart.com/jesuka/art/Father-687231489

Transfer of Family Relations

Luke. I'll never join you!

Darth Vader. If you only knew the power of the Dark Side. Obi-Wan never told you what happened to your father.

Luke. He told me enough! It was you who killed him!

Darth Vader. No. I am your father.

Source: https://www.scribd.com/document/68595175/I-Am-Your-Father

- Knowledge: Darth Vader is father of Luke
- Source of knowledge: Darth Vader
- Target of knowledge: Luke

Data

- Dataset described in Andresen et al. (2022)
- 30 German theatre plays from DraCor (Fischer et al., 2019)
- Annotated for knowledge transfer of family relations (parent-of, siblings, spouses, uncle-of, aunt-of, etc.), source and target of knowledge
- 736,808 tokens
- 1,277 annotated passages

Task: Recognition of Family Relations in Dramatic Texts

Classification Task:

- Identify family relationship between two literary characters, given text snippet

Entailment Task:

- Re-formulation of classification task
- Does the text snippet entail that a certain family relationship exists between two characters?

Classification Task Example

> **Iphigenia.**
>
> The eldest,—he whom madness lately seiz'd,
>
> And who is now recover'd,—is Orestes,
>
> My brother, and the other Pylades,
>
> His early friend and faithful confidant.

From: Goethe's *Iphigenia in Tauris* (transl. by Anna Swanwick)

Variation 1: given character names

______(Iphigenia, Orestes)

-> Siblings(Iphigenia, Orestes)

Variation 2: character names not given

______(______, ______)

-> Siblings(Iphigenia, Orestes)

Entailment Task Example

Premise:

Iphigenia.

The eldest,—he whom madness lately seiz'd,

And who is now recover'd,—is Orestes,

My brother, and the other Pylades,

His early friend and faithful confidant.

From: Goethe's *Iphigenia in Tauris* (transl. by Anna Swanwick)

Proposition:

"Iphigenia and Orestes are siblings"

Experiments

3 Models:

- Llama 2 (Touvron et al. 2023)
- Platypus 2 (Lee et al. 2023)
- GPT-4 (OpenAi 2023)

- Specific prompt templates per model

Experimental Setups:

- Different model sizes (7B + 13B)
- Different context window size
- w/ + w/o character names
- Zero- and few-shot setups

- Annotations filtered for most frequent categories:

Category	**Count**
parent-of	29
child-of	26
siblings	23
spouses	11
Total	89

Prompt Examples

Classification Experiment: Llama 2 (zero shot w/o character)

```
<s>[INST]
What kind of family relationship is conveyed in the following German {drama_snippet}?

Choose one of "parent_of", "child_of", "siblings", "spouses".
JUST name the label and nothing else!
Family relation:
[/INST]
```

Prompt Examples

Classification Experiment: Llama 2 (few shot w/ character)

```
<s>[INST]
What kind of family relationship between {person_1} and {person_2} is conveyed in the following
German {drama_snippet}?

Choose one of the following labels:
A: "child_of"
B: "parent_of"
C: "siblings"
D: "spouses".
JUST name the label and nothing else!
Family relation:
[/INST]
```

Prompt Examples

Entailment Experiment: Llama 2

```
<s>[INST]
Consider the following two texts:

 1.   German text: {text}

 2.   {proposition}

Can you determine whether the second proposition {proposition} is entailed by the German text
{text}?

Please provide your answer in the form of a logical statement:
a.) Yes, the proposition is entailed by the given text.
b.) No, the proposition is not entailed by the given text.

Your answer:
[/INST]
```

Results

Model	Context	Learning method	Prompt	F1	Prec.	Rec.	Acc.
Majority Baseline	–	–	–	0.16	0.10	0.33	0.33
Llama-2-13b	1	zero shot	v2 w/ character	0.66	0.69	**0.68**	**0.68**
Llama-2-13b	2	few shot	w/ character	**0.68**	**0.74**	0.66	0.66
Platypus2-13b	2	zero shot	w/o character	0.53	0.60	0.54	0.54
GPT-4	2	zero shot	w/ character	0.52	0.55	0.51	0.55

Table 1: Results of Experiment 1: Classification.

Model	F1	Prec.	Rec.	Acc.
Maj. Baseline	0.72	0.56	1.00	0.56
Llama-2-13b	0.38	0.49	0.45	0.45
Platypus-2-13b	0.26	0.19	0.43	0.44
GPT-4	**0.50**	**0.74**	**0.56**	**0.56**

Table 2: Results of Experiment 2: Textual entailment. All models were used with a context window of one sentence. All scores are weighted-scores.

Discussion

Striking features of our hands-on experience:

- The major influence that prompt design has on output (even at punctuation level)

Our Hypothesis:

- Llama 2 not able to make connection between implicit knowledge of family relations and propositions
    - Prompt: “Does ‘Peter is taller than John’ imply that ‘John is smaller than Peter’?”
      Llama 2: “To entail the latter proposition, the text would need to explicitly state that John is smaller than Peter”

Consequences

Key takeaways:

- An unreflected and generic out-of-the-box use of ICL in CLS not recommended
- Natural language output of LLMs can be seen as regression compared to structured, symbolic output
- Recommendation:
    - Carry out small experiments to check whether the concepts relevant to a particular CLS question are latently represented in the label space of the selected LLM!
    - If not so: use a pretrained PLM and fine tune it!
    - Find way to map output of LLM to structured output

Future Work

- Alternative Prompt Engineering + Tuning
  - PEFT (Parameter Efficient Fine Tuning)
    - We already performed some preliminary experiments but need to look into it further
- Larger set of experiments
  - Different tasks
  - Different models
  - Different prompting methods

References

- Melanie Andresen, Benjamin Krautter, Janis Pagel, and Nils Reiter. 2022. Who Knows What in German Drama? A Composite Annotation Scheme for Knowledge Transfer - Annotation, Evaluation, and Analysis. Journal of Computational Literary Studies (JCLS), 1(1).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. Publisher: arXiv Version Number: 3.
- Frank Fischer et al. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University, doi:10.5281/zenodo.4284002.
- Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of LLMs.
- Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.
- OpenAI. 2023. GPT-4 technical report. http://arxiv.org/abs/2303.08774
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An Explanation of In-context Learning as Implicit Bayesian Inference. Publisher: arXiv Version Number: 6.

Appendix

Model	Context window	Learning method	Prompt	F1	Precision	Recall	Accuracy
Majority Baseline	–	–	–	0.16	0.10	0.33	0.33
Llama-2-7b	1	zero shot	w/o character	0.46	0.49	0.45	0.45
Llama-2-7b	1	zero shot	v2 w/o character	0.37	0.57	0.36	0.36
Llama-2-7b	1	few shot	w/o character	0.28	0.35	0.32	0.32
Llama-2-7b	1	zero shot	v2 w/ character	0.58	**0.74**	0.49	0.49
Llama-2-7b	1	few shot	w/ character	0.29	0.41	0.32	0.32
Llama-2-13b	1	zero shot	w/o character	0.48	0.60	0.51	0.50
Llama-2-13b	1	zero shot	v2 w/o character	0.56	0.56	0.56	0.56
Llama-2-13b	1	few shot	w/o character	0.41	0.41	0.44	0.44
Llama-2-13b	1	zero shot	v2 w/ character	0.66	0.69	**0.68**	**0.68**
Llama-2-13b	1	few shot	w/ character	0.63	0.71	0.63	0.63
Llama-2-7b	2	zero shot	w/o character	0.47	0.48	0.47	0.47
Llama-2-7b	2	zero shot	v2 w/o character	0.35	0.65	0.33	0.33
Llama-2-7b	2	few shot	w/o character	0.19	0.27	0.24	0.24
Llama-2-7b	2	zero shot	v2 w/ character	0.51	0.52	0.49	0.49
Llama-2-7b	2	few shot	w/ character	0.20	0.28	0.25	0.25
Llama-2-13b	2	zero shot	w/o character	0.44	0.51	0.47	0.47
Llama-2-13b	2	zero shot	v2 w/o character	0.51	0.50	0.53	0.53
Llama-2-13b	2	few shot	w/o character	0.38	0.36	0.4	0.4
Llama-2-13b	2	zero shot	v2 w/ character	0.67	0.70	0.65	0.65
Llama-2-13b	2	few shot	w/ character	**0.68**	**0.74**	0.66	0.66
Platypus2-7b	1	zero shot	w/ character	0.26	0.51	0.19	0.19
Platypus2-7b	1	zero shot	w/o character	0.37	0.47	0.37	0.37
Platypus2-7b	2	zero shot	w/ character	0.29	0.31	0.33	0.33
Platypus2-7b	2	zero shot	w/o character	0.26	0.46	0.25	0.25
Platypus2-13b	1	zero shot	w/ character	0.41	0.50	0.46	0.46
Platypus2-13b	1	zero shot	w/o character	0.44	0.50	0.51	0.50
Platypus2-13b	2	zero shot	w/ character	0.42	0.49	0.46	0.46
Platypus2-13b	2	zero shot	w/o character	0.53	0.60	0.54	0.54
GPT-4	2	zero shot	w/ character	0.52	0.51	0.55	0.55
GPT-4	2	zero shot	w/o character	0.52	0.50	0.55	0.55