

LexKeyPlan: Planning with Keyphrases and Retrieval Augmentation for Legal Text Generation: A Case Study on European Court of Human Rights Cases

Santosh T.Y.S.S, Elvin Quero Hernandez

School of Computation, Information, and Technology
Technical University of Munich, Germany
{santosh.tokala, elvin.quero}@tum.de

Abstract

Large language models excel at legal text generation but often produce hallucinations due to their sole reliance on parametric knowledge. Retrieval-augmented models mitigate this by providing relevant external documents to the model but struggle when retrieval is based only on past context, which may not align with the model’s intended future content. We introduce LexKeyPlan, a novel framework that integrates anticipatory planning into generation. Instead of relying solely on context for retrieval, LexKeyPlan generates keyphrases outlining future content serving as forward-looking plan, guiding retrieval for more accurate text generation. This work incorporates planning into legal text generation, demonstrating how keyphrases—representing legal concepts—enhance factual accuracy. By structuring retrieval around legal concepts, LexKeyPlan better aligns with legal reasoning, making it particularly suited for legal applications. Using the ECHR corpus as case study, we show that LexKeyPlan improves factual accuracy and coherence by retrieving information aligned with the intended content.

1 Introduction

Recent advancements in large language models (LLMs) have led to their adoption in the legal domain for tasks such as sifting through case briefs, expediting legal research, drafting contracts, and formulating litigation strategies (Dahl et al., 2024). Despite their ability to pass bar and law school exams (Katz et al., 2024; Martínez, 2024), perform statutory reasoning and interpretation (Blair-Stanek et al., 2023; Engel and Mcadams, 2024), and apply legal reasoning frameworks like Issue-Rule-Application-Conclusion (IRAC) (Kang et al., 2023; Guha et al., 2024), LLMs still struggle with factual inaccuracies and hallucinations of legal knowledge.

Traditional LLMs generate responses based on the input context and their parametric knowledge

(Radford et al., 2019; Zhang et al., 2022; Jiang et al., 2023; Touvron et al., 2023), but this approach often fails to ensure factual accuracy. Retrieval-augmented generation (RAG) addresses this issue by retrieving external documents based on the context, allowing the model to condition its responses on more reliable information (Lewis et al., 2020; Izacard et al., 2023; Borgeaud et al., 2022; Ram et al., 2023). However, RAG relies entirely on the retrieval mechanism’s ability to fetch relevant knowledge, which may not always align with the content the model intends to generate. This limitation is particularly critical in legal text generation, where context alone may not provide sufficient signals to retrieve legally relevant information (Magesh et al., 2024; Santosh et al., 2025).

To overcome this challenge, we propose LexKeyPlan, a novel framework that introduces an anticipatory planning stage into the generation process. Instead of relying solely on context for retrieval, LexKeyPlan first generates keyphrases that represent legal concepts outlining the intended content of the response. These keyphrases serve as a forward-looking content plan, guiding the retrieval of relevant documents from external sources. By shifting the retrieval process from purely reactive to proactively structured around anticipated content, LexKeyPlan better aligns retrieval with legal reasoning, reduces reliance on misleading contextual information, and improves coherence in long-form legal text generation.

This work incorporates content planning into legal text generation, highlighting how keyphrases—representing legal concepts—can enhance retrieval, factual grounding, and structured reasoning. Using the European Court of Human Rights (ECHR) case law corpus, we demonstrate that LexKeyPlan significantly improves legal text generation by reducing hallucinations and enhancing retrieval relevance. Our findings suggest that anticipatory content planning through legal con-

cepts or keyphrases is crucial for legal applications, where precise retrieval and structured reasoning are essential for generating legally sound text.

2 LexKeyPlan

We propose a three-step framework for legal text generation, LexKeyPlan. Unlike traditional models that generate responses directly from context, LexKeyPlan introduces an intermediate planning stage using keyphrases. These keyphrases serve as a forward-looking blueprint, outlining the intent of the future response. By anticipating the necessary content, they guide the retrieval of relevant external documents, ensuring that the generated text is both factually accurate and grounded. This approach contrasts with existing retrieval augmentation methods, which typically use context alone to retrieve relevant information, potentially limiting the relevance of the retrieved content for response generation, in long form text generation.

Formally, let x be the input context. The model first generates a content plan c for the response, which can be represented as $p(c | x)$. This content plan, consisting of keyphrases, succinctly captures the essence of the future response. Then the generated plan c is used as query to retrieve relevant documents d from external knowledge sources. The final response y is produced by conditioning the model on the input context, the content plan and the retrieved documents, represented as $p(y | c, p, d)$.

Training the Content Plan Generation Module:

We train a language model using the input context followed by the content plan. To obtain the target content plan for each input, we utilize the target response text and extract keyphrases from it. We explore two keyphrase extraction algorithms to create the supervision dataset for training this content plan generation module: TextRank (Mihalcea and Tarau, 2004), a graph-based ranking algorithm inspired by PageRank, and KeyBERT (Grootendorst, 2020), leveraging embeddings from pre-trained models.

Retriever Module: We investigate two off-the-shelf retrievers: 1) BM25 (Robertson et al., 2004), a sparse lexical-based model, and 2) GTR (Ni et al., 2021), an embedding-based dense retriever model built on T5-XXL (Raffel et al., 2020). We provide the concatenated list of keyphrases (content plan) as query to retrieve the relevant documents from datastore, built from ECHR judgement documents.

Training the Response Generation Module: The language model is trained using the input con-

text, content plan, retrieved relevant documents followed by the target response. Since we do not have a gold-standard content plan or relevant documents for training, we extract keyphrases from the target response to serve as the content plan and use the target response as a query to retrieve pseudo-golden relevant documents. During inference, the model uses the actual generated content plan to retrieve relevant documents.

This three-step framework offers controllability, enabling users to adjust the content plan or retrieved documents to steer the model’s response.

3 Experiments

3.1 Dataset & Metrics

ECHR CaseLaw consists of case judgments heard by the European Court of Human Rights. We use the latest cleaned version of this dataset from Santosh et al. (2024a), which contains 15,729 cases in English from 1960 to July 28, 2022. We partition the dataset chronologically into training, validation, and test splits, following the provided temporal boundaries: the training set includes 11,615 documents from 1960 to 2016, the validation set contains 1,688 documents from 2017–2018, and the test set comprises 2,426 judgments spanning 2018–2022. Each case is segmented into two main sections: *The Facts*, which details the factual background of each case and *The Law*, which presents the legal reasoning justifying the outcome concerning alleged violations of specific ECHR articles. Our task involves generating the *The Law* (reasoning) section based on the *The Facts* section. While generating an entire section would be ideal, evaluating such large outputs against reference texts poses significant challenges. Therefore, we focus on paragraph-level generation to facilitate more accurate evaluations. Specifically, each paragraph in the *The Law* section serves as the target response, while the input context consists of the *The Facts* section and all preceding paragraphs from *The Law* section up to the target paragraph. This paragraph-level setup results in 82,597 training samples, 12,709 validation samples, and 17,470 test samples. During training, each paragraph is appended with an $\langle |end_of_paragraph| \rangle$ token, allowing us to generate the next paragraph up to this marker during open-ended generation. We use Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) as our base LM for both plan and response generator and follow parameter efficient fine-tuning using

	Plan Generator	Retriever	R-1 / R-2 / R-L	BERT Score	Align Score	Coh. / Flu.
a	-	-	0.32 / 0.13 / 0.21	0.77	0.52	0.69 / 0.61
b	TextRank	-	0.33 / 0.13 / 0.22	0.72	0.54	0.74 / 0.66
c	KeyBERT	-	0.34 / 0.14 / 0.21	0.74	0.58	0.72 / 0.64
d	-	BM25	0.37 / 0.14 / 0.23	0.78	0.64	0.63 / 0.61
e	-	GTR	0.37 / 0.15 / 0.24	0.79	0.66	0.64 / 0.63
f	TextRank	BM25	0.37 / 0.17 / 0.26	0.79	0.7	0.72 / 0.65
g	TextRank	GTR	0.39 / 0.15 / 0.24	0.78	0.72	0.75 / 0.65
h	KeyBERT	BM25	0.40 / 0.15 / 0.23	0.78	0.73	0.74 / 0.66
i	KeyBERT	GTR	0.39 / 0.16 / 0.24	0.80	0.75	0.76 / 0.65

Table 1: Comparison of LexKeyPlan with Baseline Approaches. The "Plan Generator" column specifies the keyphrase extraction algorithm used to generate supervision signals for training both the planning and response generation modules. The "Retriever" column indicates the retrieval method employed during inference and for obtaining relevant documents during training of the response generator. Note that the keyphrase-based content plan is automatically generated by the model during inference.

LoRA (Hu et al., 2021). Implementation details are provided in App. B.

We evaluate the quality of the generated paragraphs: ROUGE-1,2,L (Lin, 2004) for lexical overlap with the reference paragraph, BERTScore (Zhang et al., 2019) for semantic similarity between generated and reference paragraph, AlignScore (Zha et al., 2023) for factual consistency based on a unified alignment function between the reference and generated text and UniEval (Zhong et al., 2022) that evaluates coherence and fluency of the generated paragraph with respect to the context.

3.2 Results

To demonstrate the effectiveness of LexKeyPlan, we compare our method with several baselines: (a) traditional fine-tuning without keyphrase-based content planning or retrieval augmentation; (b, c) models with keyphrase-based planning, followed by response generation using the context and the generated plan. Two keyphrase extraction algorithms are used as supervision signals for training the planning and response generators; (d, e) models without keyphrase planning that directly retrieve relevant documents based on context and generate responses using the retrieved documents. The response generation models are trained with supervision from their respective retrievers; (f, g, h, i) our three-staged LexKeyPlan method, with different combinations of plan generators and retrievers for training the response generation.

Effect of Keyphrase-based Content Planning: As seen in Table 1, keyphrase-based content planning (b, c) outperforms traditional fine-tuning (a) in terms of coherence, fluency, and faithfulness (AlignScore), while remaining comparable on lexical similarity (ROUGE). This keyphrase planning

mechanism allows the model to anticipate and structure its response around key concepts in the upcoming text. By separating the planning and generation stages, the model’s cognitive load is reduced, as it no longer needs to generate relevant information and maintain coherence simultaneously. Instead, it can focus on each task independently, resulting in more coherent and contextually accurate outputs. Additionally, this structured planning mitigates hallucinations by providing a clear blueprint for the text, helping the model stay grounded in the key content and reducing the likelihood of producing irrelevant information. Between the two keyphrase algorithms, KeyBERT outperforms on content-related metrics, while TextRank excels in stylistic aspects like coherence and fluency.

Effect of Retrieval Augmentation: Retrieval augmentation (d, e), which utilizes context as the query, effectively addresses hallucination issues by grounding the generated text in external knowledge, resulting in improved ROUGE scores and faithfulness (AlignScore). However, this approach also leads to a decrease in coherence and fluency, likely due to the model’s challenge in integrating necessary information from the text while maintaining coherent output. Among the retrieval methods, embedding-based GTR consistently outperforms BM25 across all metrics, revealing the limitations of lexical-based retrievers in capturing semantic nuances. Nonetheless, the smaller gap for BM25 highlights the high lexical overlap and repetitive nature of legal concepts in the documents (Santosh et al., 2024a), making BM25 a scalable and efficient option for legal text retrieval.

LexKeyPlan: Integrating keyphrase-based content planning with retrieval augmentation significantly improves both content quality and stylistic

	Plan	Retriever	R-1 / R-2 / R-L	BERT Score	Align Score	Coh. / Flu.
a	-	-	0.23 / 0.06 / 0.15	0.72	0.44	0.70 / 0.64
b	-	BM25	0.25 / 0.08 / 0.15	0.74	0.56	0.60 / 0.62
c	-	GTR	0.25 / 0.08 / 0.17	0.75	0.58	0.62 / 0.62
d	✓	-	0.24 / 0.08 / 0.16	0.74	0.50	0.72 / 0.62
e	✓	BM25	0.26 / 0.10 / 0.18	0.76	0.59	0.58 / 0.61
f	✓	GTR	0.27 / 0.10 / 0.16	0.75	0.58	0.59 / 0.60

Table 2: Effect of Integrating Keyphrase-Based Content Planning and Retrieval Augmentation in Zero-Shot Setting. The ‘Plan’ column specifies whether the model is prompted to generate keyphrase-based content plan for the next response. The Retriever column identifies the retrieval method employed.

	Similarity	Exact Match
Zeroshot	0.67	0.18
Fine-tuned with TextRank	0.72	0.36
Fine-tuned with KeyBERT	0.78	0.42
TextRank (Oracle)	0.96	1.0
KeyBERT (Oracle)	0.95	1.0

Table 3: Evaluation of Content Plan Quality.

tic metrics (statistically significant improvements as per Wilcoxon signed-rank test at a 95% confidence interval over the baselines without planning or retrieval augmentation). By using keyphrases as queries to anticipate forthcoming text, LexKeyPlan offers two key advantages. First, it enhances retrieval accuracy by focusing on key concepts, leading to more relevant documents compared to using context alone, which may miss crucial information for the anticipated text. Second, separating the planning and generation tasks allows the model to produce more coherent text. Rather than performing retrieval, integration, and generation simultaneously, LexKeyPlan enables the model to first plan around keyphrases, thereby reducing cognitive load and resulting in more coherent and contextually accurate outputs. Among the methods evaluated, KeyBERT provided superior supervision for plan generation, while GTR outperformed BM25, highlighting the benefits of embedding-based models over traditional lexical approaches.

Zero-shot Experiments: To evaluate whether integrating content planning and retrieval augmentation into pre-trained models in a zero-shot setting enhances performance, we compare several approaches in Table 2: (a) prompting the pre-trained model to generate a response based solely on the context; (b, c) prompting the model to generate a response based on both the context and relevant documents retrieved from the context; (d) prompting the model to first create a content plan in the form of keyphrases and then generate a response based on this plan; (e, f) prompting the model to

generate a content plan, use this plan to retrieve relevant documents, and then generate a response based on both the plan and documents.

As expected, retrieval augmentation (b, c) directly enhances faithfulness (AlignScore) and content quality but results in a decrease in coherence. This drop underscores the difficulty of maintaining coherence while incorporating relevant information. Introducing content planning (d) helps the model manage these tasks by breaking them into sequential steps, which enhances both faithfulness and coherence. When combining content planning with retrieval augmentation (e, f), there is a slight improvement in content quality metrics, though stylistic scores decrease. This suggests that zero-shot models may face challenges with the complexity of handling multiple instructions—such as following a content plan and integrating relevant information. **Analysis of Keyphrases in Content Planning:** We evaluate the quality of keyphrases generated through (a) Zero-shot prompting the pre-trained model, and (b, c) using a fine-tuned model supervised with TextRank and KeyBERT, by computing the embedding similarity of each keyphrase to the actual target (i.e., the next paragraph). We report both average similarity and exact match (if the keyphrase appears in the target). Additionally, we present oracle metrics (d, e), where keyphrases are directly extracted from the target paragraph using TextRank and KeyBERT. Results show that fine-tuning with KeyBERT-extracted keyphrases leads to more relevant keyphrase generation with respect to future content, providing informed plans for document retrieval and response generation.

4 Conclusion

We introduce LexKeyPlan, a three-step framework for legal text generation integrating keyphrase-based content planning with retrieval augmentation. Unlike traditional methods that rely solely on context for retrieval and generation, LexKeyPlan first

generates a content plan outlining the intended future content. This plan not only guides to produce coherent response but also the retrieval of relevant documents. LexKeyPlan through its anticipatory planning ensures that the generated text is both contextually coherent and factually accurate. as demonstrated with the ECHR case law corpus.

Limitations

LexKeyPlan currently utilizes a relatively simple content planning approach based on keyphrase extraction using general-purpose algorithms such as TextRank and KeyBERT. These methods, while effective in domains like news and scientific literature, were not specifically developed for legal language. As a result, they may fail to capture the nuanced terminologies, hierarchical concepts, and contextual dependencies intrinsic to legal texts. Future work could explore legal-domain-specific keyphrase extraction techniques—such as those proposed in [Mandal et al. \(2017\)](#)—which better accommodate the structural and semantic idiosyncrasies of legal discourse.

Our use of general extractors and retrievers was intentional: we aimed to demonstrate LexKeyPlan’s core contribution—anticipatory retrieval guided by future intent—without conflating gains from domain adaptation. Employing legal-specific strategies such as regular-expression-based legal phrase extractors or retrievers fine-tuned on legal corpora may have artificially inflated performance, obscuring the effectiveness of our planning and retrieval coordination. Instead, we used BM25 and GTR as retrievers in addition to TextRank and KeyBERT as keyphrase extractors to validate that our approach can generalize across jurisdictions and is not overly reliant on domain-specific tuning. That said, this choice also introduces limitations. General retrievers often struggle with legal-specific vocabulary, temporal relevance, and the authoritativeness of retrieved materials. In legal contexts, it is not enough to retrieve documents that are merely topically related—their precedential weight, binding status, and jurisprudential relevance must also be considered. For instance, a retrieved case may be outdated, overruled, or of low precedential value, which undermines its utility in supporting legal reasoning. These subtleties are particularly critical in legal domains, where the quality and authority of supporting material can significantly affect argument strength. Future work should explore

authority-aware and time-sensitive retrieval mechanisms, potentially through metadata-aware filtering, precedent strength modeling, or temporal reranking that prioritizes current and legally influential sources. ([Santosh et al., 2024c,b](#)).

More sophisticated content planning approaches could significantly enhance the effectiveness of LexKeyPlan. One promising avenue is the use of graphical representations derived from legal concept networks. These networks visually map out legal concepts and their interrelationships, creating a structured framework that can guide content generation. Incorporating a specialized legal thesaurus into the planning process can further refine content generation. A legal thesaurus provides a curated vocabulary of legal terms and concepts, enhancing the precision of keyphrases and improving the contextual relevance of the generated content. Improving the integration of content plans during the text generation phase may also involve adopting constrained decoding techniques. Constrained decoding methods enforce adherence to predefined content plans, ensuring that the generated text aligns closely with the planned structure and content. This approach can help maintain coherence and fidelity, reducing deviations from the intended content and improving output quality.

Currently, LexKeyPlan’s performance is evaluated using established metrics like ROUGE, BERTScore, and AlignScore. While these metrics offer quantitative insights, they may not fully capture the complexities of legal content. Future research could develop domain-specific evaluation metrics tailored to the legal field. Another limitation of our study is the absence of direct validation by legal experts in the assessment of outputs, which we could not perform due to lack of access to legal experts. Lastly, we selected the ECHR dataset for evaluation due to its explicit segmentation between case facts and legal reasoning. This clean structure enabled a well-defined simulation of a real-world task: generating legal reasoning given only the case facts. Such a setup mirrors copilot-style applications where LLMs assist lawyers and judges in drafting legal arguments. Other legal datasets often lack this granularity, intermixing facts and reasoning, which makes it difficult to isolate and measure the model’s reasoning capabilities. However, this focus on ECHR also narrows the scope of our evaluation. Future research should assess LexKeyPlan’s adaptability across multiple legal systems and languages, including those with different document

structures and procedural norms.

Ethics Statement

The ECHR dataset used in this study is publicly available and widely employed in legal NLP research. However, it is important to note that this dataset is not anonymized and contains real names and details of the parties involved in the cases. Although we do not foresee direct harm resulting from our experiments, handling such sensitive data necessitates careful consideration of privacy and ethical implications. Researchers must be diligent in protecting individual privacy and managing sensitive information responsibly, even when using publicly accessible datasets.

LLMs utilized in legal contexts have significant implications for both legal professionals and the public. Despite their advanced capabilities, these models carry inherent risks. They may reproduce and amplify biases inherent in their training data, leading to outputs that reflect historical inequalities or stereotypes. Additionally, LLMs may produce incorrect or misleading information, particularly in complex legal scenarios where precision is essential. Our research focuses on enhancing rather than replacing human expertise. It is crucial to apply LLM-generated outputs with careful scrutiny, ensuring that these tools are complemented by thorough human oversight and critical evaluation. Our aim is to improve legal text generation while recognizing that AI must be a supportive tool, used with an awareness of its limitations and potential impact on legal practice.

References

- Griffin Adams, Alexander R Fabbri, Faisal Ladhak, Kathleen McKeown, and Noémie Elhadad. 2023. Generating edu extracts for plan-guided summary re-ranking. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 2680. NIH Public Access.
- Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization.
- Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.
- Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models journal of legal analysis (forthcoming).
- Christoph Engel and Richard H Mcadams. 2024. Asking gpt for the ordinary meaning of statutory terms. *MPI Collective Goods Discussion Paper*, (2024/5).
- Ameya Godbole, Nicholas Monath, Seungyeon Kim, Ankit Singh Rawat, Andrew Mccallum, and Manzil Zaheer. 2024. Analysis of plan-based retrieval for grounded text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13101–13119.
- Maarten Grootendorst. 2020. [Keybert: Minimal keyword extraction with bert](#).
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.
- Markus Guhe. 2020. *Incremental conceptualization for language production*. Psychology Press.
- Eduard H Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial intelligence*, 63(1-2):341–385.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuezhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Fantine Huot, Joshua Maynez, Shashi Narayan, Reinald Kim Amplayo, Kuzman Ganchev, Annie Priyadarshini Louis, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Text-blueprint: An interactive platform for plan-based conditional generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 105–116.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? *arXiv preprint arXiv:2310.14880*.
- Nikiforos Karamanis. 2004. Entity coherence for descriptive text structuring.
- Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.
- Rodger Kibble and Richard Power. 2004. Optimizing referential coherence in text generation. *Computational Linguistics*, 30(4):401–416.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Anirban Laha, Parag Jain, Abhijit Mishra, and Karthik Sankaranarayanan. 2020. Scalable micro-planned generation of discourse from structured data. *Computational Linguistics*, 45(4):737–763.
- Willem JM Levelt. 1993. *Speaking: From intention to articulation*. MIT press.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zhengyuan Liu and Nancy F Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. *arXiv preprint arXiv:2109.13070*.
- Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.
- Arpan Mandal, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2017. Automatic catchphrase identification from legal court case documents. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2187–2190.
- William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Eric Martínez. 2024. Re-evaluating gpt-4’s bar exam performance. *Artificial Intelligence and Law*, pages 1–24.
- Kathleen McKeown and Dragomir R Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82.
- Kathleen R McKeown. 1985. Discourse strategies for generating natural-language text. *Artificial intelligence*, 27(1):1–41.
- Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O’Donnell. 1998. Experiments using stochastic search for text planning. In *Proceedings of the 9th International General Workshop*, pages 98–107. ACL Anthology.
- Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.
- Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. *arXiv preprint arXiv:1904.03396*.
- Shashi Narayan, Joshua Maynez, Jakub Adamek, Daniele Pighin, Blaz Bratanić, and Ryan McDonald. 2020. Stepwise extractive summarization and planning with structured transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4143–4159.
- Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.
- Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.
- Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. *arXiv preprint arXiv:1906.03221*.
- Ratish Puduppully, Yao Fu, and Mirella Lapata. 2022. Data-to-text generation with variational sequential planning. *Transactions of the Association for Computational Linguistics*, 10:697–715.

- Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.
- Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49.
- TYSS Santosh, Rashid Gustav Haddad, and Matthias Grabmair. 2024a. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. *arXiv preprint arXiv:2404.00596*.
- TYSS Santosh, Kevin D Ashley, Katie Atkinson, and Matthias Grabmair. 2024b. Towards supporting legal argumentation with nlp: Is more data really all you need? In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 404–421.
- TYSS Santosh, Isaac Misael Olguan Nolasco, and Matthias Grabmair. 2025. Lecopcr: Legal concept-guided prior case retrieval for european court of human rights cases. *arXiv preprint arXiv:2501.14114*.
- TYSS Santosh, Tuan-Quang Vuong, and Matthias Grabmair. 2024c. Chronoslex: Time-aware incremental training for temporal generalization of legal classification tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3022–3039.
- Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Appendix

A Related Work

Content Planning for Text Generation: Structured planning is considered a critical link in organizing content effectively before realization (Reiter and Dale, 1997), much like humans plan at a higher level than individual words, as evidenced by psycholinguistic studies (Levelt, 1993; Guhe, 2020). Earlier approaches incorporated various planning representations, such as Rhetorical Structure Theory (Mann and Thompson, 1988; Hovy, 1993) and MUC-style representations (McKeown and Radev, 1995), discourse trees (Mellish et al., 1998), entity transitions (Kibble and Power, 2004; Barzilay and Lapata, 2008), sequences of propositions (Karamanis, 2004), schemas (McKeown, 1985) and lexical chains (Barzilay and Elhadad, 1997).

Recent works in the data-to-text generation task divide it into two distinct phases: planning and realization of natural language text (Moryossef et al., 2019). Puduppully et al. (2019) and Laha et al. (2020) proposed micro-planning strategies, where they first establish a content plan based on a sequence of records and then generate a summary conditioned on that plan. Similar planning approaches have been explored for entity realization (Puduppully et al., 2019). Additionally, Puduppully and Lapata (2021) advocated for macro-planning as

a means of organizing high-level document content, either in text form or within latent space (Puduppully et al., 2022).

In summarization, various plan representations have been investigated. Narayan et al. (2020) treat step-by-step content selection as a planning component, generating sentence-level plans through extract-then-abstract methods, and even sub-sentence-level plans using elementary discourse units (Adams et al., 2023). Further research by Narayan et al. (2021); Liu and Chen (2021) introduced intermediate plans using entity chains—ordered sequences of entities mentioned in the summary. More recently, Narayan et al. (2023); Huot et al. (2023) conceptualized text plans as a sequence of question-answer pairs, which serve as blueprints for content selection (i.e., what to say) and planning (i.e., in what order) in summarization tasks.

Recently, Godbole et al. (2024); Shao et al. (2024) have explored planning-based approaches for text generation in tasks such as expository writing and biographical summaries. However, legal text generation presents unique challenges that remain unexplored in this context. Unlike traditional QA tasks, which involve focused information-seeking queries, legal text generation requires reasoning over long, complex case facts to construct well-grounded arguments to derive conclusions. This makes planning particularly valuable, as it can help structure retrieval and generation to ensure coherence, factual consistency, and legal validity. To the best of our knowledge, no prior work has investigated planning-based approaches for legal text generation.

B Implementation Details

We implemented KeyBERT¹ with n-gram range of (1, 3). We enable Maximal Marginal Relevance (MMR) with a diversity value 0.7 to balance relevance and diversity. For TextRank (Mihalcea and Tarau, 2004), we utilize the implementation provided in the [GitHub repository](#).

We perform LoRA fine-tuning with Mistral model for both our plan and response generator with an alpha and rank configuration in {8, 16} and dropout of 0.1. We sweep over learning rates { $1e-5$, $3e-5$, $5e-5$, $1e-4$, $3e-4$ } and evaluating its performance on the validation set using 10K validation steps. We train the model end-to-end for

¹<https://github.com/MaartenGr/KeyBERT>

one epoch using Adam Optimizer (Kingma and Ba, 2014). We perform inference over models with a temperature of 0.7 and nucleus sampling of 0.95.

For retrieval, we dynamically populate the candidate datastore by excluding documents dated after the query case and including only those available up to the query date, thereby simulating a realistic setting. We segment these documents into paragraph chunks, to index them for retrieval. For retrieval, we use the top-3 retrieved paragraphs for both methods. Prompt 1 and 2 detail the prompts used for content plan generation and response generations modules respectively.

C Case Study

We present three case studies that demonstrate how LexKeyPlan performs in complex legal scenarios. These examples show how keyphrase planning enables doctrinally grounded retrieval and controllable generation, while also revealing important limitation of its inability to account for temporal relevance.

Case Study A: Enhancing Doctrinal Accuracy Through Keyphrase-Guided Retrieval

We examine a scenario where an individual is convicted for inflammatory statements made on social media and brings a claim under Article 10 (freedom of expression). The government argues that the speech in question incited hatred and thus falls outside the protection of the Convention. This case hinges on the distinction between political criticism and hate speech—a critical threshold question that determines whether the speech is covered by Article 10 or excluded under Article 17, which bars the abuse of rights. When the model operates with context-only retrieval, it relies solely on the case facts—namely, that the applicant was prosecuted for a speech act and invokes Article 10. This prompts the retriever to surface general freedom of expression jurisprudence such as *Handyside v. UK* (1976), *Sunday Times v. UK* (1979), and *Oberschlick v. Austria* (1997), all of which affirm the importance of free expression but apply a proportionality framework under Article 10(2). As a result, the generated response reasons as follows: "The Court applies the proportionality test under Article 10(2) to determine whether the restriction was necessary in a democratic society. Given the precedent set in *Handyside v. UK*, the speech must be examined in light of public morals and state discretion. Therefore, the speech in this case may

be protected under Article 10." This response is legally incorrect: it prematurely applies a balancing test without addressing the more fundamental question of admissibility under Article 10. Specifically, if the speech constitutes hate speech or denial of historical atrocities, it may be excluded altogether under Article 17, rendering Article 10 inapplicable. The model's failure to retrieve the right precedents results in doctrinal error. In contrast, with keyphrase-guided planning, the model is given key concepts such as "legitimate aim (public safety vs political criticism)," "ECHR Article 17," "abuse of rights doctrine," and "margin of appreciation." These guide retrieval toward cases that directly address the legal test relevant to hate speech and the abuse of rights. The retrieved case set now includes *Garaudy v. France* (2003), *Perinçek v. Switzerland* (2015), and *Leroy v. France* (2008)—cases that explicitly delineate when speech is excluded under Article 17 or subjected to Article 10 balancing. With this doctrinally appropriate retrieval, the model produces the following improved response: "The Court first determines whether the applicant's speech falls within the scope of Article 10 or is excluded under Article 17, which bars the abuse of Convention rights. In *Garaudy v. France*, the Court held that speech denying historical atrocities such as the Holocaust is incompatible with the values of the Convention and is not protected under Article 10. If the applicant's statements incite hatred or deny established facts, they may fall under Article 17 and be excluded from protection. However, if the speech qualifies as political criticism—even if provocative—it may still enjoy Article 10 protection, subject to a proportionality assessment under Article 10(2), as seen in *Leroy v. France*." This output demonstrates significantly improved legal reasoning: it reflects the correct structure of analysis (i.e., first evaluating exclusion under Article 17 before engaging in Article 10 balancing), references the right cases, and avoids premature or inappropriate application of proportionality. It also recognizes the margin of appreciation doctrine and the need to distinguish incitement from legitimate political critique. This case study underscores a core strength of LexKeyPlan: by prompting retrieval based on key doctrinal concepts rather than surface-level case similarity, the model can reason more accurately and in line with legal standards.

Case Study B: Controlling Legal Reasoning via Structured Planning In a case where a political figure challenged the deplatforming of their ac-

count by a private social media company, alleging a violation of Article 10. The case raised ambiguity about whether private actors performing public functions are subject to free speech obligations. LexKeyPlan was able to demonstrate controllable reasoning by varying the content plan. When the keyphrases included "de facto public forum," "chilling effect," and "corporate control over speech," the model retrieved cases such as *Appleby and Others v. UK* (2003), which address the public forum doctrine in private spaces. The generated response argued that: "Social media platforms serve as a de facto public forum, where political discourse primarily unfolds. As established in *Appleby v. UK*, while private entities are not always bound by public free speech obligations, their restrictions may be scrutinized if they effectively eliminate a person's ability to communicate elsewhere. Given the dominance of social platforms in modern public debate, banning a political figure from these spaces constitutes an undue restriction, creating a chilling effect that stifles democratic discourse." When the content plan was changed to include keyphrases such as "platform autonomy," "harm prevention," "corporate free expression," "content moderation policies," and "disinformation regulation," the retriever instead surfaced *Delfi AS v. Estonia* (2015), which upholds the rights of private platforms to moderate harmful content. The corresponding generated response shifted the framing: "Social media platforms, as private entities, have the right to moderate content under their terms of service. Prior rulings, including *Delfi AS v. Estonia*, affirm that private actors are not bound by the same restrictions as state actors. Given the risks of disinformation and harm to democratic integrity, content moderation is a necessary safeguard. The platform's decision aligns with legitimate policy goals and does not constitute a violation of Article 10." This illustrates LexKeyPlan's strength in enabling users to explore competing legal framings by adjusting the planning stage, allowing the system to retrieve and ground responses in support of different legal arguments, rather than being constrained by the framing present in the input context alone.

Case Study C: Limitations in Capturing Temporal Relevance Despite these benefits, LexKeyPlan remains limited in its ability to account for temporal and relevance. In a case involving the extradition of an individual to a country with a poor human rights record, the applicant claimed that the transfer would violate Article 3, citing the

risk of inhuman or degrading treatment. While the keyphrases “inhuman treatment,” “extradition risk,” and “diplomatic assurances” did help retrieve landmark cases like *Soering v. UK* (1989), they failed to surface more recent jurisprudence such as *Othman (Abu Qatada) v. UK* (2012) or *Paposhvili v. Belgium* (2016), which significantly recalibrated the legal standards for assessing risk and evaluating the sufficiency of diplomatic assurances. The limitation here arises from LexKeyPlan’s reliance on topical keyphrase matching without accounting for the temporal dimension of legal knowledge. As legal doctrines evolve, retrieval systems must prioritize not only relevance but recency, particularly in areas where the law is in flux.

You are a judge at the European Court of Human Rights (ECHR), drafting a judicial judgment. The paragraphs below reflect your reasoning thus far. Your task is to outline the next topics you intend to address in the next paragraph of the judgment.

Instructions:

- Generate a maximum of 5 topics. Place higher-priority topics at the beginning of the list.
- Ensure the list is formatted as a comma-separated string.
- Ensure the topics align closely with the content of the previous paragraphs and the overall context of the judgment.

Input:

{{PREVIOUS PARAGRAPHS}}

Response:

Prompt 1: Prompt for Keyphrase based content plan generation

You are a judge at the European Court of Human Rights (ECHR), drafting a judicial judgment. Below are paragraphs you have already written as part of your reasoning. Your task is to write the next paragraph to continue developing the judgment.

To assist you, content plan for the next paragraph in the form of keyphrases and excerpts from relevant external cases have been provided. These cases may offer supporting arguments or precedents to strengthen your reasoning, but their use is optional. Your primary focus should be on ensuring the logical flow and coherence of the judgment.

Instructions:

- Refer to external cases only if they add value to your reasoning or support the legal principles being applied. When citing an external case, include a proper citation at the end. For example: (Case of X v. Y). If referencing a specific section or paragraph from an external case, denote it clearly using the section symbol and the paragraph number.
- Ensure that the new paragraph is directly related to the content of the previous paragraphs and the broader context of the ECHR judgment.
- Use formal, precise, and structured language typical of ECHR judgments.

Input:

External Cases:

{EXTERNAL CASES}

Previous Paragraphs:

{PREVIOUS PARAGRAPHS}

Response:

Prompt 2: Prompt for Response Generation Module