Findings of the JUST-NLP 2025 Shared Task on Summarization of Indian Court Judgments

Debtanu Datta[1], Shounak Paul[1], Kshetrimayum Boynao Singh[2], Sandeep Kumar[2], Abhinav Joshi[3], Shivani Mishra[3], Sarika Jain[4], Asif Ekbal[2], Pawan Goyal[1], Ashutosh Modi[3], Saptarshi Ghosh[1]

[1]IIT Kharagpur, India, [2]IIT Patna, India, [3]IIT Kanpur, India, [4]NIT Kurukshetra, India

Correspondence: debtanudatta04@gmail.com

Abstract

This paper presents an overview of the *Shared Task on Summarization of Indian Court Judgments* (L-SUMM), hosted by the JUST-NLP 2025 Workshop at IJCNLP-AACL 2025. This task aims to increase research interest in automatic summarization techniques for lengthy and intricate legal documents from the Indian judiciary. It particularly addresses court judgments that contain dense legal reasoning and semantic roles that must be preserved in summaries. As part of this shared task, we introduce the **In**dian **L**egal **Sum**marization (**InLSum**) dataset, comprising 1,800 Indian court judgments paired with expert-written abstractive summaries, both in English. Therefore, the task focuses on generating high-quality abstractive summaries of court judgments in English. A total of 9 teams participated in this task, exploring a diverse range of methodologies, including transformer-based models, extractive-abstractive hybrids, graph-based ranking approaches, long-context LLMs, and rhetorical-role-based techniques. This paper describes the task setup, dataset, evaluation framework, and our findings. We report the results and highlight key trends across participant approaches, including the effectiveness of hybrid pipelines and challenges in handling extreme sequence lengths.

1 Introduction

Legal case judgments are often lengthy, intricate, and densely packed with domain-specific terminology and complex judicial reasoning. As a result, legal professionals must manually review extensive case documents to identify relevant precedents, which is crucial in Common Law systems, such as in the Indian Judiciary (Bhattacharya et al., 2019). This is a process that is not only time-consuming but also cognitively demanding domain-expert knowledge. Prior research highlights that case judgments are substantially longer than documents in most other domains, and expert-written summaries are very expensive to obtain, leading to limited availability of high-quality supervised data for the legal summarization task (Shukla et al., 2022; Datta et al., 2023). Therefore, automatic summarization of case judgments has emerged as a crucial task in the Legal-NLP field aimed at reducing the burden on law practitioners and improving access to essential case information (Zhong et al., 2019; Datta et al., 2023). Recent developments in transformer-based abstractive models, including legal-domain adaptations in Large Language Models (LLMs), have further accelerated interest in building reliable, faithful, and coherent summaries of judicial rulings (Shukla et al., 2022; Sharma et al., 2023; Deroy et al., 2025).

Motivated by these challenges, the Shared Task on Legal Summarization (L-SUMM)[1] has been organized as part of the JUST-NLP 2025 Workshop, co-located with IJCNLP–AACL 2025. The primary aim of this task is to foster increased research interest in Legal-NLP, particularly in summarization methodologies for lengthy judicial texts, reflecting real-world requirements faced by legal information retrieval systems and practitioners.

In this regard, we introduce **InLSum** (**In**dian **L**egal **Sum**marization), a dataset comprising 1,800 case judgments from prominent Indian courts, along with expert-written abstractive summaries, both in English. Participants were required to produce abstractive summaries that capture the core reasoning and factual elements of these court rulings under a unified evaluation framework. Beyond benchmarking, this shared task offers valuable insights into the practical challenges of legal summarization, including handling extreme input lengths, ensuring factual consistency, and maintaining coherence across lengthy narrative structures.

A total of 9 teams participated in the L-SUMM

[1]https://exploration-lab.github.io/JUST-NLP/task/

Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 5–11

InLSum dataset	Train		Validation		Test	
	Judgment	Summary	Judgment	Summary	Judgment	Summary
Number of Samples	1200	1200	200	200	400	400
Average Number of Words	8294	606	6815	603	7920	621
Mean Compression Ratio	1:14		1:12		1:12	

Table 1: Dataset Statistics of **InLSum** dataset.

Shared Task. In this paper, we provide an overview of the task in §2, including details of the dataset and evaluation metrics. We present the approaches adopted by the participating teams in §3 and §4 presents the results and discusses key insights. Finally, §5 concludes the paper.

2 Task Description

The L-SUMM Shared Task focuses on generating concise and coherent abstractive summaries of Indian court judgments in English, addressing the lengthy and intricate nature of these documents. Participants were provided with the **InLSum** dataset, which has dedicated train, test, and validation splits for developing models and tailoring them to legal summarization. Detailed description of the dataset is provided in §2.1. Participants used the provided *train* split to train and fine-tune their language models. During the Training Phase, they were required to submit predictions on the *validation* split to evaluate model performance and perform necessary tuning. In the Testing Phase, participants submitted predictions on the held-out *test* split to assess their models on unseen data. The final evaluation was conducted on this *test* set (the test data was released only after the completion of the Training Phase). The shared task was hosted on the CodaBench[2] platform, which facilitated access to the dataset, submission management, leaderboard evaluation, and standardized comparison of participating systems.

2.1 Dataset

For this **InLSum** Shared Task, we introduce the **InLSum** (**In**dian **L**egal **Sum**marization) dataset, which consists of court judgments from prominent Indian courts paired with expert-written abstractive summaries, both in English. To facilitate model development and evaluation, the dataset is divided into 3 splits: a *train* set of 1,200 datapoints, a *validation* set of 200 datapoints, and a *test* set of 400 datapoints. Each datapoint refers to a (judgment–summary) pair. Many judgments span several thousand words, whereas the summaries are substantially shorter, requiring high compression. Detailed statistics, including average document length, average summary length, and compression ratios (the ratio between the length of the summaries to that of the corresponding judgments), are reported in Table 1.

2.2 Evaluation

System outputs in the L-SUMM Shared Task were evaluated using these 3 widely adopted automatic relevance metrics: ROUGE-2, ROUGE-L, and BLEU.

ROUGE (Lin, 2004): It stands for *Recall-Oriented Understudy for Gisting Evaluation*, a crucial metric for assessing the n-gram overlap between the model-generated summaries and the reference summaries. In this task, In this study, we employed the ROUGE-2 score for measuring the bi-gram textual overlap and ROUGE-L for measuring the longest matching sequence of words using the Longest Common Subsequence (LCS) between the model-generated and reference summaries.

BLEU (Papineni et al., 2002): It stands for *Bilingual Evaluation Understudy*. It measures overlap between model-generated and reference summaries by considering n-gram-based precision.

To ensure uniform comparability across metrics, all scores were scaled to the range $[0, 100]$ (with higher values indicating better performance). The final leaderboard ranking was determined by computing the arithmetic mean of the F1-scores of ROUGE-2, ROUGE-L, and BLEU, for each system.

3 Shared Task Submissions

The L-SUMM Shared Task attracted a total of 9 participating teams. Brief summaries of the modeling approaches taken by these teams are described in this section.

[2] https://www.codabench.org/

Juris-Summ (Sheik et al., 2025). This system adopts an adaptive pipeline built upon the *Longformer Encoder–Decoder* (*LED*)[3] model (Beltagy et al., 2020). For shorter judgments (<8,000 words), LED is applied in a single-pass setting, while longer documents are processed hierarchically: they are split into overlapping 5,000-word chunks (20% overlap) that are individually summarized, after which a meta-summarization stage fuses the chunk summaries into a coherent final output. This strategy leverages LED's long-context capabilities while mitigating boundary effects and preserving global coherence.

BLANCKED (Parada et al., 2025). This team adopts a hybrid extractive-abstractive pipeline designed for extremely long legal judgments. Their approach first segments documents into 512-token semantic chunks and applies the extractive summarization method, *PACSUM* (Zheng and Lapata, 2019), to rank and select the most salient content, yielding a 1000-token condensed extract. This extract is then passed to the proprietary LLM, *Gemini-2.5-Pro*, for zero-shot abstractive refinement, utilizing optimized prompts to enhance coherence and minimize redundancy. The method balances efficiency with coverage, with the LLM primarily enhancing fluency rather than adding new content, and achieves consistent gains over a purely extractive baseline.

TLDR-Uniandes (Chica, 2025). This team investigates a spectrum of prompting strategies with multi-agent architectures over widely popular proprietary LLM, *GPT-4.1*. They evaluate multiple prompt families, progressing from simple `TL;DR` baselines to structured few-shot instructions that enforce target length, retention of legal terms, and n-gram density. Their strongest results come from a Reward System prompt, which incorporates progressive rewards for long exact spans, bonuses for legal transition phrases (e.g., 'held that', 'dismissed the appeal'), contextual multipliers for sentence placement optimizing for ROUGE and BLEU simultaneously. Beyond prompting, they design three multi-agent pipelines: a two-stage extract–abstract workflow using verbatim extraction; a domain-aware pipeline where legal-domain classification informs the structure and content emphasis of later stages; and an expansive 20-stage sequential pipeline that processes ten legal rhetorical roles (facts, arguments, reasoning, orders, citations, etc.) with dedicated extraction and abstraction agents. A synthesis agent then fuses all partial outputs into a coherent final summary, prioritizing high-fidelity n-gram matching. This combination of reward-driven prompting and modular multi-agent flows represents one of the interesting approaches in the shared task.

BDS-Lab (Sonowal and Sadhu, 2025). This team introduces a *Structure-Aware Chunking* (SAC) pipeline that explicitly aligns the summarization process with the rhetorical structure of legal judgments. Their method segments each document into *Facts*, *Arguments & Analysis*, and *Conclusion* using either (i) a heuristic rule-based system (SAC-H) built from lexical indicators observed in the training corpus or (ii) a zero-shot LLM-based segmentation method (SAC-LLM) using *Gemini-2.5-Pro*, which identifies section boundaries through structured prompts. After segmentation, the system allocates the summarization token limits proportionally across sections, leveraging empirical distributions of section lengths in gold summaries. Each segment is then summarized independently using *Legal-Pegasus*[4] and concatenated to form the final output. Their analysis highlights a key trade-off: while section-aligned chunking improves global coherence (ROUGE-L), it can reduce local n-gram fluency (ROUGE-2), revealing structural constraints inherent in long-document summarization.

SCaLAR (D and Madasamy, 2025). This team explored three systems built on the *Legal-Pegasus* model, addressing extreme document length through hierarchical summarization. Their baseline system employs naive token-based chunking (1000 tokens with overlap) followed by recursive summarization to aggregate local summaries into a final global one. The next system improves this process through *rhetorical chunking*: a BERT-based classifier assigns one rhetorical role to each sentence, and sentences sharing the same role are grouped into coherent semantic units, which are then summarized hierarchically using a role-aware fine-tuned version of the same model. System-3 extends from System-2 by incorporating *weighted rhetorical roles*, where each role is assigned an explicit importance score derived from its prominence in reference summaries of the *train* split; these scores are inserted into the input tags during fine-tuning. Collectively, their experiments investi-

[3] https://huggingface.co/allenai/led-large-16384

[4] https://huggingface.co/nsi319/legal-pegasus

Rank	Team Name	ROUGE-2↑	ROUGE-L↑	BLEU↑	AVG↑
1	FourCorners	34.91	33.34	21.49	**29.91**
2	Juris-Summ	29.62	28.56	21.67	**26.62**
3	TLDR-Uniandes	26.88	27.38	19.49	**24.58**
4	Contextors	25.13	25.59	16.80	**22.51**
5	GenAI-Lab	25.90	24.95	13.05	**21.30**
6	SCaLAR	21.86	25.93	14.43	**20.74**
7	BLANCKED	21.05	24.35	15.12	**20.17**
8	LegalAI	20.37	22.49	13.67	**18.84**
9	BDS-Lab	16.51	22.41	05.08	**14.67**

Table 2: Final Leaderboard results of the L-SUMM Shared Task in JUST-NLP 2025. AVG denotes the mean of ROUGE-2, ROUGE-L, and BLEU scores, where higher values (↑) indicate better performance.

gate how rhetorical structure and role-level importance signals influence the quality of summaries for long legal judgments.

LegalAI (Sha et al., 2025). This team explored two approaches: (i) a hybrid extractive–abstractive pipeline combining a *BART*-based model (fine-tuned on CNN/DailyMail and further adapted on InLSum) with *TextRank*-based extractive pre-selection, and (ii) a purely abstractive summarization approach using *Indian_Legal_Pegasus*[5], a domain-adapted variant of *Legal-Pegasus* fine-tuned on the Indian legal domain. In the first approach, sentence embeddings from *all-MiniLM-L6-v2* are used to build a similarity graph, and the top 20 TextRank-selected sentences are fed to BART (within a 1024-token limit). The second approach directly fine-tunes and applies the model under similar input and output length constraints. They illustrate the differences between extractive-guided and fully abstractive strategies for long-form legal summarization.

FourCorners (Chaksangchaichot and Akarajaradwong, 2025). This team presents a multi-stage alignment pipeline tuned by Reinforcement Learning (RL) built on *Qwen3-4B-Instruct-2507*[6] that ultimately secured the *top position* on the L-SUMM Shared Task leaderboard. Their pipeline begins with data filtering based on regression between judgment and summary lengths to remove noisy pairs. The model, *Qwen3-4B-Instruct-2507*, is then adapted using a two-stage supervised finetuning (SFT) strategy: an initial high-rank LoRA phase on medium-length inputs to shape task-specific behavior, followed by a long-context finetuning phase (10k–30k tokens) with lightweight adapters. A final Reinforcement Learning with Verifiable Rewards (RLVR) step further aligns generation quality by directly optimizing BLEU and ROUGE using **G**roup **S**equence **P**olicy **O**ptimization (GSPO) (Zheng et al., 2025) or **D**ecoupled Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization (DAPO) (Yu et al., 2025) style updates, sequence-level importance sampling, and low-rank LoRA for stable optimization. Finally, they demonstrate an RL-based approach to long-form legal summarization and show that their entire pipeline remains highly compute-efficient.

Contextors (Neelamegam and Nirmala, 2025). They fine-tune multiple pretrained Seq2Seq models (*BART*, *Legal-Pegasus*, *T5*, and *LED*) over InLSumm and integrate them through an ensemble framework. Pairwise and three-way ensembles (BART-Pegasus-LED) generate candidate summaries, which are then semantically ranked using *InLegalBERT* model to select the most contextually aligned output. They further propose a Retriever-Driven framework that identifies the most semantically relevant document chunks before fine-tuning each model and again uses semantic similarity for final summary selection.

GenAI-Lab (Jadav et al., 2025). This team also investigates retrieval–driven summarization for long legal documents, comparing three retrieval strategies – *Dense Passage Retrieval* (DPR) (Karpukhin et al., 2020), *Maximum Cosine Similarity* (MCS) (Shukla et al., 2022), and *Maximum Marginal Relevance* (MMR) (Xie and Liu, 2008) – to construct semantically aligned training pairs. Judgment texts are chunked into

[5] https://huggingface.co/akhilm97/pegasus_indian_legal

[6] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

passages (e.g., 1024 tokens depending on the input context size of the model), and relevant passages are selected for each summary sentence using embedding-based similarity. Summaries are generated using both zero-shot LLMs (e.g., *Qwen-32B*, *GPT-OSS-120B*, *LlaMa-3*, and *LlaMa-4*) and fine-tuned encoder–decoder models, particularly LED and Legal-Pegasus. The fine-tuned models are trained on datasets produced by each retrieval method, enabling controlled evaluation of how retrieval quality impacts abstractive summarization performance.

4 Results and Findings

Table 2 presents the final leaderboard of the L-SUMM Shared Task. The results reveal clear performance variation across methodological families, reflecting the diversity of modeling strategies explored by participating teams.

The top-performing teams employ different strategies, including reinforcement learning (FourCorners), hierarchical long-context modeling (Juris-Summ), multi-agent prompting (TLDR-Uniandes), and retrieval-based multi-generator ensembles (Contextors and GenAI-Lab), while also emphasizing mechanisms that preserve global coherence in extremely long legal documents. The top performer, **FourCorners** (AVG: 29.91), couples supervised finetuning with GSPO/DAPO-based RL alignment, achieving the highest scores across major metrics despite using a relatively small (4B) model. Their results illustrate the effectiveness of reward optimization for long-form legal summarization. **Juris-Summ** (AVG: 26.62) and **TLDR-Uniandes** (AVG: 24.58) follow hierarchical LED summarization and multi-agent reward-driven prompting using GPT-4.1, respectively. The mid-ranking systems, including **Contextors** (AVG: 22.51) and **GenAI-Lab** (AVG: 21.30), employ retrieval-enhanced finetuning. Their moderate performance suggests that retrieval enhances factual grounding, although gains depend strongly on the choice of generator and selection heuristics. The lower-ranking systems, including **SCaLAR** (AVG: 20.74), **BLANCKED** (AVG: 20.17), **LegalAI** (AVG: 18.84), and **BDS-Lab** (AVG: 14.67), generally relied on more traditional pipelines such as extractive pre-processing or direct finetuning of legal-domain models.

Across systems, a central finding emerges: *handling document length and structure is more critical than the choice of backbone model*. Teams that explicitly address long-context challenges via hierarchical routing, retrieval augmentation, structured chunking, or reinforcement alignment, consistently outperform those relying solely on standard summarization architectures or zero-shot prompting.

5 Conclusion

The JUST-NLP 2025 L-SUMM Shared Task provides a benchmark for evaluating long-document summarization within the challenging domain of legal judgments. The task attracted 9 participating teams, who collectively explored a wide range of modeling paradigms, including long-context encoder-decoder architectures, retrieval-augmented pipelines, multi-agent prompting strategies, hybrid extractive-abstractive designs, and reinforcement learning–based alignment methods. The final leaderboard results demonstrate that the top-performing system leveraged a multi-stage SFT+RL pipeline to secure the leading position. Several other approaches have shown that retrieval quality, rhetorical role-based chunking strategies, and domain adaptation all play vital roles in shaping summary quality. Overall, the L-SUMM Shared Task highlights both the progress and the challenges that remain in long-form legal summarization. The benchmark, dataset resource, and system descriptions released through this shared task aim to support future research on long-form summarization in high-stakes domains, such as law. As long-context architectures, RL-strategies, alignment methods, and structure-aware techniques continue to improve, future iterations of this task are likely to further advance the capabilities of automated legal reasoning and summarization.

Limitations

In this shared task, our evaluation framework relies primarily on automatic metrics such as ROUGE and BLEU, which measure surface-level lexical overlap and provide only a syntactic assessment of summary quality. Semantic evaluation metrics (e.g., BERTScore) were not included in the official scoring pipeline. Also, the model-generated abstractive summaries are highly susceptible to factual inconsistencies, particularly when dealing with complex judicial reasoning and domain-specific terminology. However, our evaluation setup does not independently assess factual accuracy using factual consistency measures (e.g., SummaCONV or

other hallucination-sensitive consistency metrics). Further, a major limitation is the absence of expert-driven human evaluation. Legal summarization requires domain knowledge, and expert review by legal professionals is essential for assessing correctness, completeness, and potential misinterpretations in generated summaries. Due to the high cost and limited availability of legal experts, we were unable to conduct a human evaluation phase. Consequently, the final leaderboard does not reflect expert assessments.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval*, pages 413–428, Cham. Springer International Publishing.

Chompakorn Chaksangchaichot and Pawitsapak Akarajaradwong. 2025. A Budget Recipe for Finetuning a Long-form Legal Summarization Model. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Santiago Chica. 2025. Automatic Legal Judgment Summarization Using Large Language Models: A Case Study for the JUST-NLP 2025 Shared Task. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Arjun T D and Anand Kumar Madasamy. 2025. SCaLAR_NITK @ JUSTNLP Legal Summarization (L-SUMM) Shared Task. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Singapore. Association for Computational Linguistics.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2025. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, 33(4):1007–1050.

Nita Jadav, Ashok Urlana, and Pruthwik Mishra. 2025. NIT-Surat@L-Sum: A Semantic Retrieval-Based Framework for Summarizing Indian Judicial Documents. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pavithra Neelamegam and S Jaya Nirmala. 2025. Contextors at L-SUMM: Retriever-Driven Multi-Generator Summarization. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Erich Giusseppe Soto Parada, Carlos Manuel Muñoz Almeida, and David Cuevas Alba. 2025. Combining Extractive and Generative Methods for Legal Summarization: BLANCKED at JUST-NLP 2025. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Sayed Ayaan Ahmed Sha, Sangeetha Sivanesan, Anand Kumar Madasamy, and Navya Binu. 2025. Integrating Graph based Algorithm and Transformer Models for Abstractive Summarization. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Saloni Sharma, Surabhi Srivastava, Pradeepika Verma, Anshul Verma, and Sachchida Nand Chaurasia. 2023. A comprehensive analysis of indian legal documents summarization techniques. *SN Comput. Sci.*, 4(5).

Reshma Sheik, Noah John Puthayathu, Fathima Firose A, and Jonathan Paul. 2025. Hierarchical Long-Document Summarization using LED for Legal Judgments. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods

and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.

Himadri Sonowal and Saisab Sadhu. 2025. Structure-Aware Chunking for Abstractive Summarization of Long Legal Documents. In *Proceedings of the First Workshop on NLP for Empowering Justice (JUST-NLP)*, Mumbai, India. Association for Computational Linguistics.

Shasha Xie and Yang Liu. 2008. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4985–4988.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476.

Chujie Zheng, Shixuan Liu, Mingze Li, Xionghui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *ArXiv*, abs/2507.18071.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D. Ashley, and Matthias Grabmair. 2019. Automatic summarization of legal decisions using iterative masking of predictive sentences. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, page 163–172, New York, NY, USA. Association for Computing Machinery.