SurveyDash Final Report

Jim Yang, Muhan Li, Yutong Wang, Hongyuan Hua

Dec. 10, 2024

Word Count: 2000

Jim
permission to post video: yes
permission to post final report: yes
permission to post source code: yes

Muhan
permission to post video: yes
permission to post final report: yes
permission to post source code: yes

Yutong
permission to post video: yes
permission to post final report: yes
permission to post source code: yes

Hongyuan
permission to post video: yes
permission to post final report: yes
permission to post source code: yes

1. Introduction

Survey papers are invaluable in academic research. They help researchers identify underlying trends in the field and provide a guide for future investigations. Writing these papers, however, is a complex and time-consuming task. This is due to the sheer volume of literature that must be reviewed and analyzed. Recent advancements in large language models (LLMs) have enabled them to efficiently summarize vast inputs and generate insightful analytical outputs. This capability positions LLMs as a powerful tool to support the creation of comprehensive survey papers. Thus, the goal of the project is to develop a multi-agent LLM system that harnesses the power of LLMs to generate cohesive, analytical, and comprehensive survey papers based on provided reference materials.

2. Illustration

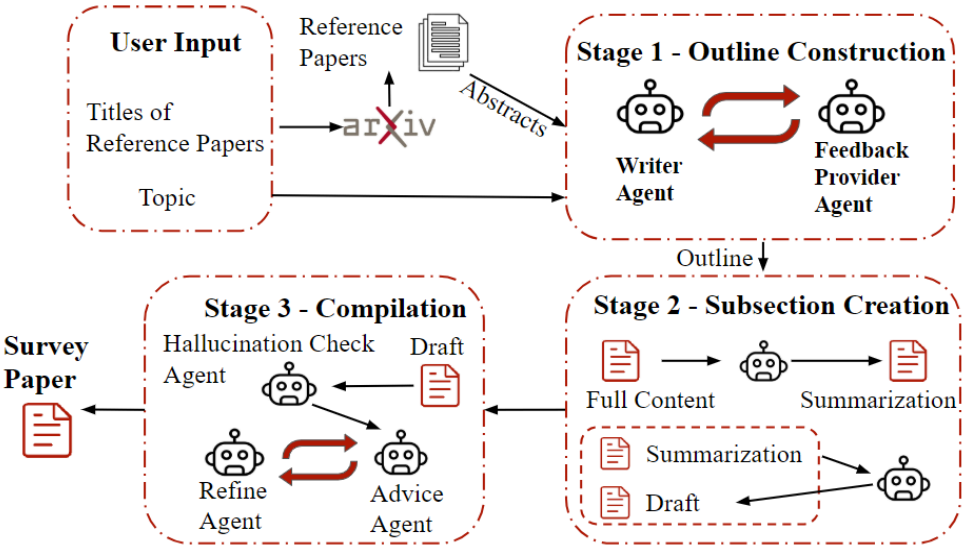

Figure 1. Architecture of System of SurveyDash

3. Background & Related Work

Automated Survey Creation. AutoSurvey [1] combines multiple specialized LLM agents to produce comprehensive academic survey papers, and its pipeline includes three stages. The first stage involves generating an outline using the top relevant papers' subsections titles and brief descriptions. The second stage involves LLM agents generating drafts for each section in parallel. The last stage involves LLMs editing each draft section for final refinements and hallucination checks. The refined sections are merged together to form the final paper.

Iterative Consensus. Diverse pre-trained models are assembled together via a closed-loop iterative consensus optimization to solve complex multimodal problems in a zero-shot manner [2]. This process involves a "generator" model that constructs proposals and a "scorer" model that iteratively provides feedback to refine the generated result. The feedback loop continues until a consensus is reached on the optimal solution, allowing for corrections of errors caused by other models. This methodology has demonstrated strong performance in

zero-shot settings and versatility across various applications.

4. Data and Data Processing

The data collection process consists of three steps. First, the team sets a research topic for the system to focus on. Then, the team manually selects an existing survey paper covering the chosen topic. We find that the reference list effectively covers different scopes within the field. The team then parses the reference list of the selected paper to extract the titles of the papers it cites, which yields a list of related papers. The initial data, therefore, consists of one research topic and the titles of the extracted related papers. This is the data that the user would provide to generate the final survey paper using SurveyDash.

Our target topic is the application of LLMs in the legal domain, and the selected paper [3] with 30 references. This reference list includes foundational papers on large language models like [4], as well as papers that apply large language models to the legal domain, such as [5]. These papers are important for understanding the topic.

To extract information from the reference papers, the team finds and downloads their contents from arXiv. The papers not found on arXiv are discarded, since most papers can be found on arXiv. The title, abstract, and main content for each reference paper is extracted and stored in a JSON file and is ready for further processing.

5. Architecture and Software

5.1 Stage 1 - Outline Construction

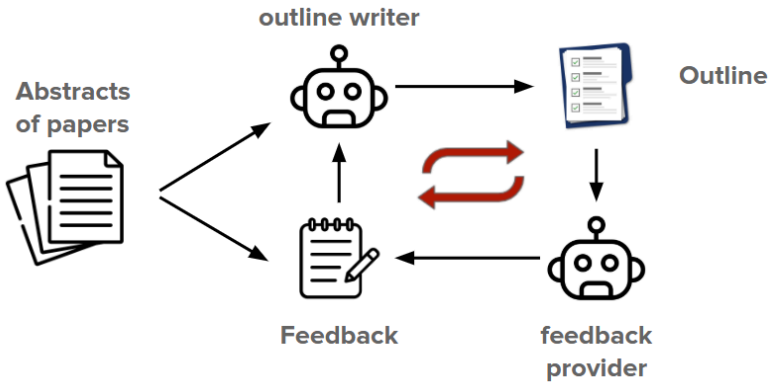

Figure 2. Agents' interaction to generate the survey paper's outline

The outline construction stage, designed to follow an iterative consensus approach, utilizes two GPT-4o agents to construct the outline as well as a brief description for each section of the survey paper. The outline writer agent is tasked with constructing an initial outline using the abstracts from the reference papers. It is also responsible for later refining the outline based on the feedback from the feedback provider agent. The feedback provider agent analyzes the generated outline based on its clarity, coherence, and depth of analysis; then it offers specific, actionable suggestions for refinement.

The iteration process concludes either after a maximum number of discussions are reached or a mutual consensus is achieved. Throughout our experiment, we observed that limiting the process to two rounds of conversation optimizes the performance; additional rounds of conversations tend to introduce irrelevant or made-up details in the outline.

5.2 Stage 2 - Subsection Generation

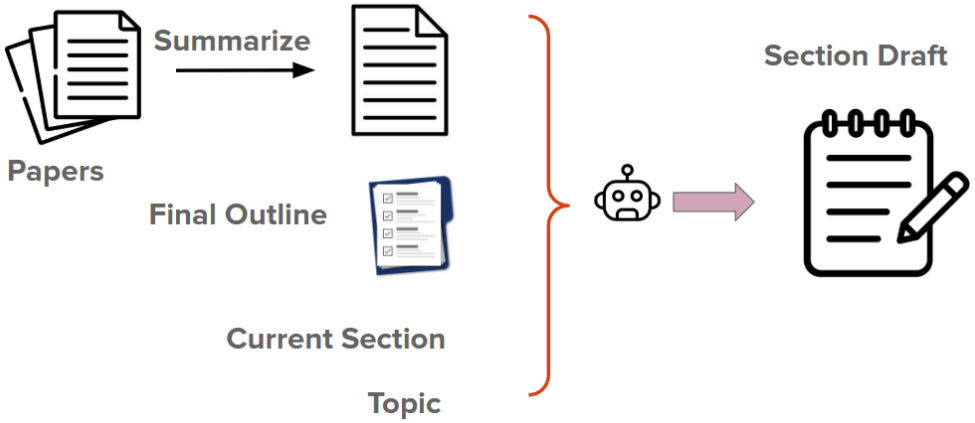

Figure 3. Display of Agent's interaction to Generate Section Draft

The step starts with feeding each reference paper's full content into the summarization agent. Based on the content, it will generate concise summaries of key contributions, methodologies, results, and findings. These summaries, along with the survey's final outline, section names, and topic of the survey, guide the section drafting agent. This agent will use this information to draft each section sequentially until all sections of the survey paper are generated. This iterative method ensures sections are contextually grounded, aligned with the survey's objectives, and integrate insights effectively.

5.3 Stage 3 - Compilation

5.3.1 Hallucination Check

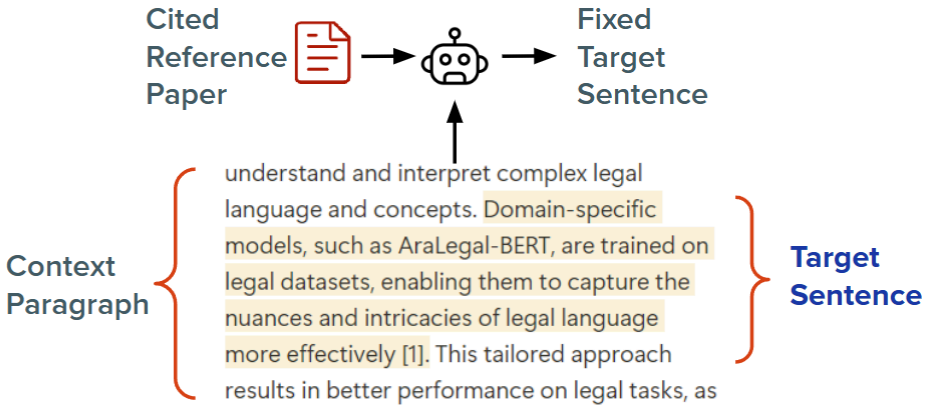

Figure 4. Display of Agent's interaction to check and correct hallucination

In this step, the hallucination check agent verifies the accuracy of citations and corrects any errors. The process involves identifying the target sentence containing the citation and providing a short context paragraph, along with the full content of the cited paper, to the GPT-4o model. The model determines whether the citation is appropriate and suggests any necessary fixes to the target sentence, ensuring the integrity and accuracy of the survey paper.

5.3.2 Refine Draft

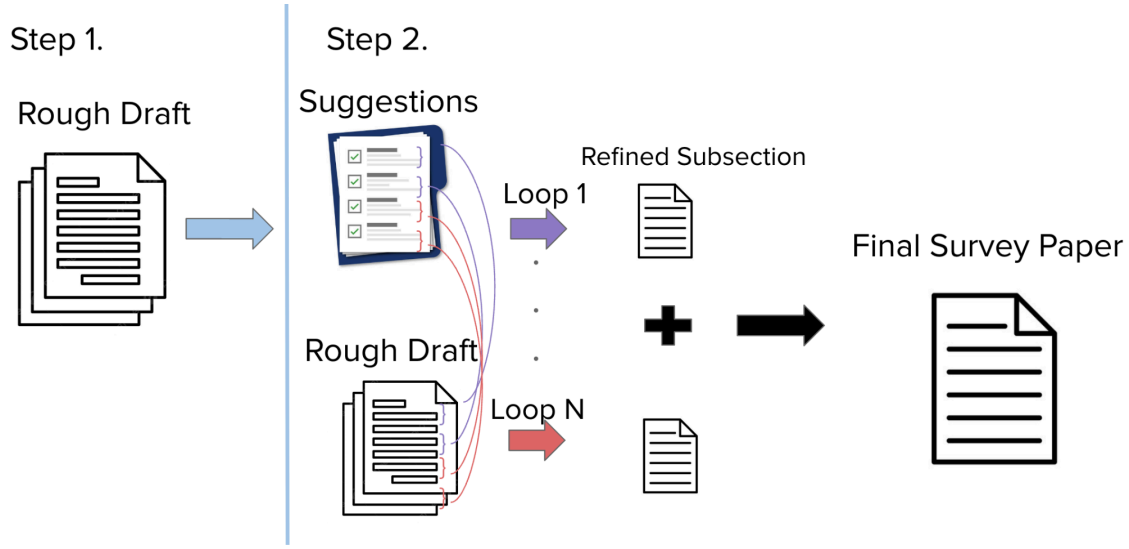

Figure 5. Display of Agent's interaction to refine draft

This system generates the final survey paper and includes two agents with different prompts. In step one, the input would be the rough draft generated by the hallucination checker system. The first agent, represented by the blue arrow in Figure 5, is responsible for reviewing the entire rough draft and outputting tailored suggestions on how to revise each section. In step two, another agent processes the suggestions and the corresponding rough draft, two sections at a time. For each pair of sections, the agent edits the sections' content based on the provided suggestions. Once completed, the agent moves on to the next two sections. This process is repeated until all sections of the draft have been edited. The outputs from each iteration are combined to form the final output of the overall system.

6. Baseline Model

The baseline is a ChatGPT GPT-4o agent with one customized user input prompt. The team chose to use ChatGPT because ChatGPT has a built-in RAG system, which can retrieve content of reference papers. The system prompt provides the background of the task. For example, the agent will know that "it is a highly knowledgeable technical writer and researcher with expertise in analyzing and presenting information in an academically rigorous manner." The user prompt provides specific instructions on how to generate the survey paper. For example, "you need to include citations of the papers that you referenced in the survey paper." Then the titles of the reference papers are also provided for the agent to retrieve relevant information.

7. Evaluation

The team developed five metrics to evaluate the generated survey papers. The metrics include: correctness, relevance to the field, depth of analysis, comprehensiveness, and coherence. The team also devised a rubric containing specific instructions to score each of the metrics from one to five. The complete rubric is displayed in Appendix B.

The team experimented with using GPT-4o to evaluate the output survey paper. However, the results from GPT-4o were unreliable, as it scored both the baseline and SurveyDash outputs very high, despite significant differences between the two generated survey papers. This is likely due to OpenAI's underlying prompt design, which encourages agents to provide overly optimistic evaluations [6]. Consequently, the team opted for individual scoring by team members using the rubric. The final result is the average of all team member's evaluations.

8. Quantitative Results

Metrics	Correctness	Relevance to the field	Depth of Analysis	Comprehensiveness	Coherence	Overall
Baseline (GPT-4o)	3.25	3.25	1.75	2.75	3	2.8
Survey Dash	**4**	**3.75**	**4.125**	**4**	**3.625**	**3.9**

Table 1: Quantitative Results of Baseline(GPT-4o) and SurveyDash

According to Table 1, the baseline model demonstrates moderate correctness and relevance but falls short in terms of depth of analysis and comprehensiveness, indicating limitations in generating a nuanced and detailed survey paper. In contrast, SurveyDash significantly outperforms the baseline. The improvements across all metrics highlight SurveyDash's ability to generate more accurate, relevant, and comprehensive content. Notably, the depth of analysis and comprehensiveness scores indicate a substantial enhancement in the ability to synthesize and integrate information from multiple sources, leading to more insightful and detailed survey papers. These results underscore the advantages of SurveyDash in delivering high-quality survey papers.

9. Qualitative Results

Rosa et al. (*2022*) investigated whether large-scale pretraining outweighs domain-specific data in legal case entailment tasks. Their findings suggest that while pretraining provides a strong baseline, fine-tuning on curated legal datasets remains essential for achieving optimal accuracy.

Figure 6. Output of one section by baseline model

2.2 Case Study: Legal Case Entailment with LLMs

The application of LLMs in legal case entailment tasks, as explored in [13], provides a practical example of how legal reasoning can be embedded into AI systems. In this study, a large-scale LLM was employed to predict the entailment of legal cases, demonstrating the model's ability to process and analyze complex legal texts. The results indicated that scaling the number of parameters in the LLM improved its performance, suggesting that larger models may possess stronger capabilities for legal reasoning.

Despite these promising results, several challenges remain in embedding legal reasoning into LLMs. One key issue is the model's ability to understand and apply legal principles consistently across different cases. Legal reasoning often involves interpreting statutes, precedents, and case-specific details, which require a deep understanding of legal concepts and context. Additionally, the model's performance may be influenced by the quality and diversity of the training data, highlighting the need for comprehensive and representative legal datasets.

Figure 7. Output of the same section as Figure 6 by SurveyDash model

Figures 6 and 7 address the same topic, but SurveyDash provides a much more detailed analysis compared to the baseline model. SurveyDash delivers a comprehensive examination of applying LLMs to legal case entailment tasks, going beyond merely summarizing results. It explores challenges such as embedding legal reasoning into AI, addressing case-specific details, and the impact of training data diversity. Additionally, it emphasizes the importance of scaling LLM parameters and their effect on performance, offering broader insights into AI's potential in legal reasoning. In contrast, the baseline offers only a brief summary of the novel findings. This example aligns with the quantitative results, demonstrating that SurveyDash excels in depth of analysis and comprehensiveness.

1.1 Multilingual Models in Legal Text Analysis

Multilingual models like BLOOM and mT5, trained on diverse linguistic datasets, excel in processing legal documents across languages, crucial for multilingual legal systems like the EU. BLOOM, for instance, is a 176B-parameter open-access language model trained on the ROOTS corpus, which includes data from 46 natural languages and 13 programming languages. Similarly, mT5 extends the T5 architecture to 101 languages, demonstrating state-of-the-art performance on numerous multilingual benchmarks [18].

These models excel in legal tasks by enabling the processing of legal documents in various languages without the need for separate monolingual models. This capability is particularly beneficial in legal systems where multiple languages are used, such as the European Union, where legal texts must be accessible in all member state languages.

Figure 8. Analysis by SurveyDash

Although SurveyDash offers valuable insights for many sections, it occasionally produces superficial analyses. Figure 8 presents an analysis section of the survey paper generated by SurveyDash. In this section, the system identified two models and described their functionalities, as well as their utility. However, it failed to include critical information, such as identifying the current state-of-the-art model for multilingual tasks, which would be highly relevant and informative.

10. Discussion and Learnings

Overall, our system demonstrates superior results compared to the baseline method in generating survey papers, especially in the criteria of depth of analysis and comprehensiveness; it synthesizes complex information rather than merely summarizing or listing concepts.

However, there are some limitations with our current implementation of the system that require addressing. The system's dependency on user-provided titles of reference papers', rather than allowing for topic-only user inputs, presents a usability constraint. Another limitation lies in our evaluation method: the absence of a standardized, specialized expert model for evaluating the system's performance may expose the evaluation process to potential biases or risks of overlooking certain aspects in our internal assessment.

Some key learnings have emerged during the design and implementation of our system. For instance, breaking down a complex task into smaller, less-complex, agent-specific tasks allows for more focused and manageable content generation. Adopting an iterative consensus approach in all stages of the system also contributes to the quality of outputs. As demonstrated in the outline construction stage, the outline generated through the iterative consensus approach is much more detailed, providing more guidance for agents in the subsequent subsection creation stage, compared to the outline generated through a single-agent approach.

11. Individual Contributions

Jim
- Picked target survey paper
- Programmed the Refine-draft system and also created the prompts for the two agents involved in the system
- Created the system and user prompt for the Baseline
- Develop the evaluation rubric
- Evaluated the generated survey paper

*Note: My local git email has a typo, so my commits can only be found if you click into the files I pushed the commit (i.e prompts.py, refined_paper.py, and enriched_sections folder). Please see the appendix for more information

Hongyuan
- Build up the pipeline
- Develop the function to summarize the papers and generate draft section
- Programmed the refine-draft system
- Evaluated the generated survey paper

Yutong
- Built up the pipeline for outline generation and hallucination check
- Created and/or refined prompts for section drafting, hallucination check, summarization
- Developed the evaluation rubric
- Evaluated the generated survey paper

Muhan
- Created the dataset including downloading and parsing the content
- Wrote the agentic prompt for writing the outline structure
- Evaluated the generated survey paper

References

[1] Y. Wang et al., ‘AutoSurvey: Large Language Models Can Automatically Write Surveys’, arXiv [cs.IR]. 2024

[2] S. Li, Y. Du, J. B. Tenenbaum, A. Torralba, and I. Mordatch, ‘Composing Ensembles of Pre-trained Models via Iterative Consensus’, arXiv [cs.CV]. 2022

[3] D. H. Anh, D.-T. Do, V. Tran, and N. L. Minh, ‘The Impact of Large Language Modeling on Natural Language Processing in Legal Texts: A Comprehensive Survey’, in 2023 15th International Conference on Knowledge and Systems Engineering (KSE), 2023, pp. 1–7.

[4] A. Vaswani et al., ‘Attention is all you need’, in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017, pp. 6000–6010.

[5] M. Al-qurishi, S. Alqaseemi, and R. Souissi, ‘AraLegal-BERT: A pretrained language model for Arabic Legal text’, in Proceedings of the Natural Legal Language Processing Workshop 2022, 2022, pp. 338–344.

[6] D. Li et al., ‘From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge’, arXiv [cs.AI]. 2024.

Appendix A

The email on my local git account (Jim) has a typo “gmil” instead of gmail, so Jim’s commit’s might not be visible by clicking on the profile icon in the contributors section. Instead, to see the commits Jim made, please visit the following files: refined_paper.py, prompts.py, and enriched sections folder.

Appendix B - Rubric

Metric	Rating Rubric
Correctness	**5**: Paper is exemplary in accuracy, with no detectable errors. **4**: Paper is largely accurate, with all key facts and methodologies. **3**: Paper is mostly correct with minor errors. **2**: Paper contains noticeable errors that affect the paper’s credibility. **1**: Paper includes numerous factual, conceptual, or methodological errors.
Relevance to the Field	**5**: Paper is entirely focused on the topic, every information mentioned in the paper contributes to a comprehensive understanding of the topic. **4**: Paper is mostly focused on the topic and effectively addresses key areas. **3**: Paper is generally relevant with a fair connection to important aspects of the field. **2**: Paper occasionally touches on relevant topics but lacks focus. **1**: Paper is not relevant to the field; the topics covered are mostly unrelated.
Depth of Analysis	**5**: Paper offers an exceptional depth of analysis, showcasing mastery of the surveyed works. It goes beyond summarization, critically synthesizing ideas to propose new perspectives, frameworks, or questions. **4**: Paper demonstrates a thorough understanding of the surveyed works, critically evaluating their contributions, limitations, and relationships. **3**: Paper provides an adequate analysis, with some critical examination and synthesis of the surveyed works. **2**: Paper’s analysis demonstrates a basic understanding of the surveyed works but lacks depth. **1**: Paper provides only surface-level information and mostly summarizes content without drawing any meaningful insights.
Comprehensive	**5**: Paper comprehensively addresses all critical topics, including emerging trends and niche areas. **4**: Paper provides an extensive review of the field, covering almost all key topics and concepts. **3**: Paper addresses most of the critical topics but may miss a few emerging or less prominent areas. **2**: Paper includes some relevant topics but misses several significant areas. **1**: Paper omits key topics, concepts, or references essential to the field.
Coherence	**5**: Paper exhibits exceptional coherence, with a clear and logical structure

	that guides the reader seamlessly from introduction to conclusion. **4**: Paper is well-structured, with logical organization and clear transitions between sections. **3**: Paper is organized into a coherent structure, though some sections may lack smooth transitions or a clear hierarchy. **2**: Paper shows some structure but frequently jumps between ideas without clear transitions. **1**: Paper is disorganized, with ideas presented in a confusing manner.

Table 2. Rubric used to evaluate the metrics

Appendix C - Outputs from Outline Generation Stage

Initial Outline - By Outline writer agent

Outline for Survey Paper on LLM Applications in Legal Texts

1. Introduction
- **Main Idea:** Introduce the significance of applying Large Language Models (LLMs) in the legal domain, highlighting the potential benefits and challenges.
- **Core Papers:** 1, 3, 5, 7
- **Possible Related Papers:** 9, 13
- **Content Structure:**
  - Overview of the legal domain's complexity and the need for advanced NLP solutions.
  - Brief introduction to LLMs and their general capabilities.
  - Importance of domain-specific models like AraLegal-BERT (1) and their impact on legal tasks.
  - Highlight the growing interest and research in Legal Judgment Prediction (3) and general NLP advancements in the legal field (5, 7).
- **Cohesion Across Sections:** Set the stage for detailed exploration of specific applications and challenges in subsequent sections.

2. Domain-Specific LLMs for Legal Texts
- **Main Idea:** Explore the development and application of domain-specific LLMs tailored for legal texts.
- **Core Papers:** 1, 3, 13
- **Possible Related Papers:** 5, 7
- **Content Structure:**
  - Discuss AraLegal-BERT's customization for Arabic legal texts and its performance improvements over general BERT models (1).
  - Review the survey on Legal Judgment Prediction, focusing on datasets, models, and challenges (3).
  - Examine the case study on legal case entailment using large-scale LLMs (13).
  - Compare and contrast the effectiveness of domain-specific versus general LLMs in legal applications.

  - Identify gaps in current research and potential areas for further exploration.
- **Cohesion Across Sections:** Connect to the broader theme of enhancing LLM capabilities for specific legal tasks and the role of domain-specific models.

3. Multilingual and Crosslingual Capabilities in Legal NLP
- **Main Idea:** Investigate the multilingual and crosslingual capabilities of LLMs in processing legal texts across different languages.
- **Core Papers:** 8, 15, 18
- **Possible Related Papers:** 1, 3
- **Content Structure:**
  - Discuss the role of multilingual models like BLOOM and mT5 in legal text processing (15, 18).
  - Explore crosslingual generalization through multitask finetuning and its implications for legal NLP (8).
  - Highlight the challenges and benefits of applying multilingual LLMs in legal contexts.
  - Address the potential for zero-shot and few-shot learning in multilingual legal tasks.
- **Cohesion Across Sections:** Link to the theme of expanding LLM applications in diverse legal systems and languages.

4. Ethical Considerations and Bias in Legal LLMs
- **Main Idea:** Address the ethical considerations and potential biases in applying LLMs to legal texts.
- **Core Papers:** 6, 9
- **Possible Related Papers:** 1, 3, 5
- **Content Structure:**
  - Present findings on biases in LLMs, such as anti-queer bias, and methods to mitigate them (6).
  - Discuss the "Law Informs Code" approach to aligning AI with human values and legal standards (9).
  - Analyze the implications of bias in legal decision-making and the importance of ethical AI in the legal domain.
  - Propose strategies for ensuring fairness and transparency in legal LLM applications.
- **Cohesion Across Sections:** Emphasize the importance of ethical considerations in the development and deployment of LLMs in legal contexts.

5. Future Directions and Open Challenges
- **Main Idea:** Outline the future directions and unresolved challenges in the application of LLMs to legal texts.
- **Core Papers:** 3, 5, 7
- **Possible Related Papers:** 9, 13
- **Content Structure:**
  - Summarize the current state of research and identify key trends in legal NLP (3, 5, 7).
  - Discuss open challenges such as data availability, model interpretability, and real-time application constraints.

  - Highlight potential research directions, including improved domain-specific models, multilingual capabilities, and ethical AI.
  - Encourage collaboration between NLP researchers and legal professionals to address these challenges.
- **Cohesion Across Sections:** Conclude by tying together the insights from previous sections and emphasizing the ongoing evolution of LLM applications in the legal domain.

6. Conclusion
- **Main Idea:** Recap the key findings and insights from the survey, emphasizing the transformative potential of LLMs in legal texts.
- **Core Papers:** 1, 3, 5, 7
- **Possible Related Papers:** 9, 13
- **Content Structure:**
  - Summarize the advancements and applications of LLMs in the legal domain.
  - Reflect on the ethical considerations and the importance of responsible AI deployment.
  - Reiterate the future directions and the need for continued research and collaboration.
- **Cohesion Across Sections:** Provide a cohesive conclusion that encapsulates the survey's narrative and highlights the importance of LLMs in advancing legal text analysis.

Feedback on Initial Outline - By feedback provider agent

General Feedback
Your outline demonstrates a clear understanding of the topic and presents a logical structure for the survey paper. However, there are areas that require refinement to ensure a more comprehensive and cohesive narrative. It is crucial to ensure that each section not only stands alone but also contributes to the overall flow of the paper. Pay attention to the connections between sections and ensure that each section builds upon the previous one. Additionally, ensure that all relevant papers are considered and that the content structure of each section is detailed and precise.

Specific Feedback

1. Introduction
- **Feedback:** The introduction sets a good foundation by highlighting the significance of LLMs in the legal domain. However, it could benefit from a more detailed explanation of the challenges faced in this field. Additionally, ensure that the introduction clearly outlines the scope and objectives of the survey paper.
- **Suggestions:** Expand on the challenges in applying LLMs to legal texts and provide a clearer roadmap of what the paper will cover.

2. Domain-Specific LLMs for Legal Texts
- **Feedback:** This section effectively discusses domain-specific models like AraLegal-BERT. However, it lacks depth in comparing these models with general LLMs.

- **Suggestions:** Provide a more detailed comparison of domain-specific and general LLMs, highlighting specific advantages and limitations. Consider discussing the implications of these differences on legal tasks.

3. Multilingual and Crosslingual Capabilities in Legal NLP
- **Feedback:** The section covers important aspects of multilingual capabilities but could be more comprehensive in addressing the challenges and solutions.
- **Suggestions:** Elaborate on the specific challenges faced in multilingual legal NLP and discuss potential solutions or ongoing research efforts. Consider including more related papers to provide a broader perspective.

4. Ethical Considerations and Bias in Legal LLMs
- **Feedback:** This section addresses a critical aspect of LLM applications but could benefit from a more structured analysis of the ethical implications.
- **Suggestions:** Structure the discussion around specific ethical issues, such as bias, transparency, and accountability. Provide examples from the core papers to illustrate these points and suggest concrete strategies for mitigation.

5. Future Directions and Open Challenges
- **Feedback:** The section outlines future directions well but lacks specificity in identifying open challenges.
- **Suggestions:** Be more specific about the challenges, such as data scarcity or model interpretability, and propose potential research avenues to address these issues. Highlight the importance of interdisciplinary collaboration in overcoming these challenges.

6. Conclusion
- **Feedback:** The conclusion effectively summarizes the survey but could be more impactful in emphasizing the transformative potential of LLMs.
- **Suggestions:** Reinforce the key findings and insights from the survey, and emphasize the importance of continued research and ethical considerations. Consider ending with a call to action for researchers and practitioners in the field.

Grade

- [ ] Accept
- [ ] Minor Revision
- [X] Major Revision
- [ ] Reject

Your outline requires major revisions to enhance its depth, coherence, and comprehensiveness. Focus on integrating the sections more effectively and providing detailed analyses and comparisons where necessary. With these improvements, your survey paper will be well-positioned to make a significant contribution to the field.

Refined Outline - By Outline writer agent

Draft Outline
Outline for Survey Paper on LLM Applications in Legal Texts

1. Introduction
- **Main Idea:** Introduce the significance of applying Large Language Models (LLMs) in the legal domain, highlighting the potential benefits and challenges.
- **Core Papers:** 1, 3, 5, 7
- **Possible Related Papers:** 9, 13
- **Content Structure:**
  - Overview of the legal domain's complexity and the need for advanced NLP solutions.
  - Brief introduction to LLMs and their general capabilities.
  - Importance of domain-specific models like AraLegal-BERT (1) and their impact on legal tasks.
  - Highlight the growing interest and research in Legal Judgment Prediction (3) and general NLP advancements in the legal field (5, 7).
  - Discuss challenges such as data scarcity, model interpretability, and ethical considerations.
- **Cohesion Across Sections:** Set the stage for detailed exploration of specific applications and challenges in subsequent sections.

2. Domain-Specific LLMs for Legal Texts
- **Title:** Customization and Performance of Domain-Specific LLMs
- **Main Idea:** Explore the development and application of domain-specific LLMs tailored for legal texts.
- **Core Papers:** 1, 3, 13
- **Possible Related Papers:** 5, 7
- **Content Structure:**
  - Discuss AraLegal-BERT's customization for Arabic legal texts and its performance improvements over general BERT models (1).
  - Review the survey on Legal Judgment Prediction, focusing on datasets, models, and challenges (3).
  - Examine the case study on legal case entailment using large-scale LLMs (13).
  - Compare and contrast the effectiveness of domain-specific versus general LLMs in legal applications.
  - Identify gaps in current research and potential areas for further exploration.
- **Cohesion Across Sections:** Connect to the broader theme of enhancing LLM capabilities for specific legal tasks and the role of domain-specific models.

3. Multilingual and Crosslingual Capabilities in Legal NLP
- **Title:** Multilingual and Crosslingual LLMs in Legal Contexts

- **Main Idea:** Investigate the multilingual and crosslingual capabilities of LLMs in processing legal texts across different languages.
- **Core Papers:** 8, 15, 18
- **Possible Related Papers:** 1, 3
- **Content Structure:**
  - Discuss the role of multilingual models like BLOOM and mT5 in legal text processing (15, 18).
  - Explore crosslingual generalization through multitask finetuning and its implications for legal NLP (8).
  - Highlight the challenges and benefits of applying multilingual LLMs in legal contexts.
  - Address the potential for zero-shot and few-shot learning in multilingual legal tasks.
- **Cohesion Across Sections:** Link to the theme of expanding LLM applications in diverse legal systems and languages.

4. Ethical Considerations and Bias in Legal LLMs
- **Title:** Addressing Bias and Ensuring Ethical AI in Legal Applications
- **Main Idea:** Address the ethical considerations and potential biases in applying LLMs to legal texts.
- **Core Papers:** 6, 9
- **Possible Related Papers:** 1, 3, 5
- **Content Structure:**
  - Present findings on biases in LLMs, such as anti-queer bias, and methods to mitigate them (6).
  - Discuss the "Law Informs Code" approach to aligning AI with human values and legal standards (9).
  - Analyze the implications of bias in legal decision-making and the importance of ethical AI in the legal domain.
  - Propose strategies for ensuring fairness and transparency in legal LLM applications.
- **Cohesion Across Sections:** Emphasize the importance of ethical considerations in the development and deployment of LLMs in legal contexts.

5. Future Directions and Open Challenges
- **Title:** Advancing Legal NLP: Future Directions and Challenges
- **Main Idea:** Outline the future directions and unresolved challenges in the application of LLMs to legal texts.
- **Core Papers:** 3, 5, 7
- **Possible Related Papers:** 9, 13
- **Content Structure:**
  - Summarize the current state of research and identify key trends in legal NLP (3, 5, 7).
  - Discuss open challenges such as data availability, model interpretability, and real-time application constraints.
  - Highlight potential research directions, including improved domain-specific models, multilingual capabilities, and ethical AI.

  - Encourage collaboration between NLP researchers and legal professionals to address these challenges.
- **Cohesion Across Sections:** Conclude by tying together the insights from previous sections and emphasizing the ongoing evolution of LLM applications in the legal domain.

6. Conclusion
- **Title:** Recap and Call to Action for LLMs in Legal Texts
- **Main Idea:** Recap the key findings and insights from the survey, emphasizing the transformative potential of LLMs in legal texts.
- **Core Papers:** 1, 3, 5, 7
- **Possible Related Papers:** 9, 13
- **Content Structure:**
  - Summarize the advancements and applications of LLMs in the legal domain.
  - Reflect on the ethical considerations and the importance of responsible AI deployment.
  - Reiterate the future directions and the need for continued research and collaboration.
- **Cohesion Across Sections:** Provide a cohesive conclusion that encapsulates the survey's narrative and highlights the importance of LLMs in advancing legal text analysis.

Feedback on revised outline - By feedback provider agent

Feedback on Outline

General Feedback
Your outline demonstrates a solid understanding of the thematic areas within the application of LLMs in legal texts. However, there are areas where you can enhance the clarity, coherence, and depth of your survey paper. Below are specific suggestions for each section:

1. Introduction
- **Content Structure:** While you have identified the core ideas, ensure that you provide a more detailed explanation of why these aspects are significant. For instance, when discussing the importance of domain-specific models like AraLegal-BERT, briefly mention what sets them apart from general models.
- **Cohesion Across Sections:** Consider explicitly stating how the introduction will lead into the next section on domain-specific LLMs. This will help in creating a seamless transition.

2. Domain-Specific LLMs for Legal Texts
- **Content Structure:** You should delve deeper into the specific contributions of each core paper. For example, what specific performance improvements does AraLegal-BERT offer? How does the survey on Legal Judgment Prediction contribute to the field?
- **Comparisons and Contrasts:** Provide more explicit comparisons between domain-specific and general LLMs. What are the specific advantages or limitations of each?

- **Gaps and Unresolved Issues:** Clearly identify any gaps in the current research that these papers highlight. This will provide a foundation for discussing future directions.

3. Multilingual and Crosslingual Capabilities in Legal NLP
- **Content Structure:** Ensure that you provide a clear explanation of how multilingual models like BLOOM and mT5 are applied in legal contexts. What specific challenges do they address?
- **Comparisons and Contrasts:** Discuss how these models compare to monolingual models in terms of performance and applicability.
- **Cohesion Across Sections:** Make sure to link this section to the previous one by discussing how multilingual capabilities can enhance domain-specific applications.

4. Ethical Considerations and Bias in Legal LLMs
- **Content Structure:** Provide a more detailed analysis of the findings on biases in LLMs. How do these biases manifest in legal applications, and what are the potential consequences?
- **Strategies for Mitigation:** Offer specific strategies for mitigating bias and ensuring ethical AI, drawing from the core papers.
- **Cohesion Across Sections:** Highlight how ethical considerations are a recurring theme throughout the survey and their importance in future research directions.

5. Future Directions and Open Challenges
- **Content Structure:** Clearly outline the key trends identified in the core papers and how they inform future research directions.
- **Open Challenges:** Provide a more detailed discussion of each challenge, supported by examples from the core papers.
- **Collaboration:** Emphasize the importance of interdisciplinary collaboration and provide examples of successful partnerships.

6. Conclusion
- **Content Structure:** Ensure that the conclusion not only summarizes the key findings but also provides a compelling call to action for researchers and practitioners.
- **Cohesion Across Sections:** Reinforce the narrative by tying together the insights from each section and emphasizing the survey's contribution to the field.

Suggestions for Writing Each Section

Introduction
- Begin with a strong opening statement that captures the reader's attention and sets the context for the survey.
- Clearly articulate the significance of LLMs in the legal domain and outline the main themes of the paper.
- Provide a roadmap for the rest of the paper, highlighting the key sections and their contributions.

Domain-Specific LLMs for Legal Texts
- Start with a brief overview of the need for domain-specific models in legal NLP.
- Discuss each core paper in detail, highlighting their contributions and findings.
- Provide a critical analysis of the effectiveness of domain-specific models compared to general models.

Multilingual and Crosslingual Capabilities in Legal NLP
- Introduce the importance of multilingual capabilities in legal NLP.
- Discuss the core papers, focusing on their methodologies and findings.
- Analyze the challenges and benefits of multilingual and crosslingual models in legal contexts.

Ethical Considerations and Bias in Legal LLMs
- Begin with an overview of the ethical challenges in applying LLMs to legal texts.
- Discuss the core papers, highlighting their findings on bias and ethical considerations.
- Propose strategies for mitigating bias and ensuring ethical AI in legal applications.

Future Directions and Open Challenges
- Summarize the current state of research and identify key trends.
- Discuss each open challenge in detail, supported by examples from the core papers.
- Highlight potential research directions and the importance of collaboration.

Conclusion
- Recap the key findings and insights from the survey.
- Reflect on the ethical considerations and the importance of responsible AI deployment.
- Provide a compelling call to action for continued research and collaboration in the field.

By addressing these points, you will be able to create a more comprehensive and well-organized survey paper. Remember to maintain a critical perspective and provide a balanced analysis of the literature.

Final Outline - By Outline writer agent

Introduction

The application of Large Language Models (LLMs) in the legal domain represents a transformative shift in how legal texts are processed and understood. The legal field is characterized by its complexity, with intricate language and domain-specific terminologies that pose significant challenges for traditional Natural Language Processing (NLP) models. As such, the advent of LLMs, which have demonstrated remarkable capabilities in understanding and generating human-like text, offers promising solutions to these challenges [1, 3, 5, 7].

LLMs, such as BERT and its variants, have revolutionized NLP by enabling state-of-the-art performance across a wide range of tasks, including question answering, language inference, and text classification [4]. However, the general-purpose nature of these models often limits their effectiveness in specialized domains like law, where nuanced understanding and domain-specific knowledge are crucial. This has led to the development of domain-specific models, such as AraLegal-BERT, which are tailored to the unique requirements of legal texts in specific languages and jurisdictions [1]. These models have shown significant improvements in accuracy and performance over their general counterparts, underscoring the importance of customization in legal NLP applications.

The growing interest in Legal Judgment Prediction (LJP) further highlights the potential of LLMs in the legal domain. LJP involves predicting judicial outcomes based on fact descriptions, a task that benefits greatly from the advanced language understanding capabilities of LLMs [3]. Recent surveys have documented the increasing sophistication of methods and the expanding scope of research in this area, reflecting a broader trend towards integrating AI into legal processes [5, 7].

Despite these advancements, several challenges remain. Data scarcity is a persistent issue, as legal datasets are often limited in size and diversity, hindering the training of robust models. Additionally, the interpretability of LLMs is a critical concern, particularly in legal contexts where transparency and accountability are paramount. Ethical considerations, such as bias and fairness, also play a significant role in the deployment of AI in legal settings, necessitating careful attention to these issues [9, 13].

This survey aims to explore the current landscape of LLM applications in legal texts, focusing on the development and performance of domain-specific models, the multilingual and crosslingual capabilities of LLMs, and the ethical considerations involved. By examining these aspects, we hope to provide a comprehensive overview of the field and identify future directions for research and collaboration. The subsequent sections will delve into these topics in detail, setting the stage for a deeper understanding of the transformative potential of LLMs in the legal domain.

Domain-Specific LLMs for Legal Texts

The application of Large Language Models (LLMs) in the legal domain has gained significant traction, particularly with the development of domain-specific models tailored to handle the unique challenges posed by legal texts. This section delves into the customization and performance of such models, highlighting their contributions and comparing them to general LLMs.

AraLegal-BERT: Customization for Arabic Legal Texts

AraLegal-BERT is a prime example of a domain-specific LLM developed to address the intricacies of Arabic legal texts. The model is a variant of BERT, fine-tuned specifically for the legal domain, and has demonstrated superior performance over general BERT models in

several natural language understanding tasks [1]. AraLegal-BERT's customization involves training on domain-relevant datasets, which enhances its ability to understand and process legal jargon, complex sentence structures, and context-specific nuances prevalent in legal documents. This targeted approach results in improved accuracy and efficiency in tasks such as legal document classification, information retrieval, and jurisprudence analysis.

Legal Judgment Prediction: Datasets, Models, and Challenges

The survey on Legal Judgment Prediction (LJP) provides a comprehensive overview of the datasets, models, and challenges associated with predicting legal judgments using NLP techniques [3]. The paper highlights the importance of domain-specific models in achieving better performance in LJP tasks. By analyzing 31 datasets across six languages, the survey underscores the need for models that can handle the diverse linguistic and contextual elements inherent in legal texts. The findings suggest that domain-specific LLMs, like those used in LJP, offer significant advantages in terms of accuracy and reliability compared to their general counterparts.

Legal Case Entailment: A Case Study with Large-Scale LLMs

A recent case study on legal case entailment demonstrates the potential of large-scale LLMs in legal applications. The study utilized a 3-billion-parameter model to tackle the legal case entailment task in the COLIEE 2022 competition, achieving remarkable results in zero-shot scenarios [13]. This case study illustrates that while larger models can enhance performance, domain-specific training remains crucial for optimal results in legal contexts. The study also highlights the trade-offs between model size and real-time application constraints, emphasizing the need for efficient domain-specific models.

Comparing Domain-Specific and General LLMs

The effectiveness of domain-specific LLMs, such as AraLegal-BERT, in legal applications is evident when compared to general LLMs. Domain-specific models are tailored to understand the unique language and context of legal texts, leading to improved performance in tasks like legal document classification and judgment prediction. In contrast, general LLMs, while versatile, may lack the nuanced understanding required for legal applications, resulting in lower accuracy and efficiency.

Gaps and Unresolved Issues

Despite the advancements in domain-specific LLMs for legal texts, several gaps and unresolved issues remain. One significant challenge is the scarcity of high-quality, annotated legal datasets, which are essential for training and fine-tuning models. Additionally, there is a need for further research into the interpretability of these models, as legal professionals require transparent and explainable AI systems. Addressing these gaps will be crucial for the

continued development and deployment of effective domain-specific LLMs in the legal domain.

In conclusion, domain-specific LLMs have shown great promise in enhancing the capabilities of NLP applications in the legal field. By focusing on the unique challenges posed by legal texts, these models offer significant improvements over general LLMs, paving the way for more accurate and efficient legal text analysis. However, ongoing research and development are necessary to address the existing gaps and ensure the successful integration of these models into legal practice.

Multilingual and Crosslingual Capabilities in Legal NLP

The legal domain, characterized by its linguistic diversity and jurisdictional variations, necessitates robust multilingual and crosslingual capabilities in Natural Language Processing (NLP) models. The advent of multilingual language models such as BLOOM and mT5 has opened new avenues for processing legal texts across different languages, addressing the inherent challenges of multilingual legal systems.

Multilingual Models in Legal Contexts

Multilingual models like BLOOM, a 176B-parameter open-access language model, and mT5, a massively multilingual pre-trained text-to-text transformer, have demonstrated significant potential in handling legal texts across various languages [15, 18]. These models are designed to process and understand multiple languages simultaneously, which is crucial for legal applications where documents may be drafted in different languages depending on the jurisdiction.

BLOOM, trained on the ROOTS corpus comprising 46 natural languages, showcases competitive performance on a wide array of benchmarks, with enhanced results following multitask prompted finetuning [15]. This capability is particularly beneficial in legal contexts where cross-jurisdictional understanding is essential. Similarly, mT5, pre-trained on a dataset covering 101 languages, achieves state-of-the-art performance on numerous multilingual benchmarks, making it a valuable tool for legal NLP tasks that require crosslingual understanding [18].

Crosslingual Generalization and Multitask Finetuning

Crosslingual generalization through multitask finetuning is a promising approach for enhancing the performance of multilingual models in legal NLP. The study by Sanh et al. [8] demonstrates that finetuning large multilingual language models on English tasks with English prompts can lead to task generalization in non-English languages present in the pretraining corpus. This approach not only improves performance on multilingual tasks but also facilitates zero-shot learning, where models can perform tasks in languages they have never explicitly encountered.

The implications of such capabilities in legal NLP are profound. Legal professionals often deal with documents in multiple languages, and the ability of models to generalize across languages without explicit training on each language can significantly streamline legal processes. Moreover, the use of machine-translated prompts to match the language of each dataset further enhances the model's performance on human-written prompts in respective languages, as evidenced by the improvements observed in crosslingual tasks [8].

Challenges and Benefits

While the benefits of multilingual and crosslingual models in legal NLP are evident, several challenges persist. One major challenge is the potential for "accidental translation," where a generative model might inadvertently translate its predictions into an unintended language. This issue, highlighted in the development of mT5, necessitates careful design and training strategies to prevent such occurrences [18].

Despite these challenges, the application of multilingual models in legal contexts offers significant advantages. They enable legal practitioners to access and analyze legal documents across different languages, facilitating a more comprehensive understanding of international legal frameworks. Additionally, the potential for zero-shot and few-shot learning in multilingual legal tasks reduces the need for extensive labeled datasets in every language, making legal NLP more accessible and scalable.

Conclusion

In conclusion, the integration of multilingual and crosslingual capabilities in legal NLP represents a significant advancement in the field. Models like BLOOM and mT5 exemplify the potential of these technologies to transform legal text processing across diverse linguistic landscapes. As research continues to address the challenges associated with multilingual models, their application in legal contexts is poised to enhance the efficiency and effectiveness of legal professionals worldwide. The exploration of these capabilities underscores the importance of expanding LLM applications to accommodate the complexities of global legal systems.
Ethical Considerations and Bias in Legal LLMs

The application of Large Language Models (LLMs) in the legal domain introduces significant ethical considerations, particularly concerning bias and fairness. As LLMs are increasingly utilized in legal contexts, it is crucial to address these ethical challenges to ensure that AI systems align with human values and legal standards.

Bias in Legal LLMs

Bias in LLMs can manifest in various forms, including gender, racial, and cultural biases, which can have profound implications in legal applications. For instance, the study by [6] highlights the presence of anti-queer bias in LLMs like BERT, demonstrating that these

models can encode significant biases against queer and trans individuals. Such biases can lead to unfair treatment and discrimination in legal decision-making processes, where impartiality is paramount.

The potential consequences of biased LLMs in legal contexts are severe. Biased models may perpetuate existing inequalities and injustices, leading to decisions that disproportionately affect marginalized groups. This is particularly concerning in legal judgment prediction and other applications where LLMs may influence or automate decision-making processes [3].

Mitigating Bias and Ensuring Ethical AI

To address these biases, several strategies have been proposed. One effective approach is finetuning LLMs on data that is representative of diverse perspectives, as suggested by [6]. By incorporating data written by and about marginalized communities, LLMs can be adjusted to reduce biases and improve fairness in their outputs.

Another promising strategy is the "Law Informs Code" approach, which seeks to align AI systems with human values through legal reasoning and standards [9]. This approach leverages legal processes, such as statutory interpretation and contract drafting, to guide AI behavior. By embedding legal knowledge and reasoning into AI systems, it is possible to create models that better reflect societal values and ethical standards.

Importance of Ethical Considerations

Ethical considerations are a recurring theme throughout the development and deployment of LLMs in legal contexts. Ensuring fairness and transparency in AI systems is not only a technical challenge but also a moral imperative. As LLMs continue to evolve and become more integrated into legal processes, it is essential to prioritize ethical AI to maintain public trust and uphold the principles of justice.

In conclusion, addressing bias and ensuring ethical AI in legal applications is critical for the responsible deployment of LLMs. By implementing strategies to mitigate bias and align AI systems with human values, we can harness the potential of LLMs while safeguarding against their risks. As the field progresses, ongoing research and collaboration between NLP researchers and legal professionals will be vital in navigating these ethical challenges.
Future Directions and Open Challenges

The application of Large Language Models (LLMs) in the legal domain is a rapidly evolving field, with significant potential to transform legal practice and research. However, several future directions and open challenges must be addressed to fully realize this potential.

Current State and Key Trends

The current state of research in legal NLP, as outlined in the core papers, highlights several key trends. There is a growing interest in developing domain-specific models, such as AraLegal-BERT, which demonstrate improved performance over general models in legal tasks [1]. Additionally, the increasing availability of large-scale public datasets and advances in NLP research have spurred interest in tasks like Legal Judgment Prediction (LJP) [3]. The field is also witnessing a shift towards more sophisticated methods and a broader range of languages covered, as documented in recent surveys [5, 7].

Open Challenges

1. **Data Availability and Quality**: One of the primary challenges in legal NLP is the scarcity of high-quality, annotated legal datasets. Many legal documents are proprietary or confidential, limiting access for research purposes. Efforts to create open-access datasets, as seen in the development of multilingual models like BLOOM, are crucial [15]. However, ensuring the datasets' quality and representativeness remains a challenge.

2. **Model Interpretability**: The complexity of LLMs often results in models that are difficult to interpret, which is a significant concern in the legal domain where transparency and accountability are paramount. Developing methods to enhance model interpretability without compromising performance is an ongoing research challenge [5].

3. **Real-Time Application Constraints**: The deployment of LLMs in real-time legal applications, such as legal case entailment tasks, is hindered by latency and computational constraints. While larger models like GPT-3 have shown promise in zero-shot scenarios, their practical application is limited by these constraints [13].

4. **Ethical Considerations and Bias**: Bias in LLMs, such as anti-queer bias, poses ethical challenges in legal applications [6]. Ensuring fairness and transparency in legal decision-making requires ongoing efforts to identify and mitigate biases. The "Law Informs Code" approach offers a framework for aligning AI with human values and legal standards, but further research is needed to implement these strategies effectively [9].

Potential Research Directions

1. **Improved Domain-Specific Models**: Continued development of domain-specific models tailored to various legal systems and languages is essential. These models should be designed to handle the nuances of legal language and context, offering more accurate and reliable results than general models.

2. **Multilingual Capabilities**: Expanding the multilingual and crosslingual capabilities of LLMs can enhance their applicability in diverse legal systems. Research should focus on improving zero-shot and few-shot learning capabilities to facilitate the processing of legal texts in multiple languages [8, 15].

3. **Ethical AI and Bias Mitigation**: Developing robust methods for bias detection and mitigation is critical. This includes creating benchmarks for bias evaluation and exploring techniques such as finetuning on diverse datasets to reduce biases in LLMs [6].

4. **Interdisciplinary Collaboration**: Collaboration between NLP researchers and legal professionals is vital to address the challenges and leverage the opportunities presented by LLMs in the legal domain. Successful partnerships can lead to the development of more effective tools and methodologies that meet the needs of legal practitioners.

In conclusion, while significant progress has been made in applying LLMs to legal texts, addressing these open challenges and pursuing these research directions will be crucial for advancing the field. By fostering collaboration and innovation, the legal NLP community can continue to enhance the capabilities and impact of LLMs in legal practice and research.
Conclusion

In this survey, we have explored the transformative potential of Large Language Models (LLMs) in the legal domain, highlighting their applications, challenges, and future directions. The integration of LLMs into legal text analysis has shown significant promise, particularly through the development of domain-specific models like AraLegal-BERT, which demonstrate enhanced performance over general models in tasks such as Legal Judgment Prediction [1, 3]. These advancements underscore the importance of tailoring LLMs to the unique requirements of legal texts, where precision and context are paramount.

Our examination of multilingual and crosslingual capabilities reveals that models like BLOOM and mT5 are pivotal in broadening the accessibility and applicability of legal NLP across diverse linguistic landscapes [15, 18]. These models facilitate zero-shot and few-shot learning, enabling legal professionals to leverage AI tools in multilingual environments without extensive retraining. However, challenges remain in ensuring that these models can consistently deliver accurate and contextually relevant outputs across different legal systems.

Ethical considerations and bias in LLMs present critical challenges that must be addressed to ensure responsible AI deployment in legal contexts. The presence of biases, such as those identified in anti-queer bias studies, highlights the need for ongoing efforts to mitigate such issues through strategies like finetuning on diverse datasets and embedding legal reasoning into AI systems [6, 9]. Ensuring fairness and transparency in AI-driven legal applications is not only a technical challenge but also a societal imperative.

Looking forward, the field of legal NLP stands at a crossroads, with numerous open challenges and opportunities for innovation. Key areas for future research include improving data availability, enhancing model interpretability, and addressing real-time application constraints [3, 5, 7]. Collaboration between NLP researchers and legal professionals will be crucial in overcoming these hurdles and advancing the state of the art.

In conclusion, the application of LLMs in legal texts holds immense potential to revolutionize legal practice and research. By continuing to address ethical considerations and fostering interdisciplinary collaboration, we can harness the power of LLMs to create more efficient, equitable, and transparent legal systems. This survey serves as a call to action for researchers and practitioners to engage in this dynamic field and contribute to its ongoing evolution.

Appendix D

Summarization Prompt:

Role
You are a knowledgeable assistant tasked with summarizing content of an academic paper.

Objective
Provide a thorough and comprehensive summary of an academic paper based on the provided full content.

Skills
1. Proficient in detailed information analysis.

Steps
1. Thoroughly analyze the content of the research paper to extract **all** key points, main ideas, and essential details.
2. Create a detailed summary that encompasses significant information in a clear, concise, and organized manner. Try to analyze and summarize the content in each section and compile the section summaries together. You should compile the section summaries into a single paragraph as your final response.
3. Ensure the summary is **error-free** and accurately reflects the website's content.

Requirement
1. The summary **must** include every key point from the research paper, capturing all important details accurately and concisely.
2. Do **not** include your thought process; output **only** the summarization.
[Content of Survey Paper]

Summarization Output Example #1:
The paper presents AraLegal-BERT, a BERT-based language model specifically pre-trained for Arabic legal texts. The study highlights the necessity of domain-specific adaptations of BERT, given the unique vocabulary and semantics in legal texts. AraLegal-BERT was developed by training from scratch using a large dataset of Arabic legal documents, encompassing legislative, judicial, and contractual texts. The model was fine-tuned and evaluated against existing Arabic BERT models on three Natural Language Understanding (NLU) tasks: legal text classification, keyword extraction, and named entity recognition. AraLegal-BERT outperformed other models, showing significant improvements in F1-Macro scores, particularly in keyword extraction and named entity recognition tasks. The model's

efficiency is attributed to its focus on domain-specific vocabulary and semantics, making it a cost-effective solution for various Arabic legal applications. The study suggests that pre-training BERT for specific domains enhances performance on related tasks, with future work aimed at exploring other models like Electra and Roberta for similar tasks in the Arabic legal domain. The research was supported by Elm Company's Research Department.

Comparison with the paper's abstract:
The effectiveness of the BERT model on multiple linguistic tasks has been well documented. On the other hand, its potentials for narrow and specific domains such as Legal, have not been fully explored. In this paper, we examine how BERT can be used in the Arabic legal domain and try customizing this language model for several downstream tasks using several different domain-relevant training and testing datasets to train BERT from scratch. We introduce the AraLegal-BERT, a bidirectional encoder Transformer-based model that have been thoroughly tested and carefully optimized with the goal to amplify the impact of NLP-driven solution concerning jurisprudence, legal documents, and legal practice. We fine-tuned AraLegal-BERT and evaluated it against three BERT variations for Arabic language in three natural languages understanding (NLU) tasks. The results show that the base version of\nAraLegal-BERT achieve better accuracy than the general and original BERT over the Legal text.

Section Drafting Prompt:
Role
You are a researcher with expertise in writing academic survey papers

Objective
You are tasked to enrich the current draft content for a subsection.

Context
1. You will be provided with a research paper, as well as its index number, that is relevant to the content of this subsection at a time as reference. The research paper provided has already been cited in the draft using the format [index of the paper].
2. You will be provided with the current draft of the section of the survey paper. The current draft of the section is written using only the abstract of the research paper.

Steps
You **must** think step-by-step when writing the subsection.
1. **Analyze**
1-1. Analyze the content of the provided research paper. Specifically, you need to analyze what sections of the provided research paper is most relevant to the current subsection of the survey paper. Make sure you fully understand the content of the research paper before writing.
1-2. Analyze the current draft of the subsection. Specifically, you need to check how you can enrich the content in the current draft using the content from the provided research paper.

2. **Write**: write the draft of the subsection using **only** the content of the provided research paper. **Do not** modify parts that are not relevant to the content of the provided research paper.
3. **Evaluate**: evaluate the draft you just created.

Requirement
1. When writing the content, you **must** ensure you are only using the content of the provided research paper.
2. You **must** cite the paper when
2-1. Summarizing Research: Cite sources when summarizing the existing literature.
2-2. Using Specific Concepts or Data: Provide citations when discussing specific theories, models, or data.
2-3. Comparing Findings: Cite relevant studies when comparing or contrasting different findings.
2-4. Highlighting Research Gaps: Cite previous research when pointing out gaps your survey addresses.
2-5. Using Established Methods: Cite the creators of methodologies you employ in your survey.
2-6. Supporting Arguments: Cite sources that back up your conclusions and arguments.
2-7. Suggesting Future Research: Reference studies related to proposed future research directions.
3. Whenever information from the provided contexts is used or referenced, you **must** cite it appropriately using the format [x], where x is the index number of the paper. If a sentence comes from multiple contexts, please list **all** applicable citations, like [x_1; x_2]. You **do not** need to provide a separate section for reference in your response.
3-1. If you are modifying an existing sentence in the draft that already contains a citation, **do not** delete this citation when modifying.
3-2. Citation Example: the emergence of large language models (LLMs) [1]
4. The draft for the subsection **must** be less than 1000 words.
5. Do not report your thought process, return the draft only.

The topic of the survey paper is: {topic}
The current section you are working on is: {section}.

You must only write the {section} section you planned to write. Do not write the entire survey paper before making sure the current section is well-written.

Outline
{outline}

Papers
"""
Section Drafting Output Example:

The application of Large Language Models (LLMs) in the legal domain represents a significant advancement in the intersection of artificial intelligence and law. As the legal field is characterized by its complexity, precision, and reliance on extensive textual data, the integration of LLMs offers unique opportunities to enhance legal text analysis, streamline legal processes, and improve access to justice. The growing interest in applying Natural Language Processing (NLP) techniques to legal texts is driven by the potential of LLMs to transform traditional legal practices through automation and improved accuracy in legal document analysis [1, 3, 5, 7].

LLMs, such as BERT and its domain-specific variants, have demonstrated remarkable success in various NLP tasks, including text classification, sentiment analysis, and language translation [4]. However, the legal domain presents distinct challenges that necessitate specialized adaptations of these models. Legal texts are often characterized by complex language, domain-specific terminology, and jurisdictional variations, which require models to be finely tuned to capture the nuances of legal language [1, 3]. The development of domain-specific models, such as AraLegal-BERT, highlights the importance of customizing LLMs to address the specific needs of legal text analysis [1].

The potential of LLMs to revolutionize legal text analysis is immense, yet it is accompanied by challenges that must be addressed to fully realize their benefits. These challenges include the need for domain-specific training data, the handling of multilingual legal systems, and the mitigation of biases inherent in language models [3, 5, 7]. Moreover, the ethical implications of deploying LLMs in legal contexts, particularly concerning fairness and transparency, are critical considerations that must be addressed to ensure just outcomes [6, 9].

This survey aims to provide a comprehensive overview of the current state of LLM applications in legal texts, exploring the advancements, challenges, and future directions in this rapidly evolving field. The paper is structured as follows: Section 2 delves into the development and application of domain-specific language models tailored for legal texts. Section 3 examines the role of multilingual and cross-lingual capabilities in legal text analysis. Section 4 addresses the ethical considerations and biases present in LLMs when applied to legal texts. Finally, Section 5 outlines the future research directions and challenges in this domain, while Section 6 concludes with a reflection on the broader implications for the legal field.

By systematically exploring these themes, this survey seeks to illuminate the transformative potential of LLMs in the legal domain and provide insights into the ongoing efforts to enhance their applicability and effectiveness. Through this exploration, we aim to facilitate further research and collaboration between NLP researchers and legal professionals, ultimately contributing to the advancement of legal technology and the improvement of legal services.

Appendix E

Hallucination Prompt:

Role
You are a researcher with expertise in analyzing academic research papers.

Objective
You are tasked to check whether the citations of reference papers in the sentence is correct.

Context
1. You will be provided with **one research paper**, as well as its **index number**, that is referenced in the sentence.
2. You will be provided with the target sentence, as well as its context.
3. Whenever information from the provided contexts is used or referenced, it is cited using the format [x], where x is the index number of the paper. If a sentence comes from multiple contexts, it is cited with **all** applicable citations, like [x_1; x_2].
A reference paper is cited when
- Summarizing Research: Cite sources when summarizing the existing literature.
- Using Specific Concepts or Data: Provide citations when discussing specific theories, models, or data.
- Comparing Findings: Cite relevant studies when comparing or contrasting different findings.
- Highlighting Research Gaps: Cite previous research when pointing out gaps your survey addresses.
- Using Established Methods: Cite the creators of methodologies you employ in your survey.
- Supporting Arguments: Cite sources that back up your conclusions and arguments.
- Suggesting Future Research: Reference studies related to proposed future research directions.

Requirement
1. You need to check if the citation of reference papers in the sentence is correct. Think step-by-step when checking.
1-1. When checking, you **must** ensure you are only looking at the index number of the provided reference paper.
2. Output your response as fixed sentence **only**.
2-1. If you determine the citation of reference paper is correct, **do not** modify the sentence, return as it is. Do not remove the citation from the sentence when returning.
2-2. If you determine the citation of reference paper is incorrect, think about why it is incorrect. Examine the context of the sentence and try to modify the sentence such that either the citation is removed or the content is modified to make the citation correct. You **must** properly updated the citation in the sentence.

Example Output
Target Sentence: By synthesizing insights from core papers [1, 3, 5, 7]

Reference Paper Index Number: 3
By synthesizing insights from core papers [1, 5, 7]

Target Sentence: the emergence of large language models (LLMs) [1]
Reference Paper Index Number: 1
the emergence of large language models (LLMs) [1]

Target Sentence: It is said that Lay's Smoky Bacon flavor chip is not gluten-free [11]
Reference Paper Index Number: 11
It is said that Lay's Smoky Bacon flavor chip is gluten-free [11]
"""

Refine System

Advice Prompt

The following are the rough draft of a survey paper. You are acting as an editor before the author can publish their paper. Provide suggestions on improving the quality of these sections. Find places that are repetitive, where not enough detail is provided, or ideas that can be combined to make the sections more concise. You are not limited to these types of suggestions. For each suggestion, clearly state what needs to be modified and how it should be modified. Also, list out your reasons for each modification. Write the title of the section that you are providing feedback on before you provide the suggestion.

Output Format: For each section, your feedback should follow this format:
Section Title: [Title of the section]
Suggestion 1:
What to modify: [Clearly state the part of the text that needs modification].
How to modify: [Provide the exact changes or general advice].
Reason: [Explain why the change is necessary].
Suggestion 2:
What to modify: [Clearly state the part of the text that needs modification].
How to modify: [Provide the exact changes or general advice].
Reason: [Explain why the change is necessary].
Other suggestions will follow the same suggestion as suggestion 1 and suggestion 2
Remember to also provide suggestions for the Conclusion paragraph
Full rough draft of survey paper:

Advice Output

Generated Combined Advice:
Section Title: Introduction
Suggestion 1:
What to modify: The repetitive explanation of AraLegal-BERT in the introduction.
How to modify: Combine the two instances where AraLegal-BERT is explained into one concise explanation.

Reason: The repetition of the same information can make the introduction seem redundant and less engaging.
Suggestion 2:
What to modify: The last paragraph of the introduction.
How to modify: Instead of listing out the structure of the paper, provide a brief overview of the main points that will be discussed in the paper.
Reason: This will make the introduction more concise and will give the reader a clear understanding of what to expect from the paper.
Section Title: 2. Domain-Specific LLMs for Legal Texts
Suggestion 1:
What to modify: The repetitive explanation of AraLegal-BERT in this section.
How to modify: Since AraLegal-BERT was already explained in the introduction, you can refer back to that explanation instead of repeating it.
Reason: This will make the section more concise and avoid redundancy.
Suggestion 2:
What to modify: The explanation of Legal Judgment Prediction Models and Datasets.
How to modify: Provide more specific examples of these models and datasets to give the reader a better understanding of what they are.
Reason: This will provide more detail and context to the reader.
Section Title: Multilingual and Cross-Lingual Capabilities
Suggestion 1: What to modify: The explanation of BLOOM and mT5.
How to modify: Provide more specific examples of how these models are used in legal tasks.
Reason: This will provide more detail and context to the reader.
Suggestion 2: What to modify: The explanation of cross-lingual generalization.
How to modify: Provide more specific examples of how cross-lingual generalization is used in legal tasks.
Reason: This will provide more detail and context to the reader.
Section Title: Legal Reasoning and Societal Values in LLMs
Suggestion 1: What to modify: The explanation of the "Law Informs Code" approach.
How to modify: Provide more specific examples of how this approach is used in legal tasks.
Reason: This will provide more detail and context to the reader.
Suggestion 2:
What to modify: The explanation of the case study.
How to modify: Provide more specific details about the case study to give the reader a better understanding of it.
Reason: This will provide more detail and context to the reader.
Section Title: Future Directions and Ethical Considerations
Suggestion 1: What to modify: The list of future research directions and ethical considerations.
How to modify: Instead of listing out these points, integrate them into the text in a more narrative form.
Reason: This will make the section more engaging and easier to read.
Suggestion 2:
What to modify: The conclusion of this section.

How to modify: Instead of summarizing the points made in the section, provide a brief overview of the implications of these points for the future of LLMs in the legal domain.
Reason: This will provide a more compelling conclusion to the section.
Section Title: Conclusion
Suggestion 1:
What to modify: The summary of the main points of the paper.
How to modify: Instead of summarizing each point, provide a brief overview of the main findings and implications of the paper.
Reason: This will make the conclusion more concise and impactful.
Suggestion 2:
What to modify: The final sentence of the conclusion.
How to modify: Instead of stating that collaboration will be crucial, provide specific examples of how collaboration can help realize the full potential of LLMs in the legal domain.
Reason: This will provide a more compelling and specific ending to the paper.

Refine Prompt
You are the author of the survey paper, and you are doing your final edits before submitting the paper to a conference. There are suggestions provided by editors. You need to follow the suggestions to modify your paper or your paper will not be accepted by the conference. You also need to write the final paper in a formal format of a survey paper. You would submit this draft to the conference. It is also important to note that the writings provided to you are only a part of the survey paper. You should edit the writing according to the section title only. You also need to provide numbered sections. Ignore the context in your memory from previous chats. Make sure you only use the information from this prompt. It is important to note that the following writing is only part of the survey paper. Do not treat it as an entire survey paper.
Section 1: {section1}
Section 2: {section2}
Suggestions

Refined Output
The output would be the final survey paper, which is displayed in Appendix E

Appendix F - SurveyDash Final Output

Introduction

The advent of Large Language Models (LLMs) has marked a transformative era in Natural Language Processing (NLP), showcasing unprecedented capabilities in understanding and generating human language across a wide array of linguistic tasks. These models, exemplified by architectures such as BERT [4] and its derivatives, have demonstrated remarkable proficiency in various applications. However, the application of LLMs in domain-specific

contexts, particularly in the legal domain, presents unique challenges and opportunities that are yet to be fully explored.

Legal texts, with their intricate syntax, specialized terminology like 'habeas corpus' or 'amicus curiae,' and the necessity for precise interpretation, often pose significant hurdles for general-purpose LLMs. As such, there is a growing interest in developing domain-specific adaptations of these models to enhance their applicability and effectiveness in legal contexts. AraLegal-BERT, a domain-specific adaptation of BERT, is pre-trained on a substantial dataset of Arabic legal documents, excelling in tasks like legal text classification and named entity recognition, demonstrating the necessity of tailored LLMs for legal contexts. This showcases the potential of customized LLMs in improving the accuracy and relevance of legal text analysis.

The importance of LLMs in the legal domain extends beyond mere text analysis. They hold the potential to revolutionize legal practice by automating routine tasks, improving access to legal information, and supporting decision-making processes. This survey aims to provide a comprehensive overview of the current state of LLM applications in legal texts, highlighting key advancements, challenges, and future directions.

In recent years, there has been a notable increase in research focused on Legal Judgment Prediction (LJP), which utilizes NLP techniques to predict judicial outcomes based on factual descriptions [3]. This surge in interest is fueled by the availability of large-scale datasets and the continuous evolution of NLP methodologies. Despite the progress made, a significant gap remains between machine and human performance, underscoring the need for further exploration and refinement of LLMs in this domain.

Moreover, the integration of legal reasoning and societal values into LLMs is a critical area of research that seeks to align AI behavior with human goals. The "Law Informs Code" approach [9] exemplifies efforts to embed legal knowledge and reasoning within AI systems, thereby enhancing their alignment with societal values and ethical standards.

This survey will delve into the various facets of LLM applications in legal texts, structured as follows: an exploration of domain-specific LLMs tailored for legal texts, an examination of multilingual and cross-lingual capabilities, an investigation into the incorporation of legal reasoning and societal values, and a discussion on future directions and ethical considerations. By synthesizing insights from core papers [1, 3, 5, 7] and related works [9,

10, 13], this paper aims to provide a valuable resource for both NLP researchers and legal professionals, fostering further collaboration and innovation in this burgeoning field.

Domain-Specific LLMs for Legal Texts

The application of Large Language Models (LLMs) in the legal domain is a burgeoning area of research, driven by the need to process and analyze complex legal texts efficiently. Domain-specific adaptations of LLMs, such as AraLegal-BERT, have emerged to address the unique challenges posed by legal texts, which often contain specialized terminology, intricate structures, and jurisdiction-specific nuances. This section explores the development and application of LLMs tailored specifically for legal texts, highlighting their advantages over general-purpose models and identifying areas for further research.

AraLegal-BERT: Customization for Arabic Legal Texts

AraLegal-BERT is a notable example of a domain-specific LLM designed to enhance the processing of Arabic legal texts. The model is a bidirectional encoder Transformer-based model that has been fine-tuned on legal datasets to improve its performance on tasks such as jurisprudence analysis, legal document classification, and legal practice applications [1]. The customization of AraLegal-BERT involves training the model from scratch using a manually collected dataset of Arabic legal documents, including legislative texts, judicial documents, contracts, and Islamic rules, resulting in a 4.5GB dataset with approximately 13.7 million sentences [1]. This extensive dataset allows AraLegal-BERT to outperform general-purpose BERT models in accuracy and relevance when applied to legal texts, particularly in tasks requiring domain-specific vocabulary and semantics.

Legal Judgment Prediction Models and Datasets

Legal Judgment Prediction (LJP) is another critical area where domain-specific LLMs have shown promise. LJP involves using NLP techniques to predict judicial outcomes based on fact descriptions, a task that requires a deep understanding of legal language and reasoning [3]. The availability of large-scale public datasets and advancements in NLP research have spurred interest in developing models that can accurately predict legal judgments. These models are evaluated using various metrics and benchmark datasets, highlighting the

importance of domain-specific training to achieve state-of-the-art results. Expanding on the specific datasets and metrics used in LJP research provides new insights into the potential of LLMs to transform legal decision-making processes by providing data-driven insights and predictions.

Advancements in Legal Text Analysis

Recent advancements in AI and NLP have significantly impacted legal text analysis, with LLMs playing a pivotal role in this transformation. The integration of sophisticated NLP techniques into legal applications has led to improved accuracy and efficiency in tasks such as document classification, information retrieval, and legal reasoning [5]. These advancements are not only enhancing the capabilities of legal professionals but also democratizing access to legal information by making it more accessible and understandable to non-experts. The comparison between domain-specific models and general LLMs reveals that while general models offer broad applicability, domain-specific models provide superior performance in specialized tasks due to their tailored training and optimization.

Comparison with General LLMs

While general LLMs like BERT and GPT-3 have demonstrated impressive capabilities across various NLP tasks, their performance in the legal domain often falls short compared to domain-specific models. This is primarily due to the specialized nature of legal texts, which require models to understand and interpret complex legal language and concepts. Domain-specific models, such as AraLegal-BERT, are trained on legal datasets, enabling them to capture the nuances and intricacies of legal language more effectively [1]. This tailored approach results in better performance on legal tasks, as evidenced by higher accuracy and relevance in legal text analysis.

Research Gaps and Future Directions

Despite the progress made in developing domain-specific LLMs for legal texts, several research gaps remain. One significant challenge is the need for more comprehensive and diverse legal datasets that cover various jurisdictions and legal systems. Additionally, there is a need for models that can handle the dynamic and evolving nature of legal language, which often changes with new legislation and judicial interpretations. Future research should focus

on creating more robust and adaptable models that can generalize across different legal contexts while maintaining high accuracy and relevance. Furthermore, exploring the integration of legal reasoning and societal values into LLMs could enhance their applicability and alignment with human goals and ethical standards.

In conclusion, domain-specific LLMs have shown significant potential in transforming legal text analysis by providing more accurate and efficient solutions tailored to the unique challenges of the legal domain. By addressing the existing research gaps and continuing to refine these models, the legal field can harness the full potential of LLMs to improve legal practice and access to justice.

Multilingual and Cross-Lingual Capabilities

The application of multilingual large language models (LLMs) in the legal domain offers significant potential for enhancing legal text analysis across various languages and jurisdictions. This section delves into the capabilities of multilingual models like BLOOM and mT5, their performance on legal tasks, and the implications of cross-lingual generalization.

1.1 Multilingual Models in Legal Text Analysis

Multilingual models like BLOOM and mT5, trained on diverse linguistic datasets, excel in processing legal documents across languages, crucial for multilingual legal systems like the EU. BLOOM, for instance, is a 176B-parameter open-access language model trained on the ROOTS corpus, which includes data from 46 natural languages and 13 programming languages. Similarly, mT5 extends the T5 architecture to 101 languages, demonstrating state-of-the-art performance on numerous multilingual benchmarks [18].

These models excel in legal tasks by enabling the processing of legal documents in various languages without the need for separate monolingual models. This capability is particularly beneficial in legal systems where multiple languages are used, such as the European Union, where legal texts must be accessible in all member state languages.

1.2 Cross-Lingual Generalization

Cross-lingual generalization refers to the ability of a model trained on tasks in one language to perform well on similar tasks in another language. This is a crucial feature for legal applications, as it allows for the transfer of legal reasoning and insights across different jurisdictions. For example, BLOOM and mT5 have shown promising results in legal document classification and contract analysis, achieving strong zero-shot performance on tasks in languages that were not explicitly included in their training datasets [8; 18].

The implications of cross-lingual generalization are profound for legal practitioners. It enhances the accessibility of legal information, allowing for more seamless integration of legal knowledge across borders. This capability can aid in the harmonization of legal standards and practices, facilitating international legal cooperation and understanding.

1.3 Challenges and Solutions

Despite their advantages, multilingual models face challenges in the legal domain. One significant issue is the handling of cultural nuances and legal terminology differences that may not be adequately captured by a model trained on general language data. Legal texts often contain specialized vocabulary and context-specific meanings that require careful adaptation of the model.

To address these challenges, researchers have explored techniques such as machine-translated prompts and domain-specific finetuning. For instance, finetuning multilingual models on legal datasets with prompts translated into the target language has been shown to improve performance on human-written prompts in those languages [8]. Additionally, integrating domain-specific knowledge into the training process can enhance the model's understanding of legal terminology and context.

1.4 Conclusion

The integration of multilingual and cross-lingual capabilities in LLMs represents a significant advancement in legal text analysis. Models like BLOOM and mT5 offer the potential to break down language barriers in the legal domain, promoting greater accessibility and understanding of legal information across different jurisdictions. However, ongoing research is needed to address the challenges of cultural nuances and specialized legal terminology, ensuring that these models can fully realize their potential in legal applications. As we

transition to the next section, we will explore how LLMs can incorporate legal reasoning and societal values to align AI behavior with human goals.

Legal Reasoning and Societal Values in LLMs

The integration of legal reasoning and societal values into Large Language Models (LLMs) is a critical area of research that seeks to align AI behavior with human goals and societal norms. This section explores the methodologies and challenges associated with embedding legal reasoning and societal values into LLMs, drawing insights from key papers in the field.

2.1 Law Informs Code: Aligning AI with Human Values

The "Law Informs Code" approach, as discussed in [9], presents a novel framework for embedding legal knowledge and reasoning into AI systems. This methodology leverages the law as a computational engine that translates opaque human values into explicit directives. For instance, in a real-world scenario, this approach could be applied to develop AI systems that assist in drafting legal contracts by ensuring compliance with current legal standards and adapting to unforeseen contingencies. The approach draws parallels between the unpredictability of future contingencies in legal contracts and the unforeseen circumstances AI systems may encounter. Legal standards, which allow for shared understandings and adaptability, serve as a model for developing AI systems that can navigate complex, real-world scenarios.

However, the practical implementation of the "Law Informs Code" approach faces several limitations. One significant challenge is the inherent complexity and variability of legal systems across different jurisdictions. Legal reasoning often involves nuanced interpretations and context-specific applications, which can be difficult to codify into AI systems. Moreover, the law is not a perfect reflection of societal values, as it is influenced by historical and political factors. Therefore, while the law can provide a structured framework for aligning AI with human values, it requires careful parsing and adaptation to ensure its relevance and legitimacy.

2.2 Case Study: Legal Case Entailment with LLMs

The application of LLMs in legal case entailment tasks, as explored in [13], provides a practical example of how legal reasoning can be embedded into AI systems. In this study, a large-scale LLM was employed to predict the entailment of legal cases, demonstrating the model's ability to process and analyze complex legal texts. The results indicated that scaling the number of parameters in the LLM improved its performance, suggesting that larger models may possess stronger capabilities for legal reasoning.

Despite these promising results, several challenges remain in embedding legal reasoning into LLMs. One key issue is the model's ability to understand and apply legal principles consistently across different cases. Legal reasoning often involves interpreting statutes, precedents, and case-specific details, which require a deep understanding of legal concepts and context. Additionally, the model's performance may be influenced by the quality and diversity of the training data, highlighting the need for comprehensive and representative legal datasets.

2.3 Potential Frameworks for Integrating Legal Standards

To address the challenges of embedding legal reasoning and societal values into LLMs, several potential frameworks can be considered. One approach is to develop hybrid models that combine LLMs with rule-based systems, allowing for the integration of explicit legal rules and principles. For example, a hybrid model could use an LLM to interpret legal texts and a rule-based system to apply specific legal standards, enhancing the model's ability to apply legal reasoning in a consistent and interpretable manner.

Another approach is to incorporate feedback mechanisms that allow for continuous learning and adaptation. By leveraging real-time feedback from legal experts and practitioners, LLMs can refine their understanding of legal concepts and improve their alignment with societal values. This iterative process can help address the dynamic nature of legal systems and ensure that AI systems remain relevant and effective in real-world applications.

2.4 Conclusion

In conclusion, the integration of legal reasoning and societal values into LLMs is a complex but essential endeavor. By leveraging frameworks like "Law Informs Code" and exploring innovative methodologies, researchers can enhance the alignment of AI systems with human

goals and societal norms. This not only improves the local usefulness of AI in legal contexts but also contributes to broader efforts toward society-AI alignment.

Future Directions and Ethical Considerations

The application of Large Language Models (LLMs) in the legal domain presents a myriad of opportunities and challenges that warrant further exploration. As the field continues to evolve, several future research directions and ethical considerations emerge as pivotal to the responsible and effective deployment of LLMs in legal contexts.

Future Research Directions

1. **Advancing Research Methodologies for Domain-Specific Adaptations**: While models like AraLegal-BERT have demonstrated the potential of domain-specific adaptations, future research should focus on specific methodologies and technologies that enhance these adaptations. This could involve leveraging advanced machine learning techniques, such as transfer learning and fine-tuning, to better handle the intricacies of legal language, including complex terminologies and context-specific interpretations. Additionally, creating more comprehensive datasets that capture the diversity of legal texts across different jurisdictions and legal systems remains crucial.

2. **Improving Multilingual and Cross-Lingual Capabilities**: The ability of multilingual models like BLOOM and mT5 to generalize across languages is promising. However, more work is needed to address the challenges posed by cultural nuances and legal terminology differences. Research should aim to enhance the accuracy and reliability of these models in translating and interpreting legal texts across various languages, thereby improving accessibility and understanding in global legal contexts.

3. **Integrating Legal Reasoning and Societal Values**: The "Law Informs Code" approach offers a framework for aligning AI behavior with human values. Future research should explore methodologies for embedding legal reasoning and societal values into LLMs more effectively. This includes developing algorithms that can interpret and apply legal standards in a manner consistent with human judgment and societal norms.

4. **Addressing Bias and Fairness**: There is a critical need to address issues of bias and fairness in legal applications. Future research should focus on developing techniques to

identify and mitigate biases in LLMs, ensuring that these models do not perpetuate or exacerbate existing inequalities in the legal system.

Ethical Considerations

Key ethical concerns include data privacy, accountability, potential misuse, and the need for transparency and fairness in LLM applications. The deployment of LLMs in legal contexts raises significant concerns regarding data privacy and security, necessitating compliance with data protection regulations to prevent unauthorized access and misuse. Additionally, the use of LLMs in legal decision-making processes requires clear accountability and transparency mechanisms to ensure interpretability and human oversight. The potential for misuse, such as generating misleading legal documents or automating biased legal decisions, underscores the need for ethical guidelines and regulatory frameworks. Community-driven efforts and open-access models play a crucial role in democratizing access to LLM technology and fostering collaborative advancements in the field.

In conclusion, the future of LLMs in the legal domain holds significant promise, but it also requires careful consideration of ethical implications and ongoing research to address existing challenges. By focusing on these future directions and ethical considerations, researchers and practitioners can harness the transformative potential of LLMs to enhance legal practice and contribute to a more just and equitable legal system.

Conclusion

In this survey, we have explored the transformative potential of Large Language Models (LLMs) in the legal domain, highlighting their significant impact on legal text analysis and practice. The survey has delved into various aspects of LLM applications, including domain-specific adaptations, multilingual capabilities, legal reasoning, and ethical considerations. Through the examination of core papers and related works, several key insights and future directions have emerged.

The development of domain-specific LLMs, such as AraLegal-BERT, underscores the necessity of tailoring models to handle the unique challenges presented by legal texts. These models have shown improved performance over general LLMs, highlighting the importance of customization in achieving accurate and reliable results in legal applications. Additionally,

the growing importance of multilingual and cross-lingual capabilities in legal text analysis is evident, with models like BLOOM and mT5 enhancing the accessibility of legal information and facilitating cross-jurisdictional understanding.

The integration of legal reasoning and societal values into LLMs remains a critical area for future research. The "Law Informs Code" approach offers a promising framework for aligning AI behavior with human goals by embedding legal knowledge and reasoning into AI systems. However, practical implementation challenges persist, particularly in embedding complex legal reasoning and ensuring alignment with societal values.

Ethical considerations are paramount in the deployment of LLMs in legal contexts. Issues such as bias, fairness, data privacy, and accountability must be addressed to ensure the responsible use of these technologies. The survey emphasizes the need for community-driven efforts and open-access models to advance the field while maintaining ethical standards.

Looking ahead, collaboration between NLP researchers and legal professionals is essential to realize the full potential of LLMs in the legal domain. For instance, interdisciplinary workshops and joint research initiatives could facilitate the exchange of knowledge and expertise, leading to the development of more robust and ethically sound legal AI systems. By focusing on enhancing domain-specific adaptations, improving multilingual capabilities, integrating legal reasoning, and addressing ethical concerns, LLMs have the potential to revolutionize legal practice, making it more efficient, accessible, and aligned with societal values.

Appendix G - Baseline

Baseline Prompt

System: You are a highly knowledgeable technical writer and researcher with expertise in synthesizing, analyzing, and presenting information in a structured and academically rigorous manner. Your task is to write a comprehensive survey paper that synthesizes the key findings, methodologies, and future directions of research based on a reference paper provided by the user. Focus on maintaining technical precision, clear organization, and advanced analysis. Ensure the paper provides added value by connecting the reference paper to broader research themes and offering insights into trends, gaps, and opportunities.

User: Your task is to generate a survey paper that includes the information in the reference papers. You should follow the conventional style of survey papers and make high quality analysis between different methods. You should also describe the evolution of each of the

relevant areas. Before you generate your outputs, you need to think over what is the outline of the survey paper. Then after you have an outline in mind, generate the survey paper based on the outline you had in mind. Before you output the survey paper, in you mind only do not output this, double check if the words you are about to output make sense. You should also include citations of the papers that you referenced in the survey paper where you used the information.

Reference paper: Al-qurishi Muhammad et al. AraLegal-BERT: A pretrained language model for Arabic Legal text. Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. URL: https://aclanthology.org/2022.nllp-1.31. Stephen Bach et al. “PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts”. In: Proceedings of ACL 2022: System Demonstrations. Dublin, Ireland, May 2022, pp. 93–104. DOI: 10.18653/v1/2022.acl-demo.9. Hyung Won Chung et al. “Scaling instruction-finetuned language models”. In: arXiv preprint arXiv:2210.11416 (2022). Jiaxi Cui et al. “Chatlaw: Open-source legal large language model with integrated external knowledge bases”. In: arXiv preprint arXiv:2306.16092 (2023). Junyun Cui et al. A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges. 2022. arXiv: 2204.04859 [cs.CL]. Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: arXiv preprint arXiv:1810.04805 (2018) Joao Dias et al. ˜ State of the Art in Artificial Intelligence applied to the Legal Domain. 2022. arXiv: 2204.07047. Virginia K. Felkner et al. Towards WinoQueer: Developing a Benchmark for Anti-Queer Bias in Large Language Models. 2022. arXiv: 2206.11484 [cs.CL] Daniel Martin Katz et al. Natural Language Processing in the Legal Domain. 2023. arXiv: 2302.12039 Niklas Muennighoff et al. Crosslingual Generalization through Multitask Finetuning. 2022. arXiv: 2211.01786. John J. Nay. Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans. 2023. arXiv: 2209.13020 [cs.CY] Alejandro Pena et al. ˜ Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs. 2023. arXiv: 2306.02864. Alec Radford et al. Learning Transferable Visual Models From Natural Language Supervision. 2021. arXiv: 2103.00020 Guilherme Moraes Rosa et al. “Billions of parameters are worth more than in-domain training data: A case study in the legal case entailment task”. In: arXiv preprint arXiv:2205.15172 (2022). Teven Le Scao et al. “Bloom: A 176b-parameter openaccess multilingual language model”. In: arXiv preprint arXiv:2211.05100 (2022). Zhenwei Shao et al. Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering. 2023. arXiv: 2303.01903. Xiaofei Sun et al. Text Classification via Large Language Models. 2023. arXiv: 2305.08377 Yi Tay et al. “Unifying language learning paradigms”. In: arXiv preprint arXiv:2205.05131 (2022). Ashish Vaswani et al. Attention Is All You Need. 2017. arXiv: 1706.03762

Baseline Final Paper Ouput

Evolution and Applications of Large Language Models in the Legal Domain

Abstract

The advent of large language models (LLMs) has revolutionized natural language processing (NLP), enabling groundbreaking applications in the legal domain. This survey explores the evolution of LLMs, their adaptation to legal tasks, and emerging challenges in ethical deployment. We provide a comparative analysis of methodologies, from pretraining architectures to domain-specific fine-tuning, and discuss future research directions.

---

1. Introduction

Large language models (LLMs) have transformed NLP, providing solutions to complex language-related tasks through innovations in neural architecture and scaling. Their applications in the legal domain are particularly promising, addressing tasks such as judgment prediction, contract analysis, and legal text classification. However, these applications face challenges unique to the legal field, including the need for precise interpretations and the integration of domain-specific knowledge.

This survey aims to outline the evolution of LLMs, highlight their role in legal NLP, and present a comprehensive analysis of recent advancements. The study also identifies challenges and ethical considerations in deploying LLMs for legal applications.

---

2. Foundations of Large Language Models

2.1 The Rise of Transformer Architectures

The transformer model introduced by Vaswani et al. (*2017*) revolutionized NLP with its attention mechanism, enabling parallelization and efficient handling of large datasets. Subsequent models, such as BERT (*Devlin et al., 2018*), expanded on this architecture, introducing bidirectional training for contextual understanding.

2.2 Scaling and Fine-Tuning

Recent advancements focus on scaling LLMs, with models like GPT-3 and Bloom (*Le Scao et al., 2022*) demonstrating the benefits of massive parameter counts and diverse training datasets. Instruction tuning further enhances model adaptability, as shown by Chung et al. (*2022*), enabling models to generalize across varied tasks.

---

3. Legal Applications of LLMs

3.1 Legal Text Classification

Domain-specific adaptations of LLMs have been instrumental in classifying legal texts. Pena et al. (*2023*) demonstrate the effectiveness of large-scale models in public affairs classification. Crosslingual fine-tuning (*Muennighoff et al., 2022*) further extends this capability to multilingual legal corpora, reducing barriers in cross-border legal systems.

3.2 Legal Judgment Prediction

Legal judgment prediction (LJP) involves analyzing case facts to predict judicial outcomes. Cui et al. (*2022*) surveyed datasets, metrics, and model architectures for LJP, highlighting challenges in interpretability and bias. Advances in LJP suggest a shift towards hybrid models integrating rule-based and data-driven approaches.

3.3 Knowledge Integration

LLMs like ChatLaw (*Cui et al., 2023*) incorporate external legal knowledge bases to enhance reasoning capabilities. By bridging pretrained representations with domain-specific knowledge, these models achieve superior performance in retrieval-based tasks.

3.4 Legal Case Entailment

Rosa et al. (*2022*) investigated whether large-scale pretraining outweighs domain-specific data in legal case entailment tasks. Their findings suggest that while pretraining provides a strong baseline, fine-tuning on curated legal datasets remains essential for achieving optimal accuracy.

---

4. Emerging Trends and Challenges

4.1 Bias Mitigation and Ethics

Bias remains a critical issue in deploying LLMs for legal tasks. Felkner et al. (*2022*) introduced WinoQueer, a benchmark for identifying anti-queer bias, demonstrating the broader implications of fairness in legal AI systems. Addressing bias requires holistic evaluation and dataset curation to ensure equitable outcomes.

4.2 Scaling and Instruction Tuning

Instruction-tuned models, such as those by Chung et al. (*2022*), highlight the potential of task adaptability. Prompt engineering (*Bach et al., 2022*) complements these models by optimizing input instructions, further enhancing task-specific performance.

4.3 Multimodal and Cross-Domain Applications

Combining textual and visual reasoning has gained traction in the legal domain. Shao et al. (*2023*) explored heuristic-based prompting for visual question answering, while Radford et al. (*2021*) emphasized transferable learning across modalities. These advancements pave the way for richer multimodal legal analyses.

---

5. Comparative Analysis of Methods

Comparing methods across benchmarks reveals a trade-off between scalability and domain specificity. While generalized models like Bloom (*Le Scao et al., 2022*) excel in multilingual scenarios, specialized models such as AraLegal-BERT (*Al-qurishi et al., 2022*) outperform in tasks requiring fine-grained legal understanding.

Table 1 presents a summary of key models, datasets, and their performance metrics:

Model	Task	Strengths	Limitations
AraLegal-BERT	Legal text classification	Domain-specific pretraining	Limited multilingual support
ChatLaw	Knowledge integration	External knowledge incorporation	Computationally expensive
GPT-3	General legal tasks	Scalability and generalization	Limited interpretability
Instruction GPT	Task-specific fine-tuning	Adaptability across diverse tasks	Relies on well-structured instructions

---

6. Conclusion and Future Directions

This survey highlights the evolution of LLMs and their transformative impact on the legal domain. While significant advancements have been made, challenges such as ethical bias, data scarcity, and computational constraints persist. Future research should focus on:

1. Developing equitable models through robust bias detection and mitigation strategies.
2. Enhancing domain-specific adaptability via knowledge integration.
3. Exploring multimodal and cross-domain synergies for richer legal analysis.

By addressing these challenges, LLMs have the potential to revolutionize legal practices, enabling fairer, faster, and more accessible legal systems.