https://doi.org/10.6092/issn.1825-1927/20421 ISSN 1825-1927

# From human-in-the-loop to LLM-in-the-loop for high quality legal dataset

Irina Carnat [1 2 3] Giovanni Comandé [1 2 3] Daniele Licari [2 3] Chiara De Nigris [4]

[1] LIDER-Lab
[2] DIRPOLIS
[3] Scuola Superiore Sant'Anna, Pisa, Italia
[4] Smartlex

**Abstract**.

Annotating legal documents with rhetorical structures is difficult and time-consuming, especially if done completely manually. This paper explores two methodologies for optimal results: first, a human-in-the-loop approach based on a multi-step annotation process with domain experts reviewing and revising datasets iteratively. To enhance interpretability, eXplainable Artificial Intelligence (XAI) models are incorporated, aiding in understanding decision-making processes. Second, an LLM-in-the-loop method has humans leveraging generative large language models (LLMs) to assist experts by automating repetitive annotation tasks under supervision. Further research is proposed to develop interaction models that effectively balance automation with human guidance and accountability.

**Keywords**:

Natural Language Processing, Data Annotation, Large Language Models, Accountability, Prompt Engineering, Generative AI.

## 1 Introduction

It is well established that the quality of machine learning models is highly dependent on the quality of the training data. The presented methodology places practitioners' and legal experts' knowledge at the center of qualitative analysis processes to obtain a high-quality annotated dataset. Pre-trained language models offer great opportunities in the domain of Natural Language Processing (NLP). However, the legal domain presents unique challenges, thus requiring more tailored solutions for complex tasks[1]. As such, training machine learning models on datasets annotated with rhetorical roles is crucial for extracting information from legal documents. Legal writing contains complex argument structures serving specific purposes like providing evidence

irina.carnat@santannapisa.it (Irina Carnat); giovanni.comande@santannapisa.it (Giovanni Comandé); d.licari@smartlex.eu (Daniele Licari); c.denigris@smartlex.eu (Chiara De Nigris);
https://orcid.org/0009-0007-3587-3847 (Irina Carnat); https://orcid.org/0000-0003-2012-7415 (Giovanni Comandé); (Daniele Licari); (Chiara De Nigris);

1. Daniele Licari and Giovanni Comandè, 'ITALIAN-LEGAL-BERT: A Pre-Trained Transformer Language Model for Italian Law', *EKAW'22: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management* (2022). See also Ilias Chalkidis and others, 'LEGAL-BERT: The Muppets Straight out of Law School' in Trevor Cohn, Yulan He and Yang Liu (eds), *Findings of the Association for Computational Linguistics: EMNLP 2020* (Association for Computational Linguistics 2020) <https://aclanthology.org/2020.findings-emnlp.261> accessed 11 December 2024.

or concluding arguments[2]. However, manually identifying these rhetorical components is time-consuming and difficult even for experts, as thoroughly discussed in paragraph 2.

Recent advancements in large language models (LLMs) and transformer architecture[3] have drawn attention to the different prompting techniques for task-specific purposes[4]. Such an approach leveraging the generative capacity of modern GPT-based[5] language models is worth exploring for data annotation purposes.

In this paper, we will first illustrate the methodology of a multi-step annotation process based on a human-in-the-loop approach for enhanced protocols for rhetorical roles annotation and a high-quality training dataset for legal sentences. In a second moment, we will address both the advantages and disadvantages of such a methodology. Finally, we will present early results of a new methodological approach that places greater emphasis on a human-LLM collaboration to achieve the same desired results in terms of enhanced annotation protocols and high-quality training dataset while realizing significant time savings and increasing the overall efficiency of the annotation process.

As a methodological note, we do not delve here into considerations of why annotation is needed but rather focus on how the generative LLM capabilities can be leveraged to assist the process. Through proper prompting techniques, LLMs can automate straightforward annotations to minimize the disadvantages of human annotation we have encountered from experience. Yet human experts still provide training data, validate results, and make final decisions for a balanced approach between automation and accountability.

The goal, as presented in the final discussion section, is to determine an optimal hybrid approach where LLMs efficiently handle routine annotation tasks and provide useful feedback for quality checks while experts guide the overall methodology and handle complex qualitative annotations. We propose further research into prompt engineering techniques and human-AI system collaboration that allow LLMs to assist with beneficial automation under expert supervision, improving annotation protocols and dataset quality.

# 2 Methodology

## 2.1 Human in-the-loop multi-step annotation process

The proposed qualitative model is composed of a setup phase consisting of the definition of tasks, labels, pilot cases, and sample selection, followed by an iterative phase of the quantitative and qualitative refinement of the dataset.

Figure 1 describes a human-in-the-loop diagram of the annotation rounds through a multi-step process, described as follows:

- Step 1 - Task definition. It consists of the definition of the number of rhetorical roles, legal subjects, and pilot projects.
- Step 2 - Sample selection. It consists of a robust selection of legal cases from the database.
- Step 3 - Guidelines Development. It consists of the drafting of the annotation protocols and subsequent amendments.

2. Gabriele Marino and others, 'Automatic Rhetorical Roles Classification for Legal Documents Using LEGAL-TransformerOverBERT', *Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023)* (2023).

3. Ashish Vaswani and others, 'Attention Is All You Need' (arXiv, 5 December 2017) <http://arxiv.org/abs/1706.03762> accessed 14 May 2023. See also Anas Belfathi, Nicolas Hernandez and Laura Monceaux, 'Harnessing GPT-3.5-Turbo for Rhetorical Role Prediction in Legal Cases' (arXiv, 26 October 2023) <http://arxiv.org/abs/2310.17413> accessed 11 December 2024.

4. Vaswani and others (n 3). See also Rui Wang and others, 'Self-Critique Prompting with Large Language Models for Inductive Instructions' (arXiv, 23 May 2023) <http://arxiv.org/abs/2305.13733> accessed 2 August 2023. Takeshi Kojima and others, 'Large Language Models Are Zero-Shot Reasoners'. Vern R Walker and others, 'Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning'.

5. OpenAI, 'GPT-4 Technical Report' (2023) <https://cdn.openai.com/papers/gpt-4.pdf> accessed 12 May 2023.

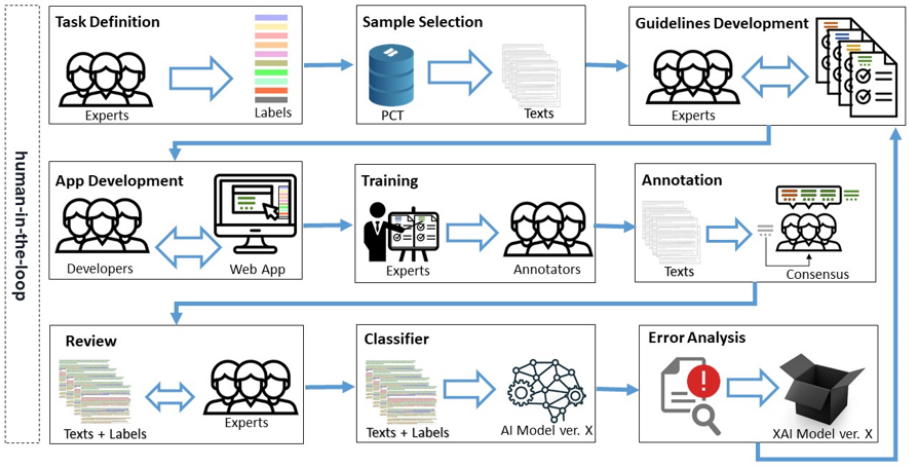



***Figure ? Human-in-the-loop diagram of the six labelling rounds. Step 1 Task definition**, definition of the number of rhetorical roles. **Step 2 Sample selection**, selection of legal cases from the database. **Step 3 Guidelines Development**, initial definition of the annotation protocols and the subsequent amendments. **Step 4 App Development**, converting the legal cases selected in LSJson format and set up of UI for the annotation phase. **Step 5 Training**, preparation phase of the annotators on the identification of the rhetorical roles using the annotation protocols and the use of the UI. **Step 6 Annotation**, the first two of three annotation steps conducted respectively by the Law students and senior lawyers. **Step 7 Review**, last step of the annotation phase, review of annotated sentences by the steering lawyers. **Step 8 Classier**, training and testing phase of the classification model. **Step 9 Error Analysis**, review prediction errors using XAI model. For each round of labelling, after the error phase, it cycles back to Step 3 to improve annotation protocols and the subsegments steps with the aim to create a gold standard annotation protocols and improve the performance of the final classification model of rhetorical roles*

Human-in-the-loop annotation process.

- Step 4 - App Development. This step involves converting the legal cases selected in LSJson format and setting up a Web User Interface for the annotation phase[6].
- Step 5 – Training. This is the preparation phase of the annotators on the identification of the rhetorical roles using the annotation protocols and the use of the Web User Interface.
- Step 6 – Annotation. It consists of the manual annotation of the sentences of the documents conducted by law students, supervised by senior jurists on the selected legal subjects[7]. A consensus through a majority vote is eventually employed to determine the final labeling for each sentence.
- Step 7 – Review. This is the last step of the annotation phase, which consists of reviewing the annotated sentences by the expert team. We utilize inter-rater agreement as a metric to gauge the quality of annotations, considering the comments provided by our annotators.
- Step 8 – Classifier. In this stage, training and testing of the classification model is carried out. It is important to note that the primary focus is not placed on maximizing the model's performance. Rather, emphasis is given to prioritizing interpretability and transparency over complexity, with the utilization of a relatively straightforward model, such as TFIDF, combined with logistic regression, especially when the primary goal is to identify potential annotation errors rather than achieving state-of-the-art performance.
- Step 9 - Error Analysis. It consists of a review of prediction errors using Explainable AI (XAI). We utilize local XAI models like LIME[8] or Anchors[9] to aid in identifying new lexical cues that can be

6. Walker and others (n 4).
7. Vern R Walker, 'The Need for Annotated Corpora from Legal Documents, and for (Human) Protocols for Creating Them: The Attribution Problem' 7.
8. Riccardo Guidotti and others, 'A Survey Of Methods For Explaining Black Box Models' [2018] arXiv:1802.01933 [cs] <http://arxiv.org/abs/1802.01933> accessed 20 December 2021.
9. Ian Covert, Scott Lundberg and Su-In Lee, 'Explaining by Removing: A Unified Framework for Model Explanation' (arXiv, 12 May 2022) <http://arxiv.org/abs/2011.14878> accessed 3 November 2023.

incorporated into the annotation protocol. This iterative process is essential for enhancing the overall quality of the protocol.

For each round of labeling, after the error phase is completed, the annotation process cycles back to Step 3 to improve annotation protocols and the subsegments steps with the aim to create a 'gold standard' annotation protocol and annotated dataset and improve the performance of the final classification model of rhetorical roles.

The annotation process consists of three rounds of labeling. At each round, the dataset is expanded and refined through a process of review and debugging using standard explainable artificial intelligence techniques. This iterative process is described as follows. Note from the outset that the goal is not to optimize the AI model's performance but to improve the quality of the final Dataset.

The Actors involved are:

- Level 1 annotators (Lv1), who are in charge of the first-round data annotation.
- Level 1+ annotators (Lv1+), who carry out a supervisory role overseeing the Lv1 annotators.
- Level 2 annotators (Lv2), domain experts who evaluate and correct any classification and interpretation errors at each annotation round.
- Data Scientists, who are responsible for training models on each annotated dataset and reporting results to domain experts.

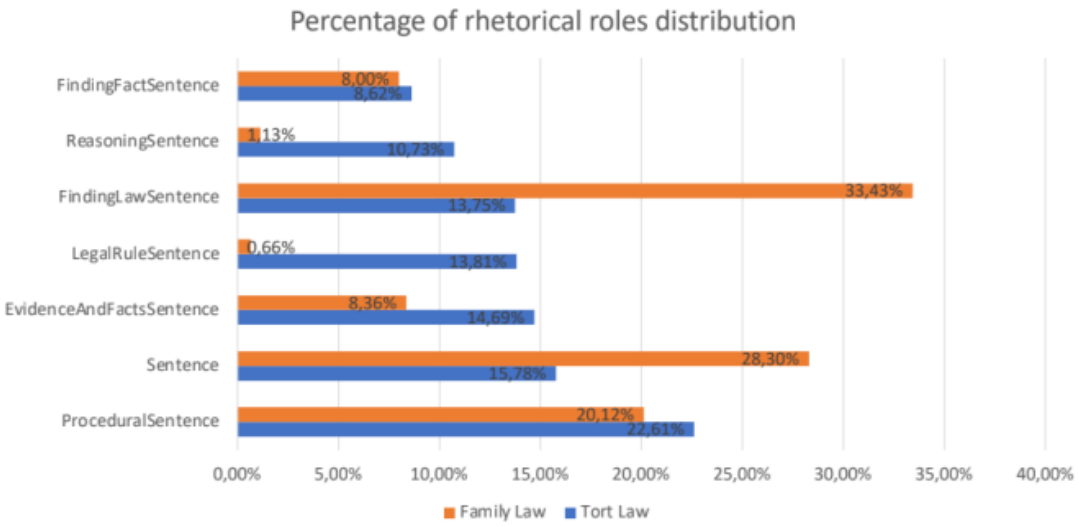
The D database is stratified into three subsets {D1, D2, D3}, each incrementally representing 15%, 35%, and 50% of the dataset, respectively. This incremental division has been chosen deliberately to enhance the robustness of our approach, enabling the application of a more comprehensive and validated protocol across a broader range of data. Each has undergone a specific human-in-the-loop annotation cycle, as follows.

- Cycle1 on D1 (18 documents): Firstly, Lv1 annotates the data in D1 under the supervision of Lv1+. Data Scientists then train a model (M1) on the annotated data (D1). Domain experts (Lv2) review the model and identify any classification and interpretation errors. Lv2 will then improve the annotation protocols based on these problems. If there are any errors in the annotations, Lv2 will then re-annotate D1 following the improvements in the annotation protocols.
- Cycle2 on D2 (40 documents): The second cycle follows a similar pattern to the first but now with the second subset of data, D2. Lv1 once again carries out the first data annotation on D2, having the annotation protocols been improved. Data Scientists now train a new model (M2) on both datasets: the already annotated D1 and newly annotated D2. Lv2 again identifies and fix any classification and interpretation mistakes in the model. They also enhance the annotation protocols based on this round of review. Lv2 then re-annotates both datasets, D1 and D2, correcting any errors that were made in the previous annotations.
- Cycle3 on D3 (60 documents): In the third cycle, Lv1 begins by annotating the data in the third and final dataset, D3. Data scientists then train a third model (M3) on all three datasets: D1, D2, and the recently annotated D3. Lv2 as usual, pinpoints any classification and interpretation mistakes and enhances the protocols for future annotations. After this, they will re-annotate all the datasets D1, D2, and D3 for any errors in earlier annotations.

It is worth explaining that for accountability purposes, logs of the annotated dataset are stored after each round to enable tracking of change. By preserving dataset versions throughout expert refinement, detailed records are maintained for full traceability and auditing of annotations as they improve through the iterative process.

## 2.2 Results

The adopted methodology allowed us to achieve promising results using a relatively small dataset of a total of 118 legal judgments (at the end of the third annotation round) from the Italian jurisdiction in the domains of tort law, specifically personal injury and family law.




Percentage of rhetorical roles distributed among the annotated legal decisions.

We employed cross-validation for both training and model evaluation. Subsequently, in the error analysis stage, we harnessed the Anchors technique for local explainability[10]. This method provided Lv2 annotators with valuable insights into potential reasons behind false positives and false negatives for each rhetorical role type, as depicted in Figure 3. Moreover, the use of confusion matrices, as shown in Figure 4, allows us to identify mismatches between human annotations and model predictions and to further enhance the clarity of the annotation protocols for optimal results.

Our methodology employed a total of 6 annotators, 3 at Lv1, 2 at Lv1+, and 1 acting as domain expert at Lv2. Based on the results and Error Analysis after each cycle of annotation, two revisions of the annotation protocols were required in total, gathering consensus among all the annotators, with a significant increase in the model's predictive accuracy.

## 2.3 Advantages

This iterative multi-step annotation process has multiple advantages.

First, the creation of a relatively small yet high-quality training dataset. Once the iterative cycles are completed, the annotated data from D1, D2, and D3 can be compiled and used to train the final model. Although the total dataset D may be small compared to the data sizes used in many modern deep learning models, the focus here is on precision over scale. By concentrating on refining a smaller dataset through extensive qualitative checks, the resulting annotations have much higher accuracy and reliability compared to a lightly validated large-scale dataset. This emphasis on expert oversight and error correction helps ensure that the final dataset, though smaller in absolute terms, contains annotations of exceptional quality.

Second, the iterative error-analysis steps described in the multi-step annotation process allow for progressive refinement and enhancement of the annotation protocols. As anticipated, at each round, the domain experts (Lv2) review the model predictions and identify any classification or interpretation errors. These errors highlight issues in the current annotation protocols, which the experts can then address by modifying and improving the protocols before the next annotation cycle. By cycling through multiple rounds of annotation, model training, error analysis, and protocol refinement, the overall quality and reliability of the final protocols is substantially increased. The repeated expert review and correction of issues found during error analysis is key to creating robust, high-quality annotation protocols through an accountable and transparent process.

10. ibid.

**EvidenceAndFactsSentence - False Negative**

- ''Ed a tale proposito va tenuto conto del fatto che la posizione dirigenziale ricoperta permette allo stesso di usufruire di vari benefit tra cui l'auto aziendale, il cellulare, il contributo per spese di locazione, benefit propri di una posizione dirigenziale.'' (docID:70421 ,sentID: 101545317_7895406P31S5)
  - **LABEL** EvidenceAndFactsSentence
  - **PREDICTION** ReasoningEvidenceSentence (prob.: 82.27%)
  - **RULE**: IF conto, fatto, ricoperta, spese, cui, che, propri, dirigenziale, una, dirigenziale, va, usufruire, benefit, l', auto, il, di, a, proposito, tenuto, la, di, stesso THEN ReasoningEvidenceSentence
- ''Irrilevante il fatto che questi abbia mantenuto la propria residenza (v. doc. n° 7), avendo i certificati anagrafici, come noto, solo un valore presuntivo.'' (docID:163740 ,sentID: 101659000_15890844P41S7)
  - **LABEL** EvidenceAndFactsSentence
  - **PREDICTION** ReasoningEvidenceSentence (prob.: 97.51%)
  - **RULE**: IF Irrilevante, fatto, avendo, che, residenza, certificati, mantenuto, doc THEN ReasoningEvidenceSentence
- ''Percepisce € 500,00 per la pensione di invalidità di Nicolò, di cui però è autorizzata a prelevare solo 300,00 euro al mese e gestisce € 500,00 per la dote di cura di Nicolò, (somma che è comunque destinata non al mantenimento bensì ai bisogni del minore in relazione alla sua disabilità).'' (docID:171393 ,sentID: 101664205_16756644P11S8)
  - **LABEL** EvidenceAndFactsSentence
  - **PREDICTION** FindingLawSentence (prob.: 87.95%)
  - **RULE**: IF a, Percepisce, €, destinata, la, di, cui, ai, e, alla, pensione, autorizzata, 50000, comunque, di THEN FindingLawSentence

Examples of false negatives for the "EvidenceAndFactsSentence" label: the report provides the domain experts with the correct label, prediction, and prediction rule extracted from the model through XAI.

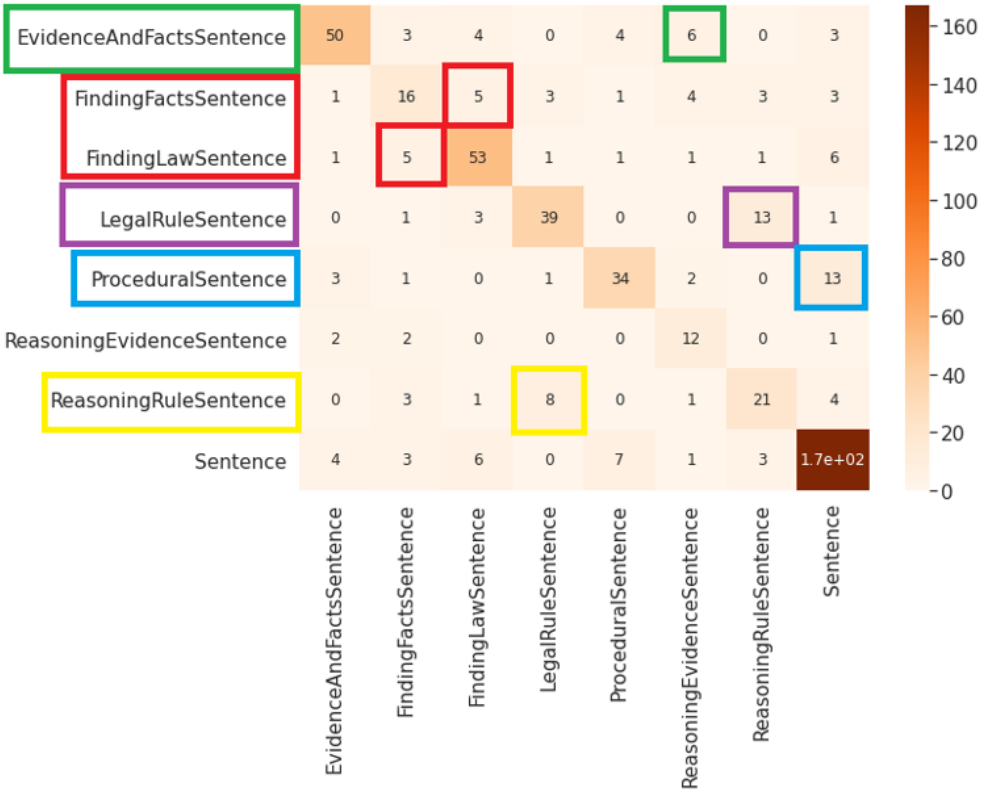
Third, the oversight and review conducted by the domain experts (Lv2) at each stage of the iterative annotation process helps ensure the overall accountability, transparency, and reliability of the final dataset. Their active participation adds a crucial layer of human judgement and domain knowledge to evaluate and enhance the annotation work. By manually checking the annotated sentences, correcting errors, and providing feedback to improve the protocols, the experts provide indispensable guidance and validation to the process. Their revisions after each round instill confidence that mistakes are being caught and addressed responsibly. Furthermore, by involving multiple levels of human annotators (Lv1, Lv1+, and Lv2), the process benefits from different perspectives to produce more balanced and well-validated results.

## 2.4 Disadvantages

Notwithstanding the above-mentioned benefits, such an iterative approach bears certain disadvantages worth exploring.

First, the Web User Interface (Web UI) for annotation introduces both benefits and drawbacks. While it is efficient for human labelers, reliance on a custom web platform requires extra preprocessing to convert data formats. The Web UI also creates a bottleneck if technical issues arise, hindering progress until resolved. Since the Web UI only enables annotation, further work is required to convert annotated data into formats needed for model training and deployment.

Second, the reliance on manual annotation is proven to be time-consuming. We estimated 0.6 full-time equivalent (FTE) for the whole annotation process. As described, the methodology relies on teams of annotators (Lv1, Lv1+) trained on the legal-relevant knowledge and also duly trained to manually label data based on the protocols. Hands-on training on the protocols and Web UI takes resources away from other productive tasks, especially from the domain experts (Lv2) in charge of training the annotators. Ongoing supervision and monitoring of annotators also adds additional management burden. Additionally, human annotation inherently does not scale as efficiently as automated approaches. Expanding the project or meeting tight deadlines becomes challenging with reliance on manual annotation.




Example of a confusion matrix for error analysis between human annotation and model prediction.

Third, ultimately, it is the domain experts who bear the final responsibility, incurring the risk of automation bias. The methodology depends heavily on the Lv2 team for identifying errors, revising annotations, and approving the final datasets. Despite their expertise, the Lv2 experts can still suffer from cognitive biases like automation bias, where excessive trust is placed in the original Lv1 annotations. In fact, the experts may grow accustomed to only minor revisions rather than thoroughly re-evaluating each case. Or they may be unconsciously influenced by the annotations they are reviewing instead of acting fully independently.

# 3 Enhancing the annotator with generative LLMs: early experiments and results

To further optimize the methodology for high-quality dataset creation, we propose incorporating generative LLMs into the annotation process[11].

To test the early results of such a methodological approach, we used GPT-4o by OpenAI through its API to annotate the legal dataset {D1+D2+D3}. The model selection was driven by two key factors: the excellent performances related to the latest GPT family models and the advantage of conducting experiments without the need for substantial computational resources. We used the most recent available version of GPT-4o (gpt-4o-2024-08-06), which, in addition to benefits related to efficiency and speed, supports requests for structured output in a customized format, thereby expediting automated output management. To achieve optimal results in annotating text with a generative model, adequate prompting is of paramount importance. For this reason, prompt writing was a highly iterative process, testing different approaches for legal data annotation, based on the annotation protocols developed by the legal tech scholarship[12]. We incorporated commonly recognized prompting techniques into our prompt engineering process,[13] with the goal of crafting succinct, precise, and clear instructions. The prompt was structured with a clear division between different parts, using special characters to indicate the separation of one instruction block from another. The language used was simple, precise, and assertive, avoiding negative commands and instead emphasizing desired behaviors.

Most promising results have been obtained with few-shot learning[14], integrating in the instruction two example sentences per label, crafted in the format of the desired output. This provides sufficient context for the LLM to understand the annotation task and expected output format. By giving practical examples of desired annotations, few-shot prompting allows the LLM to rapidly align to the annotation guidelines.

Regarding results evaluation, we employed the Cohen Kappa statistic to measure the inter-rater reliability between the LLM annotated dataset and the human-annotated dataset. With the final refined prompt, the score amounted to 0.53 for tort law judgments and 0.58 for family law judgments, which indicates a moderate level

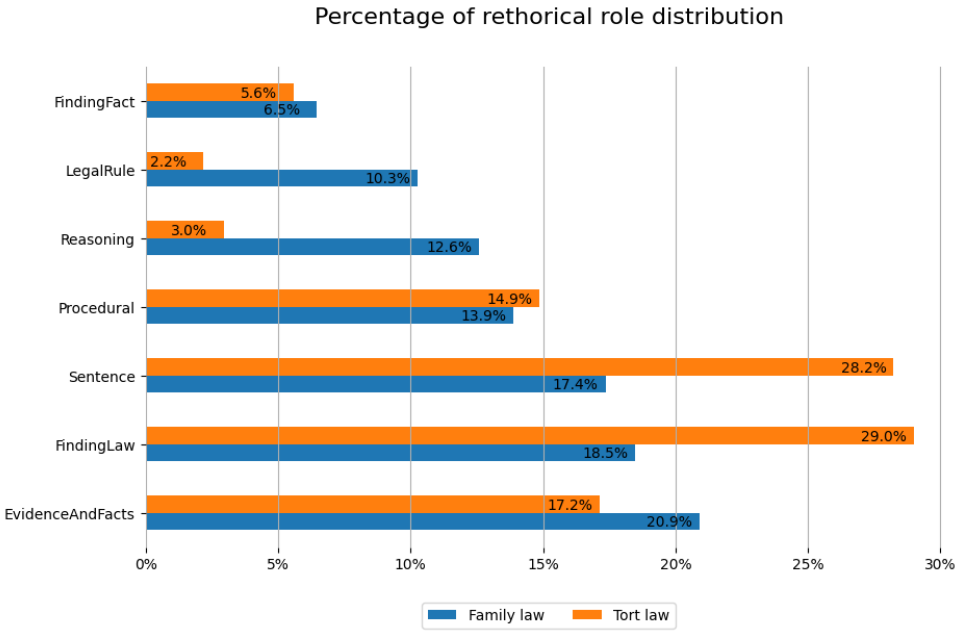
11. See, generally, Morgan Gray and others, 'Can GPT Alleviate the Burden of Annotation?' (2023) 379 Frontiers in Artificial Intelligence and Applications 157 <https://dsc.duq.edu/faculty/1396>; Zhen Tan and others, 'Large Language Models for Data Annotation and Synthesis: A Survey' (arXiv, 2 December 2024) <http://arxiv.org/abs/2402.13446> accessed 11 December 2024. See also Davide Liga and Livio Robaldo, 'Fine-Tuning GPT-3 for Legal Rule Classification' (2023) 51 Computer Law & Security Review 105864 <https://www.sciencedirect.com/science/article/pii/S0267364923000742> accessed 11 December 2024. See also Jaromir Savelka, 'Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts', *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (2023) <http://arxiv.org/abs/2305.04417> accessed 11 December 2024. See also Nataliia Kholodna and others, 'LLMs in the Loop: Leveraging Large Language Model Annotations for Active Learning in Low-Resource Languages' in Albert Bifet and others (eds), *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track* (Springer Nature Switzerland 2024). See also Surendrabikram Thapa, Usman Naseem and Mehwish Nasim, 'From Humans to Machines: Can ChatGPT-like LLMs Effectively Replace Human Annotators in NLP Tasks' (2023) 2023 Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media 15 <https://workshop-proceedings.icwsm.org/abstract.php?id=2023_15> accessed 11 December 2024.

12. Walker (n 7).

13. Sondos Mahmoud Bsharat, Aidar Myrzakhan and Zhiqiang Shen, 'Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4' (*arXiv.org*, 26 December 2023) <https://arxiv.org/abs/2312.16171v2> accessed 12 September 2024. See, generally, Sabit Ekin, 'Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices' <https://www.authorea.com/users/690417/articles/681648-prompt-engineering-for-chatgpt-a-quick-guide-to-techniques-tips-and-best-practices> accessed 12 September 2024.

14. Wang and others (n 4). See also Tom B Brown and others, 'Language Models Are Few-Shot Learners' (arXiv, 22 July 2020) <http://arxiv.org/abs/2005.14165> accessed 19 July 2023.

of agreement between the annotations. Further experiments and error analysis will contribute to the refinement of the prompting technique, as well as gathering insights for the optimization of the annotation protocols.




Percentage of rhetorical roles distributed among the LLM-annotated legal decisions.

## 3.1 Future developments: LLM-in-the-loop for legal data annotation

To address the disadvantages of the intensive training and oversight required for human annotators at Lv1 and Lv1+, we envision utilizing GPT-4o by OpenAI[15] at the Lv2 level for legal data annotation. By automating certain repetitive annotation tasks, these LLMs can reduce the annotation burden on human annotators. The domain experts in Lv2 will interact with the LLMs to develop and refine the annotation guidelines over multiple iterations with the LLMs for optimal annotation protocols.

Since generative LLMs are also trained to follow instructions and adapt based on human feedback[16], the domain expert can provide reinforcement and oversight by interacting with the LLM. As the expert reviews automated annotations, they can confirm correct labels and supply corrections to mistakes. This feedback loop enables the LLM to continuously improve annotation performance under the expert's supervision, enhancing efficiency. For increased accuracy, we also propose chain-of-thought prompting[17] to handle uncertainty between two labels. When unsure, the LLM will explain its reasoning for being hesitant between labels, much like confusion matrices and AI explainability techniques. This allows the domain expert to better understand the LLM's decisions and supply appropriate correcting feedback to resolve annotation ambiguities.

The overall multi-step annotation process remains fundamentally unvaried (see Figure 1), with human domain experts guiding the process and validating the results. However, generative LLMs now assist with routine labeling as directed through expert prompting. As such, the domain experts' oversight role is preserved, main-

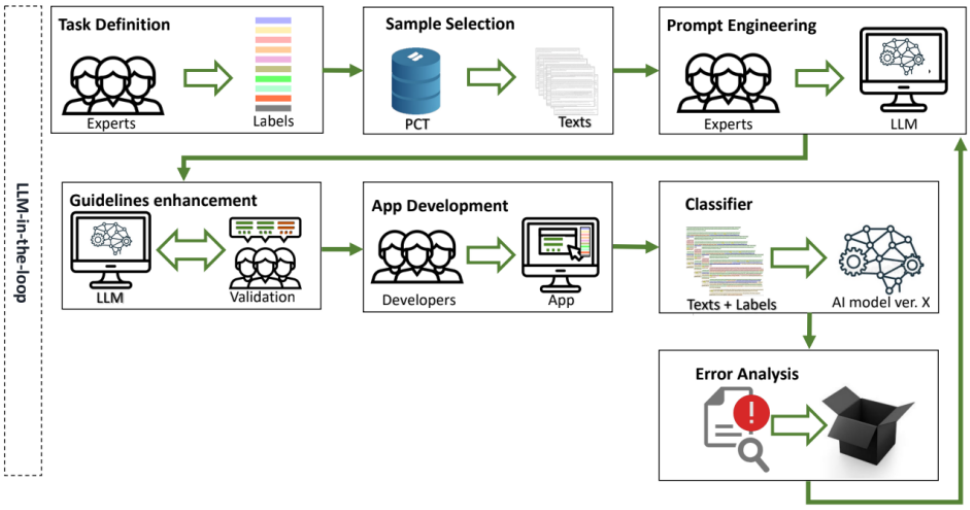
15. Daniel Schwarcz and Jonathan H Choi, 'AI Tools for Lawyers: A Practical Guide' (29 March 2023) <https://papers.ssrn.com/abstract=4404017> accessed 11 May 2023.

16. Long Ouyang and others, 'Training Language Models to Follow Instructions with Human Feedback' (arXiv, 4 March 2022) <http://arxiv.org/abs/2203.02155> accessed 12 May 2023. Aman Madaan and others, 'Self-Refine: Iterative Refinement with Self-Feedback' (arXiv, 25 May 2023) <http://arxiv.org/abs/2303.17651> accessed 21 July 2023.

17. Jason Wei and others, 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models' (arXiv, 10 January 2023) <http://arxiv.org/abs/2201.11903> accessed 31 May 2023.

taining accountability while decreasing training and monitoring demands. A tentative diagram for such an approach is depicted in Figure 6.

The core strength of this proposed methodology lies in the iterative interaction between the human domain expert and the generative LLM[18]. This collaboration enables continuous refinement of the annotation protocols without reliance on larger teams of trained human annotators. Specifically, only the domain expert (originally Lv2) is needed to guide the LLM through annotation, evaluation, validation, and protocol improvement. There is nonetheless an important error analysis phase: if the model's results are unsatisfactory, the process loops back to the prompting phase, as described in Figure 6. Here, the domain expert, perhaps assisted by a prompt engineer[19] or leveraging the LLM's own prompt generation capabilities[20], can refine the prompt to improve performance. As such, the feedback loop enables accountable expert oversight to optimize prompts until the LLM's annotations reliably meet standards.




Human-directed LLM-in-the-loop annotation process.

Another possible approach for future work could be to replicate the human-in-the-loop multi-step annotation process illustrated in 2.1, but integrate different LLMs with varying levels of capabilities and costs. The dataset would be similarly divided into parts, but in this case, the Lv1 annotators are cost-effective LLMs that are given as input the annotation protocol in its initial, unrefined version and a sentence to annotate. The model outputs the label to annotate the sentence and the reasoning behind its selection. In cases of disagreement, the output of the Lv1 LLM is then given as input to a more powerful LLM reviewer, asking it to assign a new label and further improve the protocol. At this point, the improved protocol will be used as input for the second round of annotation by the Lv1 LLM.

As with the human-in-the-loop approach, the final output of the process will be the entire annotated dataset and the improved protocol. The results obtained with this methodology can then be compared with those already obtained by automatically annotating the entire dataset with a single LLM, already exposed in paragraph 2.5.

18. Similar approaches were adopted by Xinru Wang and others, 'Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels', *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2024) <https://dl.acm.org/doi/10.1145/3613904.3641960> accessed 11 December 2024. See also Arbi Haza Nasution and Aytuğ Onan, 'ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks' (2024) 12 IEEE Access 71876 <https://ieeexplore.ieee.org/document/10534765/?arnumber=10534765> accessed 11 December 2024.

19. Aras Bozkurt and Ramesh Sharma, 'Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World' (2023) 18 i.

20. Wang and others (n 4). Yongchao Zhou and others, 'Large Language Models Are Human-Level Prompt Engineers' (2022) <https://openreview.net/forum?id=92gvk82DE-> accessed 2 August 2023.

## 3.2 Advantages

The advantages associated with such a human-directed LLM-in-the-loop approach are – indirectly – associated with the mentioned disadvantages of the human-in-the-loop approach supra at 2.4.

First, using LLMs for annotation automation eliminates the need for multi-step training of additional human annotators (Lv1, Lv1+). This removes time spent developing training programs, conducting annotator onboarding, monitoring work quality, and providing ongoing guidance. Instead, the single domain expert oversees the process end-to-end. This streamlines dataset creation through reliance on specialized expertise over crowd annotation.

Second, the expert interacts directly with the LLM to enhance protocols faster. With LLMs assisting, the annotation protocols can be refined through iterative collaboration solely between the model and domain expert. This avoids lengthy cycles of annotating sample datasets, identifying issues, and then re-annotating after improvements.

Third, dependence on the domain expert rather than larger annotation teams reduces risks of automation bias and knowledge domain divergence. The expert is directly accountable for results rather than potentially over-relying on annotator work. Only the expert's changes to prompts and protocols are logged, avoiding the storage of large intermediate datasets.

## 3.3 Disadvantages

Naturally, there are also disadvantages.

First, developing the initial annotation protocols solely through the domain expert can be limiting without the diversity of perspectives and feedback from a larger human annotation team. Important considerations may be overlooked without the debate and brainstorming that comes from collaborating with the wider Lv1 and Lv1+ annotator groups in the traditional methodology. However, we hope to be able to generalize the protocol to enable its application to different legal domains and topics in future work, which would constitute an important advancement.

Furthermore, employing generative models for structured tasks often presents significant challenges. A crucial factor in this context is the inherent difficulty of constraining output generation to predefined categories, essentially compelling a generative model to execute a structured classification task. While advanced models can yield promising results, it is imperative to anticipate the potential need for post-processing to ensure that the output aligns precisely with the required content and structure specifications.

Another relevant point is that expertise in prompt engineering is still rare and requires extensive self-learning at this stage. Properly structuring prompts and examples for optimal LLM performance remains more art than science. Attempting many prompt formulations to maximize accuracy is time-consuming and, without proper training in prompt design, the domain expert may struggle to construct effective prompts that translate their goals into high-quality annotations.

Finally, it is compulsory to consider also technical limitations related to the model adopted. One among all is that there could be constraints related to input and output lengths according to the model, which could be particularly relevant if working with long texts, forcing text division in different API calls. All these factors underline that integrating LLMs into the annotation workflow requires non-trivial software engineering efforts. The domain expert likely lacks the technical skills in LLM APIs, data pipelines, model deployments, and custom UIs needed for a usable system. Professional programming support would be imperative, adding cost and coordination requirements

Lastly, a crucial consideration pertains to the economic aspect. While proprietary models alleviate computational resource concerns, they potentially incur significant expenses. As previously highlighted, the implementation of generative AI models necessitates extensive trials and testing, thereby extending costs beyond mere model usage to encompass the entire experimental process. Such financial burdens may prove challenging for small and medium-sized enterprises (SMEs) or modest research facilities, potentially serving as a substantial constraint on their AI adoption and research capabilities.

# 4 Discussion and conclusions

This novel human-directed LLM-in-the-loop approach seems promising for enhancing legal data annotation but requires answering certain key questions before full-scale implementation.

First, how to design a new human-AI system interaction model? The methodology proposed leverages strengths of both human expertise and AI capabilities, but the collaboration framework requires careful structuring to enable beneficial symbiosis[21]. Humans remain accountable for results, but AI assists in amplifying effectiveness. Defining clear responsibilities and hand-off points between humans and machines is crucial. As such, workflows must be co-designed for smooth integration, and user interfaces should facilitate explainability and oversight to address the risk of automation bias.

Second, how to structure the system to account for possible risks from a risk management perspective? Despite advantages, risks around data quality, bias, and accountability must be addressed proactively[22]. As such, implementation protocols are needed for monitoring annotation quality and documenting AI behavior in view of complying with the proposed AI Act. Moreover, version control and logging help trace provenance and changes, along with regular audits by independent experts, can validate system operations. All this would enable a further experimental analysis exploring the possibility of segregating the legal and ethical risks of leveraging LLMs in developing LLM-based solutions since we can swiftly fulfill transparency and explainability requirements solicited by existing (e.g., GDPR) or forthcoming regulations (e.g., the AI Act) for local-specific training based on the LLM-in-the loop approach, indirectly identifying the ones generated by the original LLM used.

The latter consideration ties in with a third question: which criteria can be devised for quality checks? While multiple solutions can be envisioned, such as qualitative reviews of random sample annotations against gold standards, confusion matrices that highlight areas of uncertainty, user and LLM feedback, etc., more research is suggested in this area.

Fourth, what legal and ethical requirements exist for human oversight? Laws increasingly require transparency around AI use, including documenting development processes and algorithmic accountability[23]. This particularly applies to legal domains where stakes are high.

In conclusion, this methodology requires deliberative design weighing benefits against risks. With thoughtful development rooted in human primacy and sound ethics, LLMs offer transformative potential for amplifying legal work[24]. However, this emerging human-AI synthesis requires proactive shaping to maximize gains through responsible innovation. Further research and trials are still needed, but the possibilities merit investment in this fruitful direction.

**Bibliography**

Belfathi, Anas, Nicolas Hernandez and Laura Monceaux, 'Harnessing GPT-3.5-Turbo for Rhetorical Role Prediction in Legal Cases' (arXiv, 26 October 2023) http://arxiv.org/abs/2310.17413 accessed 11 December 2024.

Bozkurt, Aras and Ramesh Sharma, 'Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World' (2023) 18 i.

Brown, Tom B and others, 'Language Models Are Few-Shot Learners' (arXiv, 22 July 2020) http://arxiv.org/abs/2005.14165 accessed 19 July 2023.

21. Ben Green, 'The Flaws of Policies Requiring Human Oversight of Government Algorithms' (2022) 45 Computer Law & Security Review 105681 <https://www.sciencedirect.com/science/article/pii/S0267364922000292> accessed 19 October 2023.

22. Daniel J Solove and Hideyuki Matsumi, 'AI, Algorithms, and Awful Humans' (16 October 2023) <https://papers.ssrn.com/abstract=4603992> accessed 19 October 2023.

23. Rebecca Crootof, Margot E Kaminski and W Nicholson Price II, 'Humans in the Loop' (25 March 2022) <https://papers.ssrn.com/abstract=4066781> accessed 19 October 2023.

24. Harry Surden, 'The Ethics of Artificial Intelligence in Law: Basic Questions' (22 August 2019) <https://papers.ssrn.com/abstract=3441303> accessed 14 May 2023.

Bsharat, Sondos Mahmoud, Aidar Myrzakhan and Zhiqiang Shen, 'Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4' (arXiv.org, 26 December 2023) https://arxiv.org/abs/2312.16171v2 accessed 12 September 2024.

Chalkidis, Ilias and others, 'LEGAL-BERT: The Muppets Straight out of Law School' in Trevor Cohn, Yulan He and Yang Liu (eds), Findings of the Association for Computational Linguistics: EMNLP 2020 (Association for Computational Linguistics 2020) https://aclanthology.org/2020.findings-emnlp.261 accessed 11 December 2024.

Covert, Ian, Scott Lundberg and Su-In Lee, 'Explaining by Removing: A Unified Framework for Model Explanation' (arXiv, 12 May 2022) http://arxiv.org/abs/2011.14878 accessed 3 November 2023.

Crootof, Rebecca, Margot E Kaminski and W Nicholson Price II, 'Humans in the Loop' (25 March 2022) https://papers.ssrn.com/abstract=4066781 accessed 19 October 2023.

Gray, Morgan and others, 'Can GPT Alleviate the Burden of Annotation?' (2023) 379 Frontiers in Artificial Intelligence and Applications 157 https://dsc.duq.edu/faculty/1396;

Green, Ben, 'The Flaws of Policies Requiring Human Oversight of Government Algorithms' (2022) 45 Computer Law & Security Review 105681 https://www.sciencedirect.com/science/article/pii/S0267364922000292 accessed 19 October 2023.

Guidotti, Riccardo and others, 'A Survey Of Methods For Explaining Black Box Models' [2018] arXiv:1802.01933 [cs] http://arxiv.org/abs/1802.01933 accessed 20 December 2021.

Kholodna, Nataliia and others, 'LLMs in the Loop: Leveraging Large Language Model Annotations for Active Learning in Low-Resource Languages' in Albert Bifet and others (eds), Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track (Springer Nature Switzerland 2024).

Kojima, Takeshi and others, 'Large Language Models Are Zero-Shot Reasoners'.

Licari, Daniele and Giovanni Comandè, 'ITALIAN-LEGAL-BERT: A Pre-Trained Transformer Language Model for Italian Law', EKAW'22: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management (2022).

Liga, Davide and Livio Robaldo, 'Fine-Tuning GPT-3 for Legal Rule Classification' (2023) 51 Computer Law & Security Review 105864 https://www.sciencedirect.com/science/article/pii/S0267364923000742 accessed 11 December 2024.

Madaan, Aman and others, 'Self-Refine: Iterative Refinement with Self-Feedback' (arXiv, 25 May 2023) http://arxiv.org/abs/2303.17651 accessed 21 July 2023.

Marino, Gabriele and others, 'Automatic Rhetorical Roles Classification for Legal Documents Using LEGAL-TransformerOverBERT', Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023) (2023).

Nasution, Arbi Haza and Aytuğ Onan, 'ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks' (2024) 12 IEEE Access 71876 https://ieeexplore.ieee.org/document/10534765/?arnumber=10534765 accessed 11 December 2024.

Ouyang, Long and others, 'Training Language Models to Follow Instructions with Human Feedback' (arXiv, 4 March 2022) http://arxiv.org/abs/2203.02155 accessed 12 May 2023.

Savelka, Jaromir, 'Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts', Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (2023) http://arxiv.org/abs/2305.04417 accessed 11 December 2024.

Schwarcz, Daniel and Jonathan H Choi, 'AI Tools for Lawyers: A Practical Guide' (29 March 2023) https://papers.ssrn.com/abstract=4404017 accessed 11 May 2023.

Solove, Daniel J and Hideyuki Matsumi, 'AI, Algorithms, and Awful Humans' (16 October 2023) https://papers.ssrn.com/abstract=4603992 accessed 19 October 2023.

Surden, Harry, 'The Ethics of Artificial Intelligence in Law: Basic Questions' (22 August 2019) https://papers.ssrn.com/abstract=3441303 accessed 14 May 2023.

Tan, Zhen and others, 'Large Language Models for Data Annotation and Synthesis: A Survey' (arXiv, 2 December 2024) http://arxiv.org/abs/2402.13446 accessed 11 December 2024.

Thapa, Surendrabikram, Usman Naseem and Mehwish Nasim, 'From Humans to Machines: Can ChatGPT-like LLMs Effectively Replace Human Annotators in NLP Tasks' (2023) 2023 Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media 15 https://workshop-proceedings.icwsm.org/abstract.php?id=2023_15 accessed 11 December 2024.

Vaswani, Ashish and others, 'Attention Is All You Need' (arXiv, 5 December 2017) http://arxiv.org/abs/1706.03762 accessed 14 May 2023.

Vaswani and others (n 3). See also Wang, Rui and others, 'Self-Critique Prompting with Large Language Models for Inductive Instructions' (arXiv, 23 May 2023) http://arxiv.org/abs/2305.13733 accessed 2 August 2023.

Walker, Vern R, 'The Need for Annotated Corpora from Legal Documents, and for (Human) Protocols for Creating Them: The Attribution Problem' 7.

Walker, Vern R and others, 'Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning'.

Wang, Rui and others (n 4). See also Zhou, Yongchao and others, 'Large Language Models Are Human-Level Prompt Engineers' (2022) https://openreview.net/forum?id=92gvk82DE- accessed 2 August 2023.

Wang, Xinru and others, 'Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels', Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery 2024) https://dl.acm.org/doi/10.1145/3613904.3641960 accessed 11 December 2024.

Wei, Jason and others, 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models' (arXiv, 10 January 2023) http://arxiv.org/abs/2201.11903 accessed 31 May 2023.

Zhou, Yongchao and others, 'Large Language Models Are Human-Level Prompt Engineers' (2022) https://openreview.net/forum?id=92gvk82DE- accessed 2 August 2023.