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

Janis Pagel and **Axel Pichler** and **Nils Reiter**
Department for Digital Humanities
University of Cologne
{janis.pagel,axel.pichler,nils.reiter}@uni-koeln.de

Abstract

In this paper, we evaluate two different natural language processing (NLP) approaches to solve a paradigmatic task for computational literary studies (CLS): the recognition of knowledge transfer in literary texts. We focus on the question of how adequately large language models capture the transfer of knowledge about family relations in German drama texts when this transfer is treated as a classification or textual entailment task using in-context learning (ICL). We find that a 13 billion parameter LLAMA 2 model performs best on the former, while GPT-4 performs best on the latter task. However, all models achieve relatively low scores compared to standard NLP benchmark results, struggle from inconsistencies with small changes in prompts and are often not able to make simple inferences beyond the textual surface, which is why an unreflected generic use of ICL in the CLS seems still not advisable.

1 Introduction

Computational literary studies (CLS) is a subfield of Digital Humanities. CLS attempts to expand the traditional methods of literary studies to include quantitative approaches with the help of statistical methods and machine learning or natural language processing (NLP) (Piper et al., 2021; Jannidis, 2022). The latter has made enormous progress in recent years: at the latest since the development of the transformer architecture (Vaswani et al., 2017) large language models (LLMs) based on it have been breaking traditional NLP-benchmarks. Until 2020, the development and domain-adaption of these models was dominated by an approach that emerged as a result of the bidirectional encoder representations from transformer models (BERT, Devlin et al., 2018), which has been termed the *pre-training* and *fine-tuning* paradigm. This practice has now also arrived in CLS, as a recent survey on machine learning in CLS shows (Hatzel et al., 2023). However, another paradigm shift in NLP appeared when the developers of GPT-3 (Brown et al., 2020), an autoregressive LLM with around 175 billion parameters, showed in 2020 that with the help of a few examples ('few-shot learning') without fine-tuning and exclusively through natural language interaction, the model not only corresponded to the performance of predecessor models in numerous NLP tasks, but even outperformed them. This type of conditioning of a LLM to perform a specific task using only a natural language input and no gradient update is called 'in-context learning' (ICL, Dong et al., 2023). With the publication of ChatGPT, ICL has gained popularity among the general public, but its potential and limits for CLS has yet to be determined.

Use of Language Models in CLS

Especially for the often highly individualized research questions of CLS, it is tempting to provide a natural language description of the task in order to analyze literary texts. On the first sight, such an approach does not require in-depth knowledge of LLMs and NLP, and even the output – human-readable language – can be interpreted directly, without requiring quantitative and/or statistical analysis. Furthermore, this approach seemingly gets rid of the need to formulate unambigouous and precise definitions, which are tested via annotation and – due to the explication – open for critique by other researchers (cf. Reiter et al., 2019). Thus, in order to include ICL in the CLS method arsenal, it is necessary to disclose the potentials and limitations of this method properly. With this paper, we want to take a first step in this direction by testing a representative CLS-task – knowledge transfer between literary characters in German theatre plays – with the help of three different LLMs using ICL-methods.

We consider this task representative for many CLS tasks: i) It revolves around literary characters,

Proceedings of LaTeCH-CLfL 2024, pages 1–10

which are one of the most important 'anchors' for literary interpretation. Their knowledge about the world they live in is an important property if one understands literary characters as representations of human beings. ii) While some CLS tasks (e.g., authorship or genre attribution) assign properties to the entire text, many focus on smaller units such as scenes or events, which are represented by a small number of tokens. iii) Finally, corpus sizes in CLS are typically rather small, as much of the work is focused on historic data.

The paper directly links to the flourishing drama research in literary studies and CLS (Moretti, 2011; Fischer et al., 2017; Krautter, 2018; Andresen et al., 2022; Dennerlein et al., 2023). The goal of this study is two-fold: (i) evaluating if LLMs are able to sufficiently solve the task out-of-the-box, and (ii) investigating the problems and pitfalls that may arise when utilizing LLMs for such a CLS task.

2 Related Work

While prompt engineering, which is often equated with ICL, recently gained a lot of traction, studies based on it are still rare in DH in general and CLS in particular. Initial reflections and experiments can be found in Computational Science Studies (CSS). Ziems et al. (2023) investigate the potential of LLMs for CSS by investigating the viability of zero-shot-learning for sociological research. Overall, they posit that LLMs perform well in certain zero-shot classification tasks within the context of social studies research, but "do not match or exceed the performance of carefully fine-tuned classifiers" (Ziems et al., 2023, p. 2). Across their experiments, models demonstrate optimal performance in tasks related to misinformation classification, stance detection, and emotion classification. The authors attribute this success to the presence of either a ground truth (as in the case of misinformation) or an annotation schema that corresponds to (implicit) definitions of everyday concepts.[1] However, they also note that models perform worse in tasks requiring intricate expert taxonomies. This difficulty arises from the complex nature of expert-informed annotation guidelines, which may not align semantically with much of the LLM's training data.

[1]These assumptions correlate with the initial research into why and how ICL works, summarized and brought together by Xie et al. (2021) and Xie and Min (2021) who interpret this ability as Bayesian inference of a latent concept conditioned on the prompt – a capability that arises from structure in the pretraining data.

In the realm of NLP, several papers have explored ICL for solving sets of diverse tasks (cf. Ye et al., 2021; Chen et al., 2022; Wei et al., 2022; Sanh et al., 2022; Min et al., 2022), investigated methods for constructing reliable prompts (cf. Schick and Schütze, 2021; Zhao et al., 2021; Perez et al., 2021; Lu et al., 2022) and choosing suitable evaluation metrics (Schaeffer et al., 2023). In terms of related tasks, the `implicit_relations` task from the BIG-bench benchmark (Srivastava et al., 2023) seems to be closest to our setup, with Hoffmann et al. (2022) achieving an accuracy score of 49.4% using a 70B parameter Chinchilla model and Rae et al. (2021) an accuracy score of 36.4% using a 280B parameter Gopher model.

3 Task

We work on a paradigmatic task for Computational Literary Studies: the transfer of knowledge about family relationships in German theatre plays. This type of task is paradigmatic for CLS insofar as research questions from literary studies often focus on particular realizations of concepts in specific literary contexts or specific particularities of genres and single texts which are difficult to generalize and thus data sparsity becomes an issue. We conduct two experiments within the framework of ICL in order to gain insights into how LLMs can be applied here:

- **Experiment 1: Classification**: For our first approach, we aim to investigate if LLMs are able to correctly predict family relations between characters. The task for the models is to decide which family relation holds between two characters given the context in a dialogue snippet taken from a German drama. In the last scene of the last act of Lessing's *Nathan the Wise*, for instance, Nathan reveals that Recha and the Templar are siblings (see appendix A for the relevant segment). In this case, the classification goal would be to assign the class `siblings`.
- **Experiment 2: Textual entailment**: Textual entailment recognition (TER), also known as natural language inference (NLI), is a task that has been established in NLP since 2005 (Dagan et al., 2006). It refers to the ability of a model to determine whether a hypothesis is entailed by another sentence or short text. As part of this task, we reformulated the classification task from Experiment 1 so that it

becomes an entailment task. The text snippet from the play becomes the premise, and the family relationship it conveys is formulated as a proposition that serves as the hypothesis (e.g. "Iphigenia is the child of Agamemnon.").

4 Data

We make use of one pre-existing dataset, a corpus of knowledge transfer annotations on German theatre plays pertaining to family relations (Andresen et al., 2022). The authors understand knowledge in a broad way to be beliefs that are thought to be true by a certain character at a certain point in time but might later turn out to be wrong during the advance of the plot. The corpus contains 30 texts sampled from the German Drama Corpus (GerDraCor, Fischer et al., 2019). For each family relation that a character learns about, an annotation marks the source and target of the knowledge transfer, which family relation is being transferred and who is part of this relation, as well as additional, optional properties like lies or uncertainty. While the dataset in total contains 1277 annotated text passages, we removed the infrequent relation types as well as all annotations that do not represent knowledge transfer. This yields 89 annotations for our experiments, which are divided into four categories: *parent of* (29), *child of* (26), *siblings* (23) and *spouses* (11). While this is a rather small dataset, it is suitable for our premise of evaluating a typical CLS task, since data sparsity is a common — and perhaps inherent — feature of CLS tasks, as also mentioned earlier. The same dataset is used for Experiments 1 and 2, whereby in the case of Experiment 2, 39 instances are changed so that they are classified as propositions that are not entailed by the text snippet. For instance, if an annotation contains a *parent of* relationship, the proposition is changed to "[Character X] is not the parent of [Character Y]".

5 Experiments

5.1 Models

For deciding on which model to use, we looked at the rankings of the HuggingFace LLM leaderboard[2] and chose the top three performing models for their performance on the HellaSwag benchmark[3] (Zellers et al., 2019). Since HellaSwag contains common sense inferences, it appeared to be the benchmark most closely related to our task. Thus, we compare three different LLM architectures, namely the open source models LLAMA 2 (Touvron et al., 2023), in a version optimized for chatting, Platypus2 (Lee et al., 2023), which is a derivative of LLAMA 2, and the closed source model GPT-4 (OpenAI, 2023).[4]

5.2 Experimental Setup

We test different model sizes, namely LLAMA 2 7B and 13B, as well as Platypus2 7B and 13B and compare their best results to GPT-4 with roughly 1.8T parameters. In addition to the sentence containing an annotation, we provide the models with context sentences and experiment with one and two context sentences on each side of the target sentence. We experiment with different prompts and prompt formats by utilizing the structures used for initially instructing the LLMs (LLAMA 2 and Alpaca prompt formats) as well as re-formulating in order to entice the model to generate a certain output format. In particular, we test the effects of providing the models with the names of the characters involved in the relationship (w/ character name), or not (wo/ character name). The specific prompt setups we used are documented in the appendix (see C.2). Lastly, we experiment with zero shot and few shot learning and test the effects on model performance. For the few shot setups, we provide the models with four examples chosen randomly from the development set. The code to re-run all experiments can be found on https://zenodo.org/doi/10.5281/zenodo.10581289.

6 Results

Table 1 shows the best results by each model architecture for predicting family relations.[5] The best F1 and accuracy scores with values of 66% and 68% in a zero shot setting and of 68% and 66% in a few shot setting were obtained with LLAMA 2 (13b) based on prompts in which the names of the characters whose family relationship is transmitted are input to the model. They thus achieve higher performance scores than are reported for the BIG-bench benchmark (49% accuracy) and are also at the upper end of the scores achieved by

[2] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[3] Effective September 13, 2023.

[4] We believe that proprietary models such as OpenAI's GPT-4 are not suitable as a valid basis for scientific experiments due to their opacity, but can serve as a good benchmark for comparison with open source models due to their performance.

[5] For a complete list of results see Table 3 in the appendix.

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.

Ziems et al. (2023, p. 14) (58 % to 64 %). Note, however, that Ziems et al. did not aim at the classification of implicit meta-knowledge. Nevertheless, the results are still below the scores achieved for other text-classification tasks with smaller pretrained language models (PLMs). Platypus2 and GPT-4 are generally outperformed by LLAMA 2, but achieve results similar to the baselines established by Hoffmann et al. (2022) and Ziems et al. (2023).

Table 2 shows the results for the textual entailment task. The data set for this task consists of 89 text-sentence pairs, of which 50 sentences are entailed from the snippets of the drama texts and 39 are not entailed. We compare the models to a baseline that classifies all instances as "entailed" (majority baseline). All three models perform poorly at the textual entailment task: While GPT-4 achieves an accuracy of 56% (baseline: 56%), LLAMA 2 classifies a larger number of sentences as not entailed, but only has a recall of 45% and therefore achieves F1 and accuracy scores far below the baseline. Platypus2 performs worst and classifies all instances as "not entailed".

LLAMA 2 even explicitly justifies why it classifies actually entailed sentences as non-implied, as the following example shows: "The text does not explicitly state that Iphigenie and Orest are siblings. While it mentions "Schwestern" (sisters) and "Bruder" (brother) in the same sentence, it does not explicitly state their relationship. Therefore, the proposition that Iphigenie and Orest are siblings is not entailed by the given text."[6]

7 Discussion

Our experiments show that the performance of an LLM is strongly affected by its potential lack of 'understanding' of the task, or particular words in the prompt as well as minor changes in the prompt templates. This is insofar not surprising, as the dramas date from the period 1750-1910 – with the majority dating from the turn of the 18th and 19th centuries – and therefore use a language that is also literary in style and in some cases versified, for which there were most likely no examples in the training data of the models. Recent studies investigating ICL argue that both semantic priors and input-label mapping in the prompts influence the ICL competence of LLMs, although the latter only applies to very large LLMs (Zhao et al., 2021; Xie et al., 2021; Wei et al., 2023). With regard to our experiments, one can assume that in the training data of LLAMA 2 the connection between the classification of an implicit knowledge transfer of a family relation and the fact that the same text snippet entails this family relation formulated in the form of a proposition, which is self-evident for a human speaker, was not represented.

8 Conclusion

It is worth noting that one of the dangers of ICL (and generative models in general) is the seemingly straightforward use of their output. We believe that using natural language output of such models is a regression compared to properly defined, symbolic output, as it requires interpretation and naturally

[6] As experiments have shown, in the case of LLAMA 2 this also applies sometimes to very simple material inferences. For example, when asked whether "Peter is taller than Fritz" implies that "Fritz is smaller than Peter", LLAMA 2 answers: "To entail the latter proposition, the text would need to explicitly state that Fritz is smaller than Peter [...]" GPT-4 does classify these sentences correctly.

contains ambiguities and terminological vagueness. "The worst dangers may lie in the humanist's ability to interpret nearly any result" (Sculley and Pasanek, 2008, 409), and this holds even more when the result comes in the form of natural language.

It also follows from our experiments that an unreflected and generic out-of-the-box use of ICL – even with open-source LLMs – for the automation of analytical sub-steps in the CLS is not yet recommended. The accuracy and F1-scores, although respectable per se, are still too low for this. Methodologically, it follows that for each task-specific use of ICL in the CLS, it must be clarified in each specific case whether and how the selected LLMs represent the subject-specific vocabulary for this task validly and with a high degree of accuracy. For certain tasks, it should also be considered whether the breakdown of a complex concept into simpler everyday concepts potentially mastered by the model – an LLM-specific operationalization of the complex concepts – does not achieve better scores.

In the background of this methodological recommendation lie the following open research questions: Is it really the case that LLMs perform better with ICL if the models already have semantic prior knowledge of the task-specific concepts/vocabulary? What does this mean for the domain-specific vocabulary of CLS? Can this be generically categorized in such a way that a distinction is made between everyday concepts, which are likely to be represented by LLMs or can be represented in principle, and complex concepts, which are likely to be difficult to represent by LLMs? How suitable are more recent low-resource 'fine-tuning-methods' such as PEFT (Lester et al., 2021) or LoRA (Hu et al., 2021) for CLS? The highly successful instruction-fine-tuning paradigm is rarely applicable in CLS due to a lack of available data, but alternatives such as PEFT have so far only been tested on standard NLP benchmarks like SuperGLUE. The extent to which these methods increase the accuracy of LLMs in CLS tasks will have to be examined in the future.

Acknowledgements

The research in this paper has been carried out within the Q:TRACK project funded by the German Research Foundation (DFG) in the context of SPP 2207 *Computational Literary Studies*. We thank the foundation for making this possible. We would also like to thank Melanie Andresen and Benjamin Krautter as well as the anonymous reviewers for their valuable and helpful feedback.

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.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Katrin Dennerlein, Thomas Schmidt, and Christian Wolff. 2023. Computational emotion classification for genre corpora of German tragedies and comedies from 17th to early 19th century. *Digital Scholarship in the Humanities*, 38(4):1466–1481.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Publisher: arXiv Version Number: 2.

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, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable corpora: Introducing

DraCor, an infrastructure for the research on European drama. In *Proceedings of DH2019: "Complexities"*.

Frank Fischer, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, and Peer Trilcke. 2017. Network dynamics, plot analysis. approaching the progressive structuration of literary texts. In *Book of Abstracts of the DH2017 conference*.

Hans Ole Hatzel, Haimo Stiemer, Chris Biemann, and Evelyn Gius. 2023. Machine learning in computational literary studies. *it - Information Technology*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. Publisher: arXiv Version Number: 2.

Fotis Jannidis, editor. 2022. *Digitale Literaturwissenschaft: DFG-Symposion 2017*. Germanistische Symposien. J.B. Metzler, Stuttgart.

Benjamin Krautter. 2018. Quantitative microanalysis? Different methods of digital drama analysis in comparison. In *Book of Abstracts of DH 2018*, pages 225–228.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of LLMs.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. Publisher: arXiv Version Number: 2.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

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.

Franco Moretti. 2011. Network theory, plot analysis. *Pamphlets of the Stanford Literary Lab*, 2:2–11.

OpenAI. 2023. GPT-4 technical report.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Nils Reiter, Marcus Willand, and Evelyn Gius. 2019. A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics*, 4(3).

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage?

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

D. Sculley and Bradley M. Pasanek. 2008. Meaning and mining: the impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, 23(4):409–424.

Aarohi Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. Publisher: arXiv Version Number: 5.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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 and Sewon Min. 2021. How does in-context learning work? a framework for understanding the differences from traditional supervised learning. `http://ai.stanford.edu/blog/understanding-incontext/`.

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.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of International Conference on Machine Learning 2021 (ICML)*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science?

A Example Knowledge Transfer

Excerpt from Lessing's *Nathan the Wise*. English translation by W. Taylor.

> NATHAN. He called himself Leonard of Filnek, but he was no German.
>
> TEMPLAR. You know that too?
>
> NATHAN. He had espoused a German, And followed for a time your mother thither.
>
> TEMPLAR. No more I beg of you—But Recha's brother—
>
> NATHAN. Art thou
>
> TEMPLAR. I, I her brother—
>
> RECHA. He, my brother?

This segment has been annotated by Andresen et al. (2022) with the following predicate:

```
transfer(nathan, saladin,
siblings(tempelherr, recha))
```

B Complete Results

Table 3 shows the complete results for Experiment 1.

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

Table 3: Complete results for Experiment 1.

C Prompts

C.1 Used Prompts

C.1.1 Experiment 1: LLAMA 2

```
<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]
```

Listing 1: "Zero shot prompt template w/o person; v2"

```
<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]
```

Listing 2: "Zero shot prompt template w/ person"

C.1.2 Experiment 1: Platypus2

```
Instruction: You are a literary scholar.
What is the family relation in the German text {drama_snippet}?
The possible family relations are parent, child, uncle, siblings, cousins.
Answer in a single sentence in the following format: The family relation is >>correct family relation<<.
Do NOT write code.
Do NOT write anything before or after the answer sentence.
```

Listing 3: "Zero shot prompt template w/ person"

```
Instruction: You are a literary scholar.
What is the family relation between {
    person1} and {person2} in the German
     text {drama_snippet}?
The possible family relations are parent
    , child, uncle, siblings, cousins.
Answer in a single sentence in the
    following format: The family
    relation between {person1} and {
    person2} is >>correct family
    relation<<.
Do NOT write code.
Do NOT write anything before or after
    the answer sentence.
Response:
```

Listing 4: "Zero shot prompt template w/o person"

C.1.3 Experiment 2: 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]
```

Listing 5: "Textual Entailment prompt "

C.1.4 Experiment 2: Platypus2

```
<s>[INST]

A text T textually entails a proposition
     P, iff typically, a human would be
    justified in reasoning from the
    propositions expressed by T to the
    proposition expressed by H.

Is the proposition {proposition}
    entailed by the following piece of
    German text: {text}?
Answer with:
a.) Yes, the proposition is entailed by
    the given text.
b.) No, the proposition is not entailed
    by the given text.
Your answer:
[/INST]
```

Listing 6: "Textual Entailment prompt "

C.1.5 Experiment 2: GPT-4

```
Common sense reasoning exam
###
Explain your reasoning in detail than
    answer with "Yes, the proposition is
     entailed by the given text" or "No,
     the proposition is not entailed by
    the given text".
Your answer should follow this 4-line
    format:

Premise: <some sentences from a German
    play>.
Question: <question requiring logical
    deduction>.
Reasoning: <an explanation of what you
    understand about the possible
    scenarios>.
Answer: <"Yes, the proposition is
    entailed by the given text" or "No,
    the proposition is not entailed by
    the given text">.

###
Premise: German {text}
Question: {proposition}
Reasoning: Let's think logically step by
     step.
Answer:
```

Listing 7: "Textual Entailment prompt

C.2 Different Prompting Setups

C.2.1 LLAMA 2

- Use of the Llama-specific prompt templates:
  - A prompt opens with the tags <s> [INST] and ends with [/INST]. A complete user/-model interaction is contained between the <s> and </s> tags.
- Enumeration of possible labels in a sentence vs. declared list
  - Prompt Template Version 1 (v1): “Choose one of "parent_of", "child_of", "siblings", "spouses".” vs. Prompt Template Version 2 (v2): Choose one of the following labels cf. Ziems et al. (2023, p. 12):

    A : "child_of"
    B : "parent_of"
    C : "siblings"
    D : "spouses".
- Instructions for generating desired output
  - *JUST name the label, do NOT generate any more text!*

C.2.2 Platypus

- Use Alpaca-specific prompt template:
  - A prompt with *Instruction* and *Response* directives
- Instructions for generating desired output
  - *Do NOT output anything after the family relation*
  - *Do NOT output programming code*
- Inserting information about the characters in a family relation
  - *identify the type of family relation and the characters involved* vs. *identify the type of family relation between person {person1} and person {person2}*

C.2.3 GPT-4

- Here we follow the OpenAI prompting principles as taught in the prompting course with Deeplearning.ai.
  - Give the model a role: "You are a literary scholar. ".
  - Use of delimiters: ###.
  - Asking for structured output: "JUST name the label without quotation marks and nothing else!"