

Introducing Spotlight: A Novel Approach for Generating Captivating Key Information from Documents

^{1,*}Ankan Mullick

^{1,*}Sombit Bose

¹Rounak Saha

²Ayan Kumar Bhowmick

²Aditya Vempaty

²Prasenjit Dey

²Ravi Kokku

¹Pawan Goyal

¹Niloy Ganguly

{ankanm, sbcs.sombit.24, runk20}@kgpian.iitkgp.ac.in

{pawang, niloy}@cse.iitkgp.ac.in, {ayan, aditya, prasenjit, ravi}@emergence.ai

¹IIT Kharagpur, India. ²Emergence AI

Abstract

Analyzing and processing vast amounts of textual data presents significant challenges in efficiently extracting key information. In this paper, we introduce ‘Spotlight’, a novel paradigm for information extraction that produces concise, engaging narratives by highlighting the most compelling aspects of a document. Unlike highlights (fragmented key points) and traditional summaries, which prioritize comprehensive coverage, spotlights selectively emphasize intriguing content to foster deeper reader engagement with the source material. We formally differentiate spotlights from related constructs and support our analysis with a detailed benchmarking study using new datasets curated for this work. To generate high-quality spotlights, we propose a two-stage approach: fine-tuning a large language model on our benchmark data, followed by alignment via Direct Preference Optimization (DPO). Our comprehensive evaluation demonstrates that the resulting model not only identifies key elements with precision but also enhances readability and boosts the engagement value of the original document. Datasets and code are available at <https://github.com/ankan2/Spotlight-EMNLP2025>.

1 Introduction

In the field of document processing, information condensation techniques such as summarization aim to extract and present the essential content of lengthy texts in a more accessible and concise format. A summary is typically a structured and comprehensive abstraction that encapsulates all major points, providing readers with a faithful overview of the document’s primary content. In contrast, this paper introduces spotlight—a novel, self-contained narrative designed to highlight the most engaging and insightful elements of a document. Unlike

traditional summaries that prioritize full coverage, a spotlight constructs a coherent and compelling mini-narrative that remains faithful to the original text, with the goal of capturing the reader’s interest and encouraging further exploration of the complete document.

Spotlights differ from traditional forms of document processing such as highlights, headlines, and teasers, which serve specific roles in information extraction. Highlights consist of short, disconnected snippets that reflect key points (Arumae et al., 2019; Cho et al., 2020; Wei and Gao, 2018; Woodsend and Lapata, 2010), but often lack coherence and narrative flow. Headlines are single-sentence summaries that capture the central theme (Colmenares et al., 2015; Ao et al., 2021; Gavrilov et al., 2019), while teasers are crafted to entice readers by offering intriguing yet vague glimpses of the content (Karn et al., 2019; Woźny and Lango, 2023; Chakraborty et al., 2016; Deng et al., 2024), without revealing substantive information.

This paper introduces the concept of spotlights as an innovative step forward from traditional summaries and highlights. A spotlight is a concise, self-contained narrative, typically prepared by the document’s author, that highlights its most significant and engaging aspects. Unlike traditional formats, spotlights offer the narrative depth of summaries while uniquely aiming to generate interest in the full text.* They strike a careful balance between

*Note that even the summarization can be target-specific, some examples include query focused summarization (Vig et al., 2022; Baumel et al., 2016; Liu et al., 2024), layman summarization (Goldsack et al., 2023; Salaün et al., 2022; Luo et al., 2022), personal-based summarization (Mullick et al., 2024c), goal oriented (Goldstein et al., 2007; Ham et al., 2020; Jin et al., 2024), aspect-based summarization (Mullick et al., 2024d) etc. summaries with controllable readability (He et al., 2020; Zhang et al., 2023; Ribeiro et al., 2023), summarizing documents with respect to key elements (Ryu et al., 2024; Wang et al., 2023). However, the objective of these methods is generally to achieve maximum coverage and deliver a quick, faithful read—potentially sparing the reader from having to engage with the full main text, which appears to be in contrast

*Equal Contribution.

	Info-Dist	Compactness	Readability	Extractive	Informative	Interest-Generation	Faithful	Length	Mini-Story
Summary†	Uniform	High	Medium	Low	High	Low	High	Long	Yes
Teaser	Non-uniform	Low	High	Low	Low	High	Low	Short	No
Headline§	Non-uniform	Low	High	High	Low	Medium	Medium	Short	No
Highlight	Non-uniform	Low	High	High	High	Medium	High	Medium	No
Spotlight	Skewed	Low	High	High	High	High	High	Long	Yes

Table 1: Qualitative Analysis of different metrics to evaluate and characterize various methods (Summary, Teaser, Headline, Highlight, and Spotlight) for deriving information from a source document.

Information Distribution (Info-Dist): Uniformity with which information is extracted from different sections of the document; **Compactness:** Details in Section 3.1B; **Readability:** Details in Section 3.1C;

Extractive: The degree to which the generated content retains phrases or segments from the original text, rather than paraphrasing or reinterpreting them.;

Informative: Extent to which the generated content contains information relevant to the original text;

Interest-Generation: Capturing the reader’s attention and/or stimulating curiosity to encourage engagement with the source document;

Faithful: Degree to which the content is exclusively derived from the source document;

Length: Total number of words, sentences, or characters it contains;

Mini-Story: A brief compact narrative that encapsulates the main plot or theme of an article.

†: We consider the case of abstractive summary. §: For non-clickbaitish Headlines.

readability, engagement, and fidelity, functioning as standalone narratives. Much like an author’s spotlight talk at a conference, they are concise yet compelling, designed to draw readers in and encourage deeper exploration. While these five condensation techniques may exist along a continuum, they possess distinct characteristics that justify separate classification. Table 1 outlines their comparative features, such as source fidelity, level of intrigue, abstractiveness, conciseness, narrative coherence, readability, and the distribution of content coverage.

While extensive research exists on summarization and highlight extraction (see Appendix E), spotlight generation remains unexplored. In this work, we introduce the concept, quantitatively differentiate spotlights from summaries, and propose effective generation methods. Although no dedicated spotlight datasets currently exist, we identify and adapt existing sources—such as news articles, CSPubSum, Wikipedia, and conference presentations—to align with spotlight characteristics. These are compared with corresponding similar concepts (summary, teaser, headline, highlight based on the framework defined through Table 1.

We further develop a cost-effective, domain-specific spotlight generation method using a fine-tuned Large Language Model with Direct Preference Optimization (DPO). Our empirical evaluation, which includes benchmark datasets and AI-based critique mechanisms, shows that our ap-

to the objective of spotlight. Moreover, the concept of spotlights can also be extended to cater to these objectives, e.g., query-focused spotlight, persona-based spotlight etc.

proach outperforms both prompt-based techniques and larger models such as those from the GPT family across several key metrics.

The major contributions of this paper are as follows: (a) We introduce the concept of spotlight generation and formally define the problem; (b) We compile four datasets and benchmark spotlight vis-a-vis summary. (c) We propose a preference-based model for spotlight generation and perform extensive experiments to establish its effectiveness.

2 Dataset

We curate four datasets — News Category Dataset, CSPubSum, Wikipedia, and Research Presentation. These datasets were selected because each includes an additional condensed abstraction that highlights the most compelling aspects of the document, aligning well with the objectives of spotlight generation. Additionally, to highlight the distinction between a spotlight and a summary, we also created summaries for the corresponding documents.

2.1 Spotlight Datasets

News Category Dataset (Misra, 2022): This dataset has fine-grained category information for large number of news articles (published between 2012 and 2022) and we have randomly selected 14,080 news headlines for our experiment. We consider the headline along with a short description provided as a spotlight of the news article as shown in Fig. 3 (Appendix). The rationale is that headlines and the associated short descriptions provide a concise, informative structure in an engaging manner, similar to the desired spotlights.

CSPubSum (Collins et al., 2017): This dataset comprises the abstract, introduction, and conclusion sections of various research articles, accompanied by author-curated key points that highlight the core contributions of each paper. These author-curated key points are used as spotlights in our framework.

Wikipedia: We curate Wikipedia articles across four popular domains - Healthcare, Music, Education, and Life & Career—organized as detailed in Table 9 (Appendix). Each article is divided into category-specific sections, with the introductory paragraphs emphasizing the most compelling aspects of each section. To assign a portion of this paragraph to a specific category, we extract the sentence(s) most similar to the category’s focus. Thus we have each section as a separate document and the corresponding extracted sentences as its spotlight. The total number of documents obtained is 17,323.

Research Presentation: We curate this dataset by collecting 1,230 audio transcripts of NeurIPS 2022 * and 2023 * spotlight Poster presentations. These audio transcripts are derived from 5-minute self-narrative videos, which focus on presenting the most intriguing and important aspects of the research papers. This concise format is meant to capture the attention of the audience and encourage them to explore the full paper. Since these audio transcripts are generated from automated transcription services, they often contain errors that could hinder readability and interpretation. Therefore, we apply transcript cleaning techniques like contraction expansion, punctuation removal, etc. to improve the quality and utility of these transcripts.

2.2 Summary Generation

We generate summaries manually by annotators as well as by using LLMs.

Human Generated Summaries: For deeper analysis, first, we manually create a smaller set of summaries for four datasets. A total of 60 articles were randomly selected from each dataset and assigned to four annotators (one annotator per dataset). Each annotator independently generated summaries for the assigned articles, adhering to the average length of the spotlight sections of the original articles. Detailed steps are in Appendix G.2.

* <https://nips.cc/virtual/2022/events/Spotlight>

* <https://nips.cc/virtual/2023/events/spotlight-posters-2023>

LLM Generated Summaries: Since manual summary generation is time-consuming and costly, to scale up the summary generation task, we utilize GPT-4 to automate the creation of summaries for all articles of entire datasets involved in the experiments. The GPT-4-generated summaries had an average length of approximately 15% of the original articles, closely aligning with the average spotlight length of 13.76% of the original articles. To ensure the quality of these summaries, we implemented a three-step automated validation process (Table 2) followed by manual verification. Approximately 3.64% of the document-summary pairs were filtered out based on Table 2.

Additionally, we evaluate the similarity between the manually created summaries and those generated by GPT-4 for the 60 articles that were curated through both methods. The results demonstrate a high degree of alignment, with BERT-F1 scores of 81.5%, and GPT-4-based similarity scores of 89.8%. Thus, it shows that GPT-4 generated summaries are close to human-curated summaries and can be used as a replacement.

Filtering Steps with Criteria (removed)	%
Step 1: Too many special characters and other string (HTML tags and #)	1.45
Step 2: Incomplete Summary (By checking punctuations)	1.07
Step 3: If the summary contains Terms or numbers not present in the document	1.12
Overall summaries filtered out	3.64

Table 2: Step-by-Step Data Filtering

3 Spotlight: Characterization

As can be seen in Table 1, the concept of spotlight differs from the summary on the first six features, while it is similar on the last three features. We thus study the differences between the first six and then characterize the last three features.

3.1 Spotlight-Summary: Distinct Features

We do a comparative study on four datasets for 60 annotated samples (each instance has a document, corresponding spotlight and manual summary).

(A) Information Distribution over Quartiles of the document: Information distribution of a condensation (summary or spotlight) refers to the distribution of key points and essential details of the original document. First, documents are divided into four quadrants and we measure the distribution of spotlight or summary information over the four quadrants. We report the average entropy of this distribution and the fraction of documents where 50% of key points covered by the corresponding spotlight/summary are present in a single quadrant

Dataset	A. Distribution		B. Compactness↓	C. Readability (%)			D. Informativeness	E. Degree of Extraction	F. Interest-Generation (%) ↑	G. Factuality(%)	H. Length
	Entropy↓	Skew50↑		Flesch-Kincaid↓	Flesch Reading↑	Gunning↓					
News	Spot	0.69	0.61	0.52	10.57	15.46	16.81	0.061	0.021	80	91.4
	Summ	0.96	0.02	1.53	16.89	9.45	20.10	0.143	0.011	24	87.2
CSPubSum	Spot	0.62	0.47	0.62	16.28	9.67	19.73	0.042	0.109	86	85.3
	Summ	0.96	0.01	1.28	20.72	6.42	23.55	0.112	0.100	30	87.1
Wikipedia	Spot	0.76	0.29	0.82	12.62	16.26	16.37	0.030	0.032	86	88.2
	Summ	0.95	0.05	1.48	16.24	9.35	19.19	0.239	0.025	26	84.3
Research	Spot	0.65	0.36	0.28	11.84	44.10	11.85	0.069	0.052	82	75.3
Presentation	Summ	0.92	0.04	0.69	16.03	23.83	17.21	0.315	0.026	24	79.3
											675.9

Table 3: Feature level analysis of Spotlight (Spot) vs. Summary (Summ), Entropy, Compactness, Flesch Kincaid and Gunning score lower values signify better, whereas for Skew50, Flesch Reading score and Interest-Generation higher values signify a better result as discussed in Section 3, Additional Results in Appendix Table 11 and 12

of the document (denote this fraction by Skew50) as shown in Table 3. Our results reveal that spotlights exhibit significantly lower entropy and higher skew compared to summaries, as they tend to focus primarily on a specific (and presumably intriguing) portion of the document.

(B) Information Compactness: Information compactness refers to the degree to which a condensation distills essential details of the original content. We use the average IDF* score to measure compactness. We calculate the IDF scores for each word of the condensation and then do the average. From Table 3 it is clearly seen that the average IDF value is always higher for summary than spotlight indicating more information overload (compactness) in summary than spotlight.

(C) Text Readability: Text readability is best defined as the ease with which a text can be read and understood in terms of the linguistic features found within a text. We utilize six different readability scores, detailed in Table 10 (Appendix).

Table 3 shows results of three different readable metrics (due to space limitation, all other details are in the Appendix B.1) which signifies spotlight has better readability than summary. The superior performance of spotlights can be linked to our previous findings. Spotlights convey less information per sentence, which reduces cognitive load and enhances readability. In contrast, summaries tend to be more information-dense, potentially leading to higher cognitive effort for the reader. Next, we examine whether the higher compactness arises from summaries packing more information from the source document.

(D) Informativeness: To quantify the information retained from the source document in a condensation, we extract key-facts from the original documents using key-fact alignment (Song et al., 2024) approach and compare them to those present in spotlights and summaries using exact matching as

shown in Table 3 column D. It is seen that the key-facts retained from the document in summary is more than spotlight, which indicates a higher informativeness in summary.

(E) Degree of Extraction: To quantify the degree of text reuse in condensations (summaries vs. spotlights), we measure the fraction of sentences lifted near-verbatim from the source document. Specifically, we define a sentence y in the condensation as an almost-lift of a document sentence x if their ROUGE-F1 score exceeds 0.8 following (Bhandari et al., 2020). Our analysis reveals that spotlights exhibit significantly higher extractiveness than summaries, with a larger fraction of sentences qualifying as almost-lifts (see Table 3, Column E). This trend holds consistently across ROUGE metrics (R1, R2, RL), as detailed in Appendix Table 12. We hypothesize that the increased extractiveness of spotlights contributes to their enhanced readability, as they preserve the original document’s phrasing for key compelling content.

(F) Interest Generation: We conducted a human annotation study focused on interest generation. For each dataset, 50 documents with manually annotated summaries were selected, and two independent annotators were asked to indicate whether the summary or the spotlight was more likely to generate interest for deeper exploration among readers. Annotators also had the option to state that both formats generate similar levels of interest. The overall inter-annotator agreement was 0.87. As shown in Column F of Table 3, annotators overwhelmingly rated spotlights as generating more interest than summaries. We find that the score differences (spotlight vs summary) are statistically significant ($p < 0.05$) as per McNemar’s Test.

The difference in reader engagement between a spotlight and a traditional summary is illustrated in Figure 1, which presents both generated from the same document describing a polygon approximation algorithm. The summary provides a concise technical account, focusing on adaptive ver-

*IDF stands for Inverse Document Frequency. IDF is a standard metric to measure the informativeness of a word.

Summary	Spotlight	Article
The paper introduces a non-parametric polygonal approximation algorithm that enables <u>vertex relocation beyond the contour</u> , optimizing digital curve representation. By <u>adaptively estimating point neighborhoods</u> , the method <u>minimizes integral squared error</u> (ISE) while maintaining a compact dominant point set. Comparative experiments on multiple datasets demonstrate superior shape fidelity and compression efficiency, positioning this approach as an advancement in contour-based shape analysis. The proposed framework generalizes well across varying curve complexities, making it robust for real-world applications.	<i>This paper reimagines polygonal approximation of digital curves by <u>liberating vertices from strict adherence to contour points</u>. Through <u>adaptive local neighborhoods</u> and <u>strategic vertex relocation</u>, it significantly reduces approximation error without <u>increasing number of vertices, yielding a compact yet precise representation</u>. This non-parametric method truly stands out for its <u>remarkably balanced efficiency and accuracy</u> in capturing complex shapes with minimal data.</i>	Polygonal approximation is an effective way for capturing the outline of digital planar shapes, where the outline of an image is approximated by a sequence of line segments. Such approximations can be utilized for shape modeling and recognition. The main motivation for such an approximation based modeling lies in the ability of the approximation to cope with the noise on the shape contour. At the same time, a polygonal approximation is a more compact representation of the shape contour. For this reason, a lot of research has been carried out (for decades) in search of effective polygonal approximation algorithms.....

Figure 1: Comparison between a summary and a spotlight generated from the same document. The *summary* presents a concise, technical overview of the algorithm, detailing how adaptive vertex relocation minimizes approximation error with rigorous experimental validation. The *spotlight* describes a captivating mini-story, emphasizing the innovative breakthrough of liberating vertices from the contour to create elegant, engaging shape approximations. It minimizes emphasis on technical methodology and reduces the use of complex terminology, while effectively conveying the significance of the work in a way that encourages readers to explore the main article. Each phrase introduces a novel concept or method, prompting readers to wonder how exactly it works and why it improves results, thereby triggering a sense of deeper exploration by raising implicit questions in the reader’s mind.

The “interest generation phrases” are highlighted in **bold** and underlined).

tex relocation and its effectiveness in minimizing approximation error, supported by experimental validation. In contrast, the spotlight highlights the broader implications, potential impact, and limitations of the work, rather than delving into technical details. This framing encourages deeper reader engagement by implicitly prompting questions and curiosity. Additional details are available in the Appendix B.3.

3.2 Spotlight-Summary: Similar Features

Apart from the aforementioned distinctions, a spotlight shares similarities with a summary in terms of faithfulness to the source document, length, and its qualification as a mini-story. This is detailed next. **(G) Factuality:** To quantitatively evaluate the faithfulness of condensations to their source documents, we employ FactScore (Min et al., 2023), a metric designed to evaluate factual consistency by extracting and aligning key information units. Using an information extraction approach, we extract structured factual triples from both the source document and its corresponding condensation. We then leverage a Natural Language Inference (NLI) model to determine whether each extracted fact from the condensation is entailed by the document. The FactScore is computed as the proportion of facts in the condensation that are supported by the source document. Our results in Table 3 indicate that both summaries and spotlights achieve high FactScores,

demonstrating strong factual alignment with the source content.

(H) Length: We checked the length (number of words) of both summary and spotlights and found that roughly they are of similar length.

(I) Mini-story: A mini-story is a short, focused narrative that conveys the core plot or theme of an article in a clear and engaging way. It summarizes key ideas or events concisely while preserving coherence and capturing the reader’s interest. Spotlight and summary contents for 25 articles of each dataset have been provided to two annotators to mark one of the following three - (i) both spotlight and summary contents have mini-story; (ii) only spotlight has mini-story; (iii) only summary has mini-story. Inter-Annotator agreement is 81.2%. Any conflict has been resolved by another annotator. We find that for all the datasets at least 80% of respondents have certified both the condensations as mini-story (spotlight has a slight higher rating). The detail is presented in Table 18 (Appendix).

3.3 Spotlight-Highlight: Distinct Features

To explore the distinctness between spotlight and highlights, we have performed a similar mini-story study to Section 3.2. Details in Appendix Section G.3. In our analysis of distinct features (as described in Section 3.2), we observed that spotlights consistently outperform highlights in conveying a mini-story narrative. Annotator evaluations in-

Dataset	Spotlight(%)	Highlights(%)
News	80	20
CSPubSum	72	28
Wikipedia	72	28
Research Presentation	84	16

Table 4: Mini-story validation for spotlight vs. highlight analysis across all datasets

dicate that spotlights were rated as having a more engaging narrative as shown in Table 4. While both spotlights and highlights aim to distill key information from source documents, spotlights uniquely present a self-contained narrative that not only summarizes but also sparks curiosity and encourages further exploration. This narrative element is less pronounced in traditional highlights, which tend to list isolated key points rather than weaving them into a cohesive mini-story.

4 Spotlight Generation Process

For the task of spotlight generation, we propose a 2-step solution - at the first step, we fine-tune an open-sourced LLM; in the next step, we apply direct preference optimization (DPO) (Rafailov et al., 2024) framework on top of that fine-tuned LLM to generate spotlights.

In the DPO approach, the model differentiates spotlights from summaries, and learns to prioritize spotlight generation. As discussed in previous sections, while summaries and spotlights share similarities in terms of length, faithfulness, and their mini-story nature, they differ in key aspects such as information distribution, compactness, and readability. By emphasizing these distinctions, the model enhances the uniqueness of spotlights, ultimately leading to higher-quality outputs.

Dataset Preparation: For each document, we consider the ground truth spotlight as an upvoted example (indicating a preferred completion), while the corresponding summary as a downvoted example (denoting undesirable). Thus each document from the original dataset contributes one upvoted and one downvoted example.

Experiment Steps: We randomly divide each dataset into two equal parts for a two-step process. In the first step, one half is used for Supervised Fine-Tuning (SFT) on a base model (e.g., Llama2-7b-hf, Mistral-7B-v0.1, Llama3-8b) using ground truth spotlights. In the second step, the remaining half is utilized for Direct Preference Optimization (DPO) on the fine-tuned model obtained from the first step. In this stage, both spotlight and summary information are leveraged to highlight the

distinctions necessary for the DPO process.

Direct Preference Optimization: For each example in the second step, we consider the ground truth spotlight as the desired (preferred/chosen/winning) completion and the generic summary as the undesirable (rejected/losing) completion.

Loss Function: The DPO loss function is:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = - \mathbf{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{ref}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{ref}(y_l | x)} \right) \right]$$

where π_θ is the language model we are optimizing, π_{ref} is the reference model (the fine-tuned model obtained after first step) with respect to which we are increasing the likelihood of preferred responses and decreasing the likelihood of nonpreferred responses. Samples in the preference dataset \mathcal{D} look like (x, y_w, y_l) where x is the prompt (instruction with document in our case), y_w is the desired response (spotlight in our case) and y_l is the undesired response (summary in our case). β is a tunable hyperparameter.

Open Source LLMs: We use various small-size large language models (LLMs) - Mistral-7b (Jiang et al., 2023) and Llama* based - Llama2-7b and Llama3-8b models. We experiment with Mistral-7b (abbreviated as Mst-7b-FT), Llama2-7b and Llama3-8b fine tuning (abbreviated as L2-7b-FT and L3-8b-FT respectively), general instruction tuned vanilla models (Mst-7b-VA, L2-7b-VA and L3-8b-VA respectively) and corresponding DPO models. We report the best performing Llama3-8b LLM-based outcomes in Table 6 and other details in the Appendix Tables 21, 23, 25 and 27.

5 Experimental Setup

For spotlight generation task, we experiment with several methods and analyze the generated spotlights based on various evaluation frameworks.

5.1 Baseline Approaches

We consider two different approaches for baseline.

(A) Traditional Models: We employ Longformer (Beltagy et al., 2020) (LF), T5-3b fine-tuned (Raffel et al., 2020) (T5-FT), BART-large (Lewis et al., 2019), instruction-tuned Pegasus (Zhang et al., 2020), Falcon 7b-instruction tuned model (Penedo et al., 2023) and state-of-the-art extreme summarization approach (TLDR) (Cachola et al., 2020) as baselines.

*<https://ai.meta.com/llama/>

(B) Large-Size LLMs: We apply example-based prompting (1-shot) on large-sized LLMs like ChatGPT*, GPT-4* and Gemini-Pro* to check their capability to generate good quality spotlights. Due to space shortage, ChatGPT, and Gemini-Pro evaluations are in Appendix Tables 21, 23, 25 and 27.

5.2 Evaluation

We assess the quality of the generated spotlights using the metrics defined in Table 3. Besides this we measure its similarity with the reference spotlights by the following three set of evaluation metrics.

(A) Traditional: We check the competence of different models with traditional evaluations metrics like (i) Rouge1 (R1), Rouge2 (R2), (Lin, 2004) and (ii) BERT-F1 (BeF1) (Zhang et al., 2019).

(B) UniEval (Zhong et al., 2022): UniEval correlates much better with human judgments in NLG tasks. We use the ‘Relevance’ of UniEval scores to identify the quality of spotlight generation.

(C) GPT4 Critique (Mullick et al., 2024c): GPT-4 is asked to rate the spotlight – whether it is good or bad with respect to the ground truth spotlight and whether it generates interest to read the source document.

5.3 Experimental Settings

We use 80GB A100 GPU, 210MHz clock cycle with Huggingface TRL and PEFT libraries built on top of Huggingface Transformers to conduct the experiments of fine-tuning and alignment of pretrained models and for their evaluation purpose. Additionally, we use the NLTK, spaCy, and NumPy packages for further analysis and ablations. Further details are in Appendix F.

6 Experimental Results and Discussion

In this section, we analyze the results based on values of different evaluation strategies - traditional metrics, UniEval, GPT-4 critiquing and spotlight characteristics features to understand how different models perform and gain insights into the effectiveness of LLMs for spotlight generation tasks.

(A) Performance of Traditional Baseline Approaches: We explore different traditional approaches (LongFormer, T5-Fine-tune, BART-large, Pegasus, Falcon, TLDR) for spotlight generation

Model	Traditional Metrics			UniEval	GPT-4
	R1	R2	BeF1	Relevance	Critique
L2-7b-FT	28.1	8.6	81.3	66.6	77.8
Mst-7b-FT	25.7	6.9	77.4	65.1	77.0
L3-8b-FT	31.9	9.7	81.8	67.1	79.5
LF	16.0	4.8	67.4	55.2	68.7
T5-FT	20.0	5.0	71.0	63.2	75.9
BART	19.1	3.3	71.9	64.7	70.2
Pegasus	17.8	4.1	70.8	62.5	60.9
Falcon	17.0	3.0	74.0	55.2	38.4
TLDR	19.5	5.6	77.6	62.9	72.3

Table 5: Traditional, UniEval and GPT-4 Critique evaluations on CSPubSum Dataset - comparison of baselines with SFT, Additional Results in Appendix Table 20, 22, 24 and 26

and results for CSPubSum are shown in Table 5. Table 5 shows that traditional summarization models such as LongFormer, T5, BART, Pegasus, and Falcon struggle to generate effective spotlights, with lower ROUGE, BERT-F1, and UniEval scores. In contrast, even simple supervised fine-tuning of LLMs (Llama2, Mistral, Llama3) yields large improvements, confirming that LLMs can adapt well to the spotlight objective. This gap underscores that generating spotlights demands richer semantic modeling than what traditional summarization frameworks provide. Detailed results for the datasets are in Appendix Table - 20 (for News), 22 (for CSPubSum), 24 (for Wikipedia) and 26 (for Research Presentation) It is seen that even simple SFT-based approaches on LLMs perform much better than these traditional approaches across all metrics. This performance gap may be due to the advanced architectures and targeted optimizations of LLM models which are fine-tuned specifically for task performance and user preferences, whereas the older models do not benefit from the same level of refinement and cutting-edge techniques, leading to their comparatively lower performance.

(B) Performance of DPO vis-a-vis other LLM-based approaches: The results presented in Table 6, demonstrate the effectiveness of the DPO approach for spotlight generation across various evaluation criteria for Llama3-8b large language model (LLM) for different datasets where L3-8b-VA, L3-8b-FT and DPO (L3-8b) indicate Llama3-8b vanilla, supervised fine-tuning and DPO approach respectively. It is seen that

- DPO consistently outperforms vanilla models, supervised fine-tuning (SFT), and GPT-4 (1-shot) (in most of the cases) proving its superiority in spotlight generation.
- SFT approaches is a close second, however,

* <https://chat.openai.com/>

* <https://openai.com/gpt-4>

* <https://gemini.google.com/>

Model	Traditional Metrics			UniEval	GPT-4
	R1	R2	BeF1	Relevance	Critique
News					
L3-8b-VA	15.9	6.8	73.8	70.6	78.0
GPT-4 (1-shot)	31.3	12.1	79.5	83.1	90.1
L3-8b-FT	36.5	15.8	81.5	82.2	89.4
DPO (L3-8b)	36.8	16.0	83.6	84.7	91.3
CSPubSum					
L3-8b-VA	15.5	3.8	47.2	60.1	73.6
GPT-4 (1-shot)	28.7	7.4	80.8	66.8	82.1
L3-8b-FT	31.9	9.7	81.8	67.1	79.5
DPO (L3-8b)	32.3	9.9	82.3	67.8	89.0
WikiPedia					
L3-8b-VA	20.3	6.5	71.7	65.7	52.1
GPT-4 (1-shot)	34.6	13.8	80.4	70.2	73.5
L3-8b-FT	42.4	26.0	81.3	69.2	64.8
DPO (L3-8b)	42.5	26.3	83.4	67.3	67.3
Research Presentation					
L3-8b-VA	15.6	3.2	48.5	70.2	73.6
GPT-4 (1-shot)	32.3	7.1	77.2	75.2	81.6
L3-8b-FT	28.3	6.9	76.5	76.7	81.2
DPO (L3-8b)	38.2	8.9	79.6	76.8	88.8

Table 6: Traditional, UniEval and GPT-4 Critique evaluations for best performing Llama3-8b and GPT-4 approach in % (best outcome in **Bold**) for each dataset, Additional evaluation details including other LLMs are in Appendix Table 21, 23, 25 and 27

DPO consistently improves over SFT methods indicating its utility and the potency of SFT when the reference negative examples (here summaries) are not available.

- Vanilla L3-8b-VA performs the worst, struggling in both traditional and advanced evaluations, highlighting the importance of fine-tuning or DPO across all datasets.
- for *News*, *CSPubSum* and *Research Presentation* datasets, Llama3-8b fine-tuning and DPO, both perform better than GPT-4 (1-shot) for traditional metrics whereas for *Research Presentation*, GPT-4 (1-shot) performs better than the fine-tuning model but worse than DPO.
- GPT-4 (1-shot) is strong in UniEval Relevance

Model	Distribution		Compactness↓	Readability(%)		
	Entropy↓	Skew50↑		Flesch Kincaid↓	Flesch Reading↑	Gunning↓
News						
L3-8b-FT	0.65	0.51	0.62	9.09	62.39	10.67
DPO (L3-8b)	0.60	0.55	0.57	9.01	64.83	10.63
CSPubSum						
L3-8b-FT	0.60	0.53	0.65	9.38	59.36	10.72
DPO (L3-8b)	0.57	0.50	0.61	9.30	62.96	10.14
WikiPedia						
L3-8b-FT	0.63	0.55	0.71	7.41	66.31	8.67
DPO (L3-8b)	0.57	0.59	0.65	6.80	67.65	8.01
Research Presentation						
L3-8b-FT	0.49	0.40	0.37	12.0	44.74	12.12
DPO (L3-8b)	0.45	0.47	0.30	9.62	55.85	9.57

Table 7: Spotlight Characteristics Evaluation on Model Generated spotlight - Entropy, Compactness, Flesch Kincaid and Gunning score lower values signify better result, but for Skew50 and Flesch Reading score higher values signify a better result as discussed in Section 3

and Critique (it is best among all in *Wikipedia*) but falls behind in ROUGE and BERTScore, suggesting it produces more readable but not necessarily more extractive spotlights.

- For Wikipedia data, all models show higher Rouge scores with lower UniEval relevance/GPT-4 Critique suggests that while the spotlights have significant lexical overlap with the reference, they may lack deeper semantic coherence.
- We explore different DPO configurations for utilization of data and fine-tuning approach (details in Appendix I.4) and the best result is obtained in our 2-step approach when half of the data is used for SFT followed by DPO for the remaining half of the data.
- Notably, the Llama3-8b based DPO model achieves statistically significant improvements ($p < 0.05$) over other model variations.
- To validate the GPT-4 critique scores against robust human judgment, we conducted an additional human annotation study similar to (Mullick et al., 2024c) to evaluate the alignment between human judgment and automated assessments for 20 documents for each dataset, details in Appendix I.2. We obtain an overall average Point-biserial correlation coefficient of 0.751 between the GPT-4 critique scores and human judgment.

Across the datasets, DPO applied on Llama3-8b consistently outperforms both vanilla and fine-tuned versions. While GPT-4 (1-shot) is strong in certain cases, particularly in UniEval, the DPO-trained Llama3-8b achieves the best balance across all metrics. This demonstrates that direct preference optimization is especially effective for spotlight generation.

(C) DPO Conformance to Spotlight Features: We do an extensive evaluation of the model-generated spotlights of DPO pipeline with the fine-tuned version across various spotlight characteristics, as detailed in Table 7 which shows that (i) DPO (L3-8b) consistently improves readability across all datasets, producing spotlights that are easier to understand than fine-tune approach; (ii) Entropy reductions and increase in Skew50 across all datasets imply DPO maintains skewness in term distribution than fine-tune. (iii) Compactness decreases in DPO, indicating less information over-

Model	Traditional Metrics			UniEval	GPT-4 Critique
	R1	R2	BeF1	Relevance	Critique
News					
L3-8b-VA	17.2	8.3	75.8	72.5	65.0
L3-8b-FT Spot [†]	24.6	12.5	75.0	76.3	61.3
DPO (L3-8b) Spot [†]	23.6	10.6	70.7	74.9	60.8
L3-8b-FT Sum [§]	57.3	36.0	85.2	81.9	87.3
CSPubSum					
L3-8b-VA	13.5	2.5	60.2	61.8	59.6
L3-8b-FT Spot [†]	12.5	1.4	61.0	63.9	65.3
DPO (L3-8b) Spot [†]	10.1	1.8	59.8	62.5	60.8
L3-8b-FT Sum [§]	25.0	7.5	70.1	66.5	73.0
WikiMedia					
L3-8b-VA	19.2	7.8	65.7	66.1	55.1
L3-8b-FT Spot [†]	28.3	13.8	64.7	72.5	48.1
DPO (L3-8b) Spot [†]	25.9	12.5	60.7	70.3	45.3
L3-8b-FT Sum [§]	32.3	18.2	77.3	78.6	70.2
Research Presentation					
L3-8b-VA	14.8	3.8	58.5	69.5	63.6
L3-8b-FT Spot [†]	32.5	12.1	70.1	75.3	68.9
DPO (L3-8b) Spot [†]	29.3	7.1	68.8	71.8	61.9
L3-8b-FT Sum [§]	38.3	16.2	74.7	77.4	75.3

Table 8: Traditional and GPT-4 Critique Evaluations for each dataset, [†]: model trained on spotlights but evaluated on gold standard summary (GPT-4 generated), [§]: model trained on summary and evaluated on gold standard summary

load and more readability. In conclusion, DPO produces more structured, readable, and well-balanced spotlights compared to fine-tuned approach. These observations confirm that DPO yields spotlights that are not only more structured and readable but also more faithful to the intended characteristics of spotlight generation, in contrast to the less targeted outputs of a purely fine-tuned approach.

(D) Divergence of DPO-generated Spotlights from the gold standard Summary: As outlined in Section 4, our spotlight generation process consists of two steps. First, we fine-tune the base model (FT), and then we apply the DPO method to the fine-tuned model to align the output closer to a spotlight and further from a summary. While DPO produces superior spotlights (Table 6), a question remains: does the spotlight diverge from the summary characteristics, as per the DPO objective? To answer this, we compare model-generated spotlights with gold-standard summaries (GPT-4 generated) for three approaches: vanilla (L3-8b-VA), fine-tuning (L3-8b-FT Spot), and DPO (DPO (L3-8b) Spot), as presented in Table 8 for the entire datasets. Our findings indicate that the spotlights generated using SFT are closer to the reference summaries than those produced by DPO, confirming the deviation. Table 8 examines whether spotlight-trained models drift toward summary characteristics. The results indicate that spotlights produced by SFT are closer to gold summaries, whereas DPO outputs diverge more strongly, align-

ing with the intended spotlight objective. At the same time, when models are trained explicitly on summaries, they achieve higher traditional metrics but lose the engagement qualities of spotlights. This confirms that while spotlights and summaries overlap, optimizing for one does not automatically guarantee the qualities of the other. Nonetheless, both the fine-tuned and DPO outputs outperform the vanilla model, which prompts us to investigate whether the spotlight generation process inadvertently produces quality summaries.

To explore this further, we fine-tuned Llama3-8b with reference summaries (L3-8b-FT Sum) to generate concise responses. This approach consistently achieves significantly higher scores across traditional metrics (R1, R2, BeF1) as well as improved GPT-4 Critique and UniEval evaluations (Table 8). These results suggest that, although there is some overlap between spotlights and summaries, the optimization for spotlights introduces only a partial summary flavor, while remaining distinct from a true summary.

7 Conclusion

In this paper, we introduced the concept of “Spotlight,” a novel, concise representation of source documents that offers a different perspective to conventional summaries and highlights. Our work identifies a set of unique characteristics that differentiate spotlights, emphasizing their ability to present key aspects of a document in a faithful yet compelling way, effectively capturing the reader’s attention as well as drawing their attention to the source document. To validate this concept, we prepared four carefully annotated datasets, which served as the basis for a comprehensive analysis. Through rigorous quantitative evaluations and human studies, we established nine key features that differentiate spotlights primarily in terms of coverage and presentation style. Building on these findings, we proposed a Direct Preference Optimization (DPO)-based fine-tuning algorithm to generate spotlights. Extensive evaluation using both traditional metrics and LLM-based assessments demonstrate the effectiveness of our approach. We believe that, similar to summarization, there is a significant potential for developing various types of spotlights — such as aspect-based, query-focused, and personalized versions — which will be the focus of our immediate future work.

Limitations

Our datasets are neither multilingual nor multimodal. So, we need to explore how state-of-the-art approaches can be utilized in multilingual and multimodal scenarios for spotlight generation - which we aim to do as a part of future work. We have also explored some edge cases and limitations of the model generated spotlights as in Appendix J.

Ethics Statement

Our work does not reveal any personal sensitive information and we use publicly available benchmarked datasets and models in different contexts.

References

Xiang Ao, Ling Luo, Xiting Wang, Zhao Yang, Jiun-Hung Chen, Ying Qiao, Qing He, and Xing Xie. 2023. Put your voice on stage: Personalized headline generation for news articles. *ACM Transactions on Knowledge Discovery from Data*, 18(3):1–20.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. Pens: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92.

Kristjan Arumae, Parminder Bhatia, and Fei Liu. 2019. Towards annotating and creating sub-sentence summary highlights. *arXiv preprint arXiv:1910.07659*.

Tal Baumel, Raphael Cohen, and Michael Elhadad. 2016. Topic concentration in query focused summarization datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

I Cachola, K Lo, A Cohan, and DS Weld. 2020. Tldr: Extreme summarization of scientific documents. *Findings of EMNLP*.

John S Caylor et al. 1973. Methodologies for determining reading requirements of military occupational specialties. "No Title".

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 9–16. IEEE.

Roshni Chakraborty, Maitreyi Bhavsar, Sourav Kumar Dandapat, and Joydeep Chandra. 2019. Tweet summarization of news articles: An objective ordering-based perspective. *IEEE Transactions on Computational Social Systems*, 6(4):761–777.

Sangwoo Cho, Kaiqiang Song, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2020. Better highlighting: Creating sub-sentence summary highlights. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6282–6300.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. *arXiv preprint arXiv:1604.00400*.

Arman Cohan and Nazli Goharian. 2017. Scientific article summarization using citation-context and article’s discourse structure. *arXiv preprint arXiv:1704.06619*.

Edward Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205.

Carlos A Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. Heads: Headline generation as sequence prediction using an abstract feature-rich space. In *Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 133–142.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Haoxiang Deng, Yi Zhu, Ye Wang, Jipeng Qiang, Yun-hao Yuan, Yun Li, and Runmei Zhang. 2024. Prompt-tuning for clickbait detection via text summarization. *arXiv preprint arXiv:2404.11206*.

Yan Du and Hua Huo. 2020. News text summarization based on multi-feature and fuzzy logic. *IEEE Access*, 8:140261–140272.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Figen Beken Fikri, Kemal Oflazer, and Berrin Yanikoglu. 2021. Semantic similarity based evaluation for abstractive news summarization. In *Proceedings of the 1st workshop on natural language generation, evaluation, and metrics (GEM 2021)*, pages 24–33.

Rudolf Franz Flesch et al. 1960. How to write, speak, and think more effectively. (*No Title*).

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Dominik Frefel. 2020. Summarization corpora of wikipedia articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6651–6655.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II* 41, pages 87–93. Springer.

Tomsa Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. *arXiv preprint arXiv:2309.17332*.

Jade Goldstein, Gary M Ciany, and Jaime G Carbonell. 2007. Genre identification and goal-focused summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 889–892.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Souradip Guha, Ankan Mullick, Jatin Agrawal, Swetarekha Ram, Samir Ghui, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2021. Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature. *Computational Materials Science (Comput. Mater. Sci.)*, 192:110325.

Robert Gunning. 1952. The technique of clear writing. (*No Title*).

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 583–592.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Xiaojiang Huang, Xiaojun Wan, and Jianguo Xiao. 2014. Comparative news summarization using concept-based optimization. *Knowledge and information systems*, 38:691–716.

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*.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Sanjeev Kumar Karn, Mark Buckley, Ulli Waltinger, and Hinrich Schütze. 2019. News article teaser tweets and how to generate them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3967–3977.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yushan Liu, Zili Wang, and Rui Feng Yuan. 2024. Querysum: A multi-document query-focused summarization dataset augmented with similar query clusters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18725–18732.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680.

Yuning Mao, Ming Zhong, and Jiawei Han. 2022. Citessum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. *arXiv preprint arXiv:2205.06207*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

Rajdeep Mukherjee, Uppada Vishnu, Hari Chandana Peruri, Sourangshu Bhattacharya, Koustav Rudra, Pawan Goyal, and Niloy Ganguly. 2022. Mtlts: A multi-task framework to obtain trustworthy summaries from crisis-related microblogs. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 755–763.

Ankan Mullick. 2022. Multilingual intent detection and key information extraction for domain specific applications.

Ankan Mullick. 2023a. Exploring multilingual intent dynamics and applications. *IJCAI Doctoral Consortium*.

Ankan Mullick. 2023b. Novel intent detection and active learning based classification (student abstract). *arXiv e-prints*, pages arXiv–2304.

Ankan Mullick, Animesh Bera, and Tapas Nayak. 2021. Rte: A tool for annotating relation triplets from text. *arXiv preprint arXiv:2108.08184*.

Ankan Mullick, Ayan Kumar Bhowmick, Ravi Kokku, Prasenjit Dey, Pawan Goyal, Niloy Ganguly, et al. 2024a. Long dialog summarization: An analysis. *arXiv preprint arXiv:2402.16986*.

Ankan Mullick, Sombit Bose, Abhilash Nandy, Gajula Chaitanya, and Pawan Goyal. 2024b. A pointer network-based approach for joint extraction and detection of multi-label multi-class intents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15664–15680.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024c. On the persona-based summarization of domain-specific documents. *arXiv preprint arXiv:2406.03986*.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Aditya Vempaty, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024d. Leveraging the power of llms: A fine-tuning approach for high-quality aspect-based summarization. *arXiv preprint arXiv:2408.02584*.

Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2024e. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659.

Ankan Mullick, Surjodoy Ghosh D, Shivam Maheswari, Srotaswini Sahoo, Suman Kalyan Maity, and Pawan Goyal. 2018a. Identifying opinion and fact subcategories from the social web. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work*, pages 145–149.

Ankan Mullick, Pawan Goyal, and Niloy Ganguly. 2016. A graphical framework to detect and categorize diverse opinions from online news. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 40–49.

Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2017a. Extracting social lists from twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 391–394.

Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2018b. Harnessing twitter for answering opinion list queries. *IEEE Transactions on Computational Social Systems*, 5(4):1083–1095.

Ankan Mullick, Mukur Gupta, and Pawan Goyal. 2024f. Intent detection and entity extraction from biomedical literature. *LREC-COLING 2024*, page 271.

Ankan Mullick, Shivam Maheshwari, Pawan Goyal, and Niloy Ganguly. 2017b. A generic opinion-fact classifier with application in understanding opinionatedness in various news section. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 827–828.

Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R Raghav, G Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.

Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, R Raghav, and Roshni Kar. 2022a. An evaluation framework for legal document summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4747–4753.

Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, and R Raghav. 2022b. Fine-grained intent classification in the legal domain. *arXiv preprint arXiv:2205.03509*.

Ankan Mullick, Shubhraneel Pal, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2022c. Using sentence-level classification helps entity extraction from material science literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4540–4545.

Ankan Mullick, Sourav Pal, Projal Chanda, Arijit Panigrahy, Anurag Bharadwaj, Siddhant Singh, and Tammoy Dam. 2019. D-fj: Deep neural network based factuality judgment. *Technology*, 50:173.

Ankan Mullick, Sukannya Purkayastha, Pawan Goyal, and Niloy Ganguly. 2022d. A framework to generate high-quality datapoints for multiple novel intent detection. *arXiv preprint arXiv:2205.02005*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Tohida Rehman, Debarshi Kumar Sanyal, and Samiran Chattopadhyay. 2023a. Research highlight generation with elmo contextual embeddings. *Scalable Computing: Practice and Experience*, 24(2):181–190.

Tohida Rehman, Debarshi Kumar Sanyal, Samiran Chattopadhyay, Plaban Kumar Bhowmick, and Partha Pratim Das. 2023b. Generation of highlights from research papers using pointer-generator networks and scibert embeddings. *IEEE Access*.

Leonardo FR Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Geunbae Lee, and Jungseul Ok. 2024. Key-element-informed sllm tuning for document summarization. *arXiv preprint arXiv:2406.04625*.

Olivier Salaün, Aurore Troussel, Sylvain Longhais, Hannes Westermann, Philippe Langlais, and Karim Benyekhlef. 2022. Conditional abstractive summarization of court decisions for laymen and insights from human evaluation. In *Legal Knowledge and Information Systems*, pages 123–132. IOS Press.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. *FineSurE: Fine-grained summarization evaluation using LLMs*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Sajad Sotudeh, Hanieh Deilamsalehy, Franck Dernoncourt, and Nazli Goharian. 2021. Tldr9+: a large scale resource for extreme summarization of social media posts. *arXiv preprint arXiv:2110.01159*.

Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Navonil Majumder, Amir Hussain, Lotfollah Najjar, Newton Howard, Soujanya Poria, and Alexander Gelbukh. 2023. Wikides: A wikipedia-based dataset for generating short descriptions from paragraphs. *Information Fusion*, 90:265–282.

Gianluca Vico and Jan Niehues. 2022. Ted talk teaser generation with pre-trained models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8067–8071. IEEE.

Jesse Vig, Alexander Richard Fabbri, Wojciech Kryściński, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Zhongyu Wei and Wei Gao. 2018. Utilizing microblogs for automatic news highlights extraction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 277–296. World Scientific.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574. Association for Computational Linguistics.

Mateusz Woźny and Mateusz Lango. 2023. Generating clickbait spoilers with an ensemble of large language models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 431–436.

Weihan Xu, Paul Pu Liang, Haven Kim, Julian McAuley, Taylor Berg-Kirkpatrick, and Hao-Wen Dong. 2024. Teasergen: Generating teasers for long documentaries. *arXiv preprint arXiv:2410.05586*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. Macsum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

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 Dataset

Domain	Categories
HealthCare	Death, Diagnosis, Differential diagnosis, Diagnosis-Classification
Education	History, Geography, Taxonomy, Education
Life and Career	Career, Political Career, Personal Life, Life and career
Music	Production, Composition, Soundtrack, Track Listing

Table 9: Domain-wise breakdown of Categories in Wikipedia dataset

Figure 3 demonstrate how we repurpose attributes of News dataset as spotlights.

B Spotlight: Characterization Details

In this section, we further discuss different spotlight characteristics in detail. While comparing with summary, analysis has been done using two different types of summaries - Human Annotated (60 Articles of each dataset) and GPT-4 LLM generated (for all articles of each dataset).

Metric	Better Score
Flesch-Kincaid (Flesch, 1948)	Lower (↓)
Flesch Reading (Flesch et al., 1960)	Higher (↑)
Gunning Index (Gunning, 1952)	Lower (↓)
Dale-Chall Score (Dale and Chall, 1948)	Lower (↓)
McLaughlin's SMOG Score (Caylor et al., 1973)	Lower (↓)
FORCAST (Caylor et al., 1973)	Higher (↑)

Table 10: Comparison of Readability Metrics

B.1 Readability

As shown in Table 10, we utilize six different readable metrics as following:

- (i) *Flesch-Kincaid Grade Level* (Flesch, 1948): The Flesch-Kincaid Grade Level is a readability metric indicating the U.S. school grade level needed to understand a text, calculated from sentence length and syllable count, with lower scores signifying easier readability.
- (ii) *Flesch Reading Ease Score* (Flesch et al., 1960): The Flesch Reading Ease Score is a readability metric, ranging from 0 to 100, that assesses text complexity based on sentence length and syllable count, with higher scores indicating easier readability.
- (iii) *Gunning Fog Index* (Gunning, 1952): The Gunning Fog Index is a readability metric estimating the years of formal education needed to understand a text on the first reading, with lower scores indicating easier readability and less educational requirement.
- (iv) *Dale-Chall Score* (Dale and Chall, 1948): The

Dale-Chall Score is a readability metric assessing text complexity based on sentence length and the percentage of difficult words, with lower scores indicating easier-to-read texts that use common words and shorter sentences, making them accessible to a broader audience.

(v) *McLaughlin's SMOG Score* (Caylor et al., 1973): The SMOG (Simple Measure of Gobbledygook) formula estimates the readability of a text by assessing the number of polysyllabic words in a set number of sentences, with lower scores indicating easier readability and requiring fewer years of education to understand.

(vi) *FORCAST Formula Score* (Caylor et al., 1973): The FORCAST formula assesses text complexity, especially in technical and scientific documents, by counting single-syllable words in a 150-word sample, with higher scores indicating easier readability contrary to many readability formulas.

Table 11 shows complete comparative evaluations of six different readability metrics for spotlight (Spot) vs. Human Generated summary (Summ). Similarly, comparative evaluations of six different readability metrics for spotlight (Spot) vs. GPT-4 summary (Summ) are presented in Table 13. Both the Tables infer that all readability metrics indicate that spotlight is more readable than summary, similar to Table 3 assessment.

B.2 Informativeness and Degree of Extraction

Exploring further, we add Rouge-2 (R2) and Rouge-L (RL) evaluations for the Informativeness and Degree of Extraction along with the Rouge-1 (R1) score as discussed already in the main paper in Section 3.1 (D and E).

We analyze the number of sentences in the document that achieve the R1, R2, and RL score above a specified threshold th with at least one sentence in the summary or spotlight. This approach assumes that if a sentence meets the threshold, its information has been incorporated into the summary or spotlight - i.e. informativeness. However, sentences incorporate information in an abstractive fashion so the threshold th needs to be set at a relatively low value. At a value of $th = 0.3$ our findings of different Rouge (1, 2 and L) scores at the Column ‘Informativeness’ in Table 12 and 14 respectively for human-generated and GPT-4 generated summaries, indicates that a larger number of sentences from the source document are incor-

Dataset		Document										Spotlight	
News	COLUMBIA, S.C. (AP): A 63-year-old worker died in the public bathroom of a South Carolina department store, but her body was not discovered for four days, authorities said. Bessie Durham, who worked for an outside company that cleans the Belk store at Columbiana Centre, was found dead Monday, investigators said. Her cleaning cart was outside the restroom. Durham was last seen Thursday at work and her body was found shortly after her family filed a missing person report, Columbia Deputy Police Chief Melron Kelly, told WIS-TV. The Lexington County Coroner's Office said there are no signs someone killed Durham or that she was using drugs. ... Belk is helping authorities figure out what happened and is sending its deepest condolences to Durham's family, the company said in a statement.	COLUMBIA, S.C. (AP): A 63-year-old worker died in the public bathroom of a South Carolina department store, but her body was not discovered for four days, authorities said. Bessie Durham, who worked for an outside company that cleans the Belk store at Columbiana Centre, was found dead Monday, investigators said. Her cleaning cart was outside the restroom. Durham was last seen Thursday at work and her body was found shortly after her family filed a missing person report, Columbia Deputy Police Chief Melron Kelly, told WIS-TV. The Lexington County Coroner's Office said there are no signs someone killed Durham or that she was using drugs. ... Belk is helping authorities figure out what happened and is sending its deepest condolences to Durham's family, the company said in a statement.										Cleaner Was Dead In Belk Bathroom For 4 Days Before Body Found: Police. The 63-year-old woman was seen working at the South Carolina store on Thursday. She was found dead Monday after her family reported her missing, authorities said.	
Wikipedia	Peter P. Pitchlynn was born in Noxubee County, Mississippi, January 30, 1806 as the first son of Sophia Folsom, a Choctaw of partly Anglo-American descent; her mother "Natikal" was Choctaw and her father was Ebenezer Folsom, an Anglo-American trader. Sophia's Choctaw name was "Lk-lo-ha-wah" (Loved but lost). Sophia Folsom and John Pitchlynn had married in 1804. Pitchlynn was buried there in Congressional Cemetery. He was the third Native American to be buried there, after his fellow Choctaw chief Pushmataha and the Apache chief Taza. Pitchlynn was reported to have told of the origin of the Choctaw: according to the traditions of the Choctaws, the first of their race came from the bosom of a magnificent sea.	Peter P. Pitchlynn was born in Noxubee County, Mississippi, January 30, 1806 as the first son of Sophia Folsom, a Choctaw of partly Anglo-American descent; her mother "Natikal" was Choctaw and her father was Ebenezer Folsom, an Anglo-American trader. Sophia's Choctaw name was "Lk-lo-ha-wah" (Loved but lost). Sophia Folsom and John Pitchlynn had married in 1804. Pitchlynn was buried there in Congressional Cemetery. He was the third Native American to be buried there, after his fellow Choctaw chief Pushmataha and the Apache chief Taza. Pitchlynn was reported to have told of the origin of the Choctaw: according to the traditions of the Choctaws, the first of their race came from the bosom of a magnificent sea.										Educated in Choctaw culture and American schools, in 1825 Pitchlynn helped found a school for Choctaw boys: the Choctaw Academy in Kentucky. He also worked to reduce the sale of alcohol in their territory. After removal to Indian Territory in the 1830s, he was appointed by the National Council in 1845 as the Choctaw Delegate (akin to an ambassadorship) to Washington, D.C. At the time, the Nation was proposing to be recognized by the US Congress as a territory.	
Research Presentation	1 Introduction: Artificial neural networks are nowadays trained successfully to solve a large variety of learning tasks. However, a large number of fundamental questions surround their impressive success. 1.1. Main contributions 2 Setup and preliminaries 2.1 One-hidden layer neural network and training loss Model. Let us fix an integer $n \in \mathbb{N}^*$ as well as 3 Convergence and implicit bias characterisation 3.1 Main result: Theorem [1] below states our main result on the convergence and implicit bias of one-hidden layer. 3.2 Discussion: Even if the orthogonal setting we consider is quite restrictive, it carries several characteristics that may be generic, either because they have been observed empirically, shown in related contexts or simply conjectured. We discuss these important points below.	1 Introduction: Artificial neural networks are nowadays trained successfully to solve a large variety of learning tasks. However, a large number of fundamental questions surround their impressive success. 1.1. Main contributions 2 Setup and preliminaries 2.1 One-hidden layer neural network and training loss Model. Let us fix an integer $n \in \mathbb{N}^*$ as well as 3 Convergence and implicit bias characterisation 3.1 Main result: Theorem [1] below states our main result on the convergence and implicit bias of one-hidden layer. 3.2 Discussion: Even if the orthogonal setting we consider is quite restrictive, it carries several characteristics that may be generic, either because they have been observed empirically, shown in related contexts or simply conjectured. We discuss these important points below.										So the case of two layer relu neural networks remains a huge mystery to study and even more in the case of regression tasks. And so in this work we focus on the following three questions. So do this network converge to zero training loss. Do they show some implicitly by us? ... And after this first phase we have a second phase where one cluster will grow in normally fitting for example positive labels. And then during the first phase is the same happened with other cluster for the negative labels. So this is it for the summary of our work.	

Figure 2: Examples of (Document, Spotlight) pairs from different datasets, An example from CSPubSum is given in Figure 1, The documents and the example spotlight of Research Presentation have been truncated in this visualization owing to their length. Across the three examples, the spotlights consistently transform lengthy, detail-heavy texts into short, curiosity-driven narratives. In the News case, procedural details and background information are removed, leaving a striking account that emphasizes the unusual and attention-grabbing aspect of the story—the body discovered after four days. The Wikipedia example shifts from an encyclopedic biographical entry to a narrative that foregrounds milestones and memorable achievements, making the article more engaging and easier to recall. Finally, the Research Presentation example reframes dense, technical exposition into an accessible narrative by highlighting the central research question and its intrigue rather than methodological detail. Together, these examples illustrate how spotlights distill complex information into focused, compelling narratives that spark interest rather than merely summarize content.

Dataset		Distribution		Compactness↓	Readability (%)						Interest-Generation (%)↑
		Entropy↓	Skew50↑		Flesch Kincaid↓	Flesch Reading↑	Gunning ↓	Dale Chall↓	McLaughlin's↓	FORCAST↑	
News	Spot	0.69	0.61	0.52	10.57	15.46	16.81	13.21	20.69	1.96	80.0
	Summ	0.96	0.02	1.53	16.89	9.45	20.10	17.24	26.10	-0.06	24.0
CSPubSum	Spot	0.62	0.47	0.62	16.28	9.67	19.73	13.21	26.72	2.01	86.0
	Summ	0.96	0.01	1.28	20.72	6.42	23.55	17.74	30.39	-0.08	30.0
Wikipedia	Spot	0.76	0.29	0.82	12.62	16.26	16.37	13.62	23.32	1.36	86.0
	Summ	0.95	0.05	1.48	16.24	9.35	19.19	17.59	27.35	0.04	26.0
Research	Spot	0.65	0.36	0.28	11.84	44.10	11.85	13.82	25.12	2.32	82.0
Presentation	Summ	0.92	0.04	0.69	16.03	23.83	17.21	17.23	36.25	0.02	23.0

Table 11: Feature level analysis of spotlight (Spot) vs. human annotated summary (Summ) for 60 articles of each datasets, For Entropy, Compactness, Flesch Kincaid, Gunning score, Dale-Chall and McLaughlin's score lower values signifies better result, whereas for Skew50, Flesch Reading score, FORCAST and Interest-Generation higher values signifies a better result as discussed in Section 3, More Exploratory Results of Table 3.

porated into summaries compared to spotlights. It signifies that the summary is more informative than the spotlight.

However, when the threshold value is increased (at a value of $th = 0.8$) — indicating a higher level of extractiveness — the trend reverses, with fewer sentences meeting the threshold in summaries com-

pared to spotlights as shown in ‘Degree of Extraction’ column using different Rouge (1, 2 and L) scores in Table 12 and 14 respectively for human-generated and GPT-4 generated summaries. The increased extractiveness of the spotlight can also be attributed to its enhanced readability. This trend is consistent across different ROUGE metrics (R1,



By Rumni Saha, Contributor

A citizen of the world in love with life.

Feb 20, 2015, 02:55 PM EST | Updated Apr 22, 2015

This post was published on the now-closed HuffPost Contributor platform. Contributors control their own work and posted freely to our site. If you need to flag this entry as abusive, send us an email.



Spotlight



Figure 3: News Headline and the Short Description taken as spotlight for the News dataset. This figure illustrates the construction process of the News spotlight dataset. Source articles are paired with human-curated title and short description.

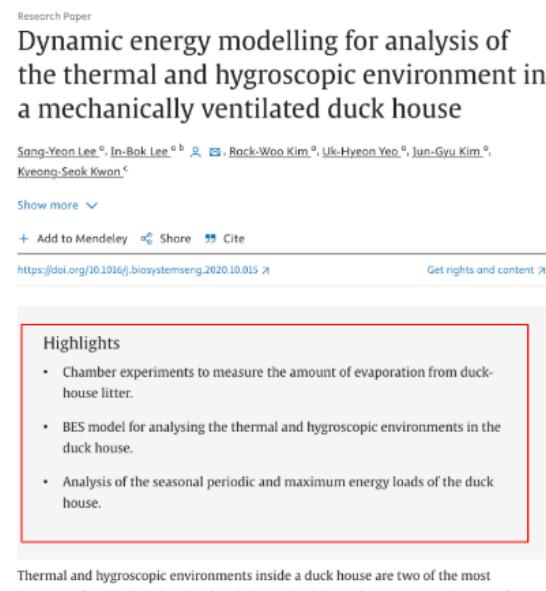


Figure 4: Author written highlights taken as spotlight for the CSPubSum dataset. From the Research Papers the author-written bullet points serve as a compelling entryway into the research paper, offering a concise preface that draws attention to the central questions and findings while setting an engaging narrative tone.

R2, RL), as shown in the mentioned Tables.

We have finalized the thresholds after manual verification. These threshold values are closely similar to the thresholds used by (Bhandari et al., 2020) for constructing Semantic Content Units (SCUs) as a basis for key fact extraction. The 0.3 threshold serves to identify minimal content overlap—indicating that a generated piece contains

Dataset		<i>threshold=0.3</i>			<i>threshold=0.8</i>		
		R1	R2	RL	R1	R2	RL
News	Spot	0.910	0.628	0.521	0.021	0.015	0.016
	Summ	2.585	1.185	2.314	0.011	0.010	0.012
CSPubSum	Spot	1.562	0.362	0.782	0.109	0.025	0.110
	Summ	3.433	0.966	2.566	0.100	0.012	0.083
Wikipedia	Spot	1.826	0.542	1.528	0.032	0.019	0.029
	Summ	4.000	1.233	3.250	0.025	0.015	0.020
Research	Spot	1.431	0.371	0.836	0.052	0.023	0.024
	Presentation	3.569	1.025	2.865	0.026	0.018	0.019

Table 12: Feature level analysis on Informativeness and Degree of Extraction: spotlight (Spot) vs. Human Annotated summary (Summ), Additional Results of Table 3

at least some of the critical information from the source—while the 0.8 threshold ensures that the content is sufficiently derived from the source, reflecting a high degree of extractiveness.

B.3 Interest Generation aspects of a spotlight

The interest generation in spotlights stems from their ability to provide a concise, engaging, and curiosity-inducing preview of the main content. They serve as a hook by emphasizing key points, novel insights, or thought-provoking questions, which draw the audience into exploring the full content. In spotlighting tasks, well-crafted information improves user engagement by offering an informative yet incomplete glimpse, triggering interest to read further. Interest in a spotlight is generated by presenting a novel insight, an unresolved question, or an impactful statement that hints at the significance of the work while leaving room for deep exploration. A well-crafted first sentence, often emphasizing surprise, controversy, or real-world implications, plays a crucial role in capturing attention. We have done an annotation to find out the reasoning behind the interest generation by the same annotators. Figure 6 shows the reasoning behind the generation of interest for a sample from the Wikipedia dataset. A spotlight sparks interest when it introduces new or unconventional ideas, methods, or results in a neutral tone that invites them to explore further. By presenting departures from typical approaches—such as relaxed constraints, adaptive techniques, or novel theoretical insights—it encourages the reader to consider how these changes work and why they matter. Rather than merely summarizing conclusions, the spotlight subtly emphasizes the underlying process, uncertainties, or implications, prompting the reader to mentally engage with open questions or deeper mechanisms, as also seen from Fig 1.

To quantitatively evaluate the effectiveness of

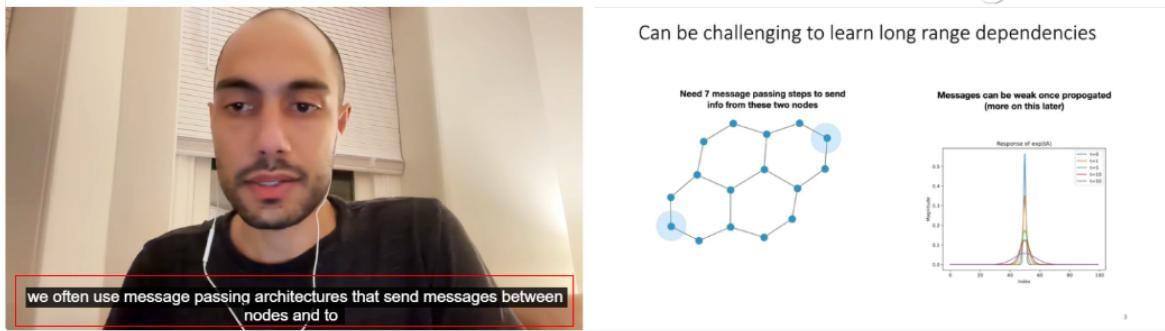


Figure 5: Conference Video Transcription taken as spotlight for the Research Presentation dataset. From slide decks and talk transcripts, concise spotlights are crafted to reflect the most engaging aspects of the research story. Unlike abstracts or slide summaries, these spotlights focus on intrigue, contributions, and memorable phrasing, ensuring that the dataset captures narrative appeal specific to academic communication.

Dataset		Distribution		Compactness↓	Readability (%)						Interest-Generation (%)↑
		Entropy↓	Skew50↑		Flesch Kincaid↓	Flesch Reading↑	Gunning ↓	Dale Chall↓	McLaughlin's↓	FORCAST↑	
News	Spot	0.58	0.43	0.68	6.57	65.86	8.46	10.58	14.35	2.82	72.0
	Summ	0.91	0.33	0.78	11.24	49.31	13.40	11.42	16.98	-0.90	72.0
CSPubSum	Spot	0.66	0.47	0.50	8.69	49.67	11.75	10.61	17.55	4.01	81.0
	Summ	0.89	0.22	0.77	14.31	23.74	16.42	12.68	20.40	-0.27	37.0
Wikipedia	Spot	0.76	0.29	0.69	8.87	59.22	10.97	11.19	16.12	1.15	83.0
	Summ	0.88	0.18	0.89	9.80	52.99	11.95	12.36	17.35	1.04	29.0
Research Presentation	Spot	0.62	0.38	0.26	11.64	38.10	13.55	11.80	25.10	1.51	75.0
	Summ	0.84	0.18	0.72	16.23	24.23	17.81	15.43	32.65	0.05	25.0

Table 13: Feature level Analysis of spotlight (Spot) vs. GPT-4 generated summary (Summ) for entire datasets, For Entropy, Compactness, Flesch Kincaid, Gunning score, Dale-Chall and McLaughlin's score lower values signifies better result, whereas for Skew50, Flesch Reading score, FORCAST and Interest-Generation higher values signifies a better result as discussed in Section 3, Additional Results of Table 3

She was the **winner of the third series of "The Voice Australia"**. In 2016, Nissen competed in "Dansk Melodi Grand Prix 2016", Denmark's national final for the Eurovision Song Contest, performing the song "Never Alone". She came in second place with 36% of the vote. In 2017, Nissen **competed in "Dansk Melodi Grand Prix 2017"** with the song "Where I Am", which she co-wrote. Anja went on to win and represented Denmark in the Eurovision Song Contest 2017, where she **finished the final in the top 20**. Nissen **performed at the 2014 Carols by Candlelight** in Melbourne, Australia.

Figure 6: Interest-Generation aspect of a spotlight on Life and Career of Anja Nissen from Wikipedia dataset, where annotators selectively identify the most engaging and narratively significant parts of a text. The **bold** segments represents the achievements, competitions and performances of Anja Nissen that capture attention and convey intrigue, rather than covering every detail. This mirrors the idea of spotlight generation which focus on a few compelling elements that can spark curiosity and draw readers into the broader narrative.

Dataset		threshold=0.3			threshold=0.8		
		R1	R2	RL	R1	R2	RL
News	Spot	0.422	0.201	0.365	0.014	0.013	0.014
	Summ	1.243	0.261	0.995	0.007	0.002	0.007
CSPubSum	Spot	0.595	0.218	0.513	0.101	0.020	0.101
	Summ	2.792	1.245	2.441	0.056	0.008	0.047
Wikipedia	Spot	1.519	0.235	0.995	0.027	0.016	0.026
	Summ	2.241	0.332	1.204	0.018	0.010	0.014
Research Presentation	Spot	1.256	0.352	0.765	0.050	0.021	0.024
	Summ	2.856	1.011	1.965	0.025	0.017	0.016

Table 14: Feature level analysis on Informativeness and Degree of Extraction: spotlight (Spot) vs. GPT-4 generated summary (Summ), Additional Results of Table 3

spotlights in generating interest, we conducted pairwise comparisons between spotlights and summaries for the same 50 documents. Table 15 shows that spotlights were preferred significantly more often than summaries for interest generation. This is evident from the win-rates of spotlight as evident from 3rd Column of Table 15 : 68% for News, 64% for CSPubSum, 62% for Wikipedia, and 60% for Research Presentation which is much higher than summary.

Dataset	Both of them	Spotlight only	Summary only	None of them
News	6	34	6	4
CSPubSum	11	32	4	3
Wikipedia	12	31	1	6
Research Presentation	11	30	1	8

Table 15: Pairwise annotation results showing preferences for spotlight vs. summary across four datasets. ‘Both’ indicates both generates an interest to read the document; ‘None’ indicates neither generated interest.

C Spotlight Generation Process

Along with other approaches, we also explore a different preference-based optimization approach - KTO as discussed below.

KTO (Ethayarajh et al., 2024): We initially use the summaries for documents in the dataset to create a binary preference Kahneman-Tversky Optimization (KTO) dataset. Each document from the original dataset contributes one upvoted and one downvoted example. The document paired with the ground truth spotlight forms the upvoted example (indicating a preferred completion), while the document paired with a summary forms the downvoted example (denoting undesirable). The rationale behind this is to generate spotlights over summary with higher rewards. This results in the KTO dataset being twice the size of the original dataset, with an equal number of upvoted and downvoted examples. We then apply KTO to the instruction-tuned versions of base models (e.g., Llama-7b-chat-hf, Mistral-7B-Instruct-v0.2) using the KTO dataset, balancing desirable and undesirable weights. Under this setup, we evaluate the performance of the KTO on three LLMs (Mistral-7b, Llama2-7b, and Llama3-8b) by comparing its generations with ground truths on the test set. We also apply KTO on SFT but the results are not better. Among all experiments with different versions of base models, we report the best outcome only (which was achieved when KTO was done directly on the instruct models). These results are denoted in Tables 21, 23, 25, 27 in rows denoted by KTO (Mst-7b), KTO (L2-7b) and KTO (L3-Bb).

D Prompts

We use prompting in two stages - data generation, finetune-inference and critique. There are two kinds of prompts - system prompt and user prompt.

D.1 Finetune and Inference prompt

system prompt - You are a helpful, respectful and honest assistant. Your task is to write spotlight of documents. A spotlight is a short, concise overview of the document. It is meant to spark curiosity in the reader to read the entire article. But it does not provide much coverage of the content of the document and that is how it differs from a summary. Below is an instruction, paired with a document that provides further context.

user prompt - ### Instruction:

Document: {document}

Response:

Specific for different datasets are:

Wikipedia. Write the spotlight of the following document based on the {category} of {article title}. The spotlight need not have detailed information coverage but should include only the key points that makes the reader curious about the entire article.

CSPubSum. Write the spotlight of the following scientific article entitled {main_title} presented as a document. The spotlight need not have detailed information coverage but should include only the key points that makes the reader curious about the entire article.

News. Write a headline for the following news article presented as document. Also include a short description of the article in not more than 4 sentences that can be presented as its highlight. The headline together with the highlight is the spotlight for the news article. It need not have detailed information coverage but should make the reader curious about the news article.

Research Presentation. Write the spotlight of the following scientific article presented as a document. The spotlight need not have detailed information coverage but should include only the key points that makes the reader curious about the entire article.

D.2 LLM Critique

For LLM (GPT-4, Gemini) Critique, we ask the LLM with system and user prompt for evaluating the model-generated spotlights.

system: You are an AI assistant who is to evaluate the spotlight of a textual document. You need to return a score between 0 and 1 reflecting the quality of the generated spotlight based on some criteria. A spotlight is a short summary of some of the most important sentences of a document that creates a curiosity to read the actual document.

user: You are given a document and the corresponding spotlight of the document generated by a language model as follows.

Document: {document}

Ground truth spotlight : {label spotlight}

spotlight generated: {model generated spotlight}

Evaluate the above spotlight for the document in terms of each of the following criteria and return only a score between 0 and 1 without any explanation:

- Rate the spotlight – whether the spotlight is good or bad. A good spotlight provides a concise, captivating narrative that emphasizes a document’s most intriguing aspects, fostering reader engagement. A bad spotlight either fragments key points without cohesion or overwhelms with excessive detail, failing to engage the reader effectively.

E Related Work

We study the Spotlight generation task in diverse directions and various concepts similar to the spotlight.

1. Summarization: Researchers focus on summarization task on different domains - news documents (Du and Huo, 2020; Goyal et al., 2022; Huang et al., 2014; Fikri et al., 2021), scientific articles (Cohan and Goharian, 2016, 2017; Cachola et al., 2020) and other different areas - social media (Mukherjee et al., 2022; Chakraborty et al., 2019; Mullick et al., 2024d), wikipedia (Ta et al., 2023; Frefel, 2020) , dialog (Mullick et al., 2024a) etc. TLDR (Cachola et al., 2020), TLDR9+ (Sotudeh et al., 2021), CiteSum (Mao et al., 2022), (Narayan et al., 2018) work on extreme summarization of main concepts of articles. People also focus on different summarization - query focused (Vig et al., 2022; Baumel et al., 2016; Liu et al., 2024), layman summarization (Goldsack et al., 2023;

Salaün et al., 2022; Luo et al., 2022), goal oriented (Goldstein et al., 2007; Ham et al., 2020; Jin et al., 2024), persona-based (Mullick et al., 2024c) etc. CTRLSum (He et al., 2020), (Zhang et al., 2023) and (Ribeiro et al., 2023) focus on summaries with controllable readability. Researchers also work on summarizing documents with respect to key elements (Ryu et al., 2024; Wang et al., 2023). But, the problem of spotlight generation differs from these summarization tasks.

2. Highlight Generation: People focus on highlight detection task in different directions - story highlight (Woodsend and Lapata, 2010), research paper highlight (Rehman et al., 2023a,b), micro-blog news highlight (Wei and Gao, 2018), Teaser generation (Karn et al., 2019), highlight span extraction (Arumae et al., 2019; Cho et al., 2020). But, these work mostly focus on extracting multiple spans as a highlight. Our works differ from the fact that we generate most important part of the article which will entice reader to read through the article.

3. Headline/Title Generation: Researchers aim at developing different approaches - sequence prediction (Colmenares et al., 2015) for headline generation, personalized headline generation (Ao et al., 2021, 2023), self-attention based headline generation (Gavrilov et al., 2019). But our task is different from traditional headline generation work which is too short.

4. Teaser Production: Teasers serve as a powerful strategy for promoting content across various domains. Earlier works cover teaser generation from long documents (Xu et al., 2024), news article (Karn et al., 2019), ted-talks (Vico and Niehues, 2022) etc.

Our work differs from all of the above as Spotlight has different features than summary, highlight, headline and teaser, and serves different perspectives. This concept of spotlight is also different from the notion of entity-extraction (Mullick et al., 2024e; Guha et al., 2021; Mullick et al., 2022c, 2021), opinion context detection (Mullick et al., 2017b, 2016, 2018a,b, 2019, 2017a), intent phrase identification task in specialized contexts (Mullick et al., 2022d,b; Mullick, 2023b,a; Mullick et al., 2022a; Mullick, 2022; Mullick et al., 2024b, 2023, 2024f).

F Experimental Settings

F.1 Time and GPU

We experiment on 80GB A100 GPU with GPU clock cycle 210 MHz. The finetuning and inference time of our finetuned models are in Table 16.

Model	Finetune Time	Inference Time
L2-7b-FT	20 hrs	2hrs 30 mins
DPO (L2-7b)	40 hrs	2hrs 30 mins
KTO (L2-7b)	96 hrs	2hrs 30 mins
Mst-7b-FT	17 hrs	2hrs 10 mins
DPO (Mst-7b)	36 hrs	2hrs 10 mins
KTO (Mst-7b)	83 hrs	2hrs 10 mins
L3-8b-FT	22 hrs	3hrs 20 mins
DPO (L3-8b)	44 hrs	3hrs 20 mins
KTO (L3-8b)	91 hrs	3hrs 20 mins

Table 16: Model Training Time [using 80GB A100 GPU]

F.2 Hyperparameter Optimization

For all fine-tuning (SFT) and alignment (DPO, KTO) experiments, we use QLoRA with 4-bit quantization, LoRA $r = 64$, optimization is done with AdamW optimizer, and an effective training batch size of 16 (with gradient accumulation) over 4 epochs. For SFT, we use a learning rate of 5×10^{-5} with linear schedule. For DPO, we use $\beta = 0.01$ and a learning rate of 5×10^{-6} with cosine schedule. For KTO, we use $\beta = 0.01$, with both desirable and undesirable weights set to default 1 and a learning rate of 5×10^{-6} with linear schedule.

F.3 Train-validation-test Split

Table 17 shows train-validation-test splits of dataset for experiments.

Dataset	Total	Train	Validation	Test
News Category	14080	13000	80	1000
CSPubSum	7591	6408	80	1103
Wikipedia	17323	14279	500	2544
Research Presentation	1230	1000	100	130

Table 17: Statistics of the spotlight datasets used

G Human Annotations

G.1 Task 1 - Identify Interest-Generation

A paragraph describing an article is provided to you and you are asked whether reading that paragraph generates further interests to read the entire document. Use your judgment and perspective to make this determination, ensuring the information is relevant and clear to you.

Annotation Guidelines for Comparative Rating:

We provide instructions with explanations of curiosity generation after reading the provided paragraph

(either a spotlight or summary). User is asked to identify if the paragraph generates curiosity to read the entire article as shown in Fig 7.

Instructions: The instructions are the following

- In this study, you will assess whether a brief description of an article sparks your interest in reading the entire article. You will be provided with a Short Description (SD) for an article.

Carefully review the Short Description. No additional software download is required. Use a browser, preferably Google Chrome, and ensure a stable internet connection. Allocate time judiciously for crafting each of the summaries based on the provided 20 SD instances.

After completion, you will be asked to provide feedback on the generation exercise, platform interaction, and details about your academic background, age, country of birth or experience.

Payment Requirements: Upon completing the study, click on the provided link containing the completion code to redirect you to the Prolific platform. Payment will be processed within one to two weeks.

Ethical Considerations: Adhere to strict confidentiality and data protection standards to ensure privacy. If you have concerns or questions, feel free to reach out, as this study aligns with ethical guidelines.

This study aims to harness diverse perspectives, including those of academic professionals, to refine the curiosity generation for enhanced utility in various contexts.

Please do not use ChatGPT/GPT4 or any Large Language Models - all responses must be produced by human annotators. It is a strict instruction and will be checked manually - if any issue: the submission will be rejected and re-doing will be required.

Participant Prescreening criteria:

- Age above 24
- Primary Language is English
- Minimum Graduation degree is either *Graduate degree (MA/MSc/MPhil/other)* or *Doctorate degree (PhD/other)*
- Approval Rate in platform > 85% and minimum number of previous submissions > 30.

G.2 Task 2 - Summary Generation

In this task, you are asked to generate a summary. You will be provided with a section name and a source document (link) – wikipedia or news article, etc. Read the section of the Source Document and craft summaries. Use your understanding and perspective to tailor the information in a way that is most relevant and comprehensible.

Instructions: Carefully review the Source Document. No additional software download is required. Use a browser, preferably Google Chrome, and ensure a stable internet connection. Allocate time judiciously for crafting each of the summaries based on the provided 10 SD instances. After completion, you will be asked to provide feedback on the generation exercise, platform interaction, and details about your academic background, age, country of birth, and any medical background or experience with the summary generation task. Your summary length (word count) must be around 15% (\pm 20 words) of the document. For example if the Source Document has 1000 words, the summary must have between 130 to 170 words. Please do not use ChatGPT/GPT4 or any Large Language Models - all response should be generated by human properly. It is a strict instruction and will be checked manually - if any issue: it will be rejected and re-doing will be required.

Payment Requirements. Upon completing the study, click on the provided link containing the completion code to redirect you to the Prolific platform. Payment will be processed within one to two weeks.

Ethical Considerations. Adhere to strict confidentiality and data protection standards to ensure privacy. If you have concerns or questions, feel free to reach out, as this study aligns with ethical guidelines.

Participant Prescreening criteria.

- Age above 24
- Primary Language is English
- Minimum Graduation degree is either *Graduate degree (MA/MSc/MPhil/other)* or *Doctorate degree (PhD/other)*
- Approval Rate in platform $> 85\%$ and minimum number of previous submissions > 30 .

G.3 Task 3 - Mini-Story Validation

In this study, you are tasked with evaluating paragraphs to determine which one aligns more closely with the characteristics of a mini story. You will be provided with two paragraphs – which highlights some research work. You will be provided with pairs of paragraphs. Your job is to carefully read both paragraphs in each pair and decide which one feels more like a mini story – that is, concise, engaging, and complete with a clear sense of narrative or resolution. Use your understanding and perspective to tailor the information in a way that is most relevant and comprehensible.

Annotation Guidelines for Comparative Rating: We provide instructions with explanations of mini-story after reading the provided paragraph (either a spotlight or summary/highlight). User is asked to identify which of the either paragraphs or both the paragraphs are a mini-story.

Instructions: The instructions are the following
Review both paragraphs in each pair attentively. Select the paragraph that you believe better captures the essence of a mini story. No prior knowledge is required, and no additional software download is needed. Use a browser, preferably Google Chrome, and ensure a stable internet connection. Allocate sufficient time for each comparison to make thoughtful decisions on the provided 10 instances.

After completion, you will be asked to provide feedback on the generation exercise, platform interaction, and details about your academic background, age, country of birth or experience.

Payment Requirements: Upon completing the study, click on the provided link containing the completion code to redirect you to the Prolific platform. Payment will be processed within one to two weeks.

Ethical Considerations: Adhere to strict confidentiality and data protection standards to ensure privacy. If you have concerns or questions, feel free to reach out, as this study aligns with ethical guidelines.

Please do not use ChatGPT/GPT4 or any Large Language Models - all response should be generated by human properly. It is a strict instruction and will be checked manually - if any issue: the submission will be rejected and re-doing will be required.

Participant Prescreening criteria:-

- Age above 24

- Primary Language is English
- Minimum Graduation degree is either *Graduate degree (MA/MSc/MPhil/other)* or *Doctorate degree (PhD/other)*
- Approval Rate in platform > 85% and minimum number of previous submissions > 30.

Dataset	Both Better(%)	Spotlight (%)	Summary (%)
News Category	72	16	12
CSPubSum	72	16	8
Wikipedia	68	20	12
Research Presentation	80	12	8

Table 18: Mini-Story Validation Study - Prolific

H Examples

One example of the human annotation interface is shown in Fig 7.

I Experimental Results

We analyzed the performance of traditional and LLM-based approaches, including the preference optimization techniques using various referenceless metrics and LLM-critiquing. We also assess different variations of DPO with held-out data.

I.1 Performance of Traditional Baseline Approaches

In addition to the results in Table 5, we have also explored different traditional approaches for spotlight generation and results for News, CSPubSum, Wikipedia and Research Presentation datasets are shown in Table 20, 22, 24 and 26. It is seen that even simple SFT-based approaches on LLMs perform much better than these traditional approaches across all metrics. This performance gap may be due to the advanced architectures and targeted optimizations of LLM models which are fine-tuned specifically for task performance and user preferences, whereas the older models do not benefit from the same level of refinement and cutting-edge techniques, leading to their comparatively lower performance.

I.2 Performance of DPO vis-a-vis other LLM-based approaches

In addition to the results presented in Table 6, we demonstrate the effectiveness of the DPO and KTO approach for spotlight generation across various evaluation criteria for Llama3-8b (L3-8b), Llama2-7b (L2-7b) and Mistral-7b (Mst-7b) large language

model (LLM) for each datasets in Table 21 (for News), 23 (for CSPubSum), 25 (for Wikipedia) and 27 (for Research Presentation). We report evaluations for Llama2-7b, Mistral-7b and Llama3-8b finetuning abbreviated as L2-7b-FT, Mst-7b-FT and L3-8b-FT, general instruction tuned vanilla model L2-7b-VA, Mst-7b-VA and L3-8b-VA, DPO DPO(L2-7b), DPO(Mst-7b) and DPO(L3-8b) and KTO KTO(L2-7b), KTO(Mst-7b) and KTO(L3-8b). It is seen that

- DPO consistently outperforms vanilla models, supervised fine-tuning (SFT), KTO and GPT-4 (1-shot) (in most of the cases) proving its superiority in spotlight generation.
- SFT only approach is a close second, however, DPO consistently improves over SFT methods indicating its utility and the potency of SFT when the reference negative examples (here summaries) are not available
- Vanilla L3-8b-VA performs the worst, struggling in both traditional and advanced evaluations, highlighting the importance of fine-tuning or DPO across all datasets.
- for *News*, *CSPubSum* and *Research Presentation* datasets, Llama3-8b fine-tuning and DPO, both perform better than GPT-4 (1-shot) for traditional metrics whereas for *Research Presentation*, GPT-4 (1-shot) performs better than the fine-tuning model but worse than DPO.
- GPT-4 (1-shot) is strong in UniEval Relevance and Critique (it is best among all in *Wikipedia*) but falls behind in ROUGE and BERTScore, suggesting it produces more readable but not necessarily more extractive spotlights.
- For Wikipedia data, all models show higher Rouge scores with lower UniEval relevance/GPT-4 Critique suggests that while the spotlights have significant lexical overlap with the reference, they may lack deeper semantic coherence.
- Notably, the Llama3-8b based DPO model achieves statistically significant improvements ($p < 0.05$) over other model variations.
- Across all the datasets, both Llama2 and Mistral models consistently show that DPO optimization leads to superior performance over KTO—evidenced by higher ROUGE,

Life and career of Anja Nissen

SD:- She was the winner of the third series of "The Voice Australia". In 2016, Nissen competed in "Dansk Melodi Grand Prix 2016", Denmark's national final for the Eurovision Song Contest, performing the song "Never Alone". She came in second place with 36% of the vote. In 2017, Nissen competed in "Dansk Melodi Grand Prix 2017" with the song "Where I Am", which she co-wrote. Anja went on to win and represented Denmark in the Eurovision Song Contest 2017, where she finished the final in the top 20. Nissen performed at the 2014 Carols by Candlelight in Melbourne, Australia.

Interested

Not Interested

Figure 7: Spotlight curiosity experiment example snapshot

BERT-F1, UniEval relevance, and GPT-4 Critique scores. This aggregated trend reinforces the robustness of the DPO approach in spotlight generation across diverse domains.

- The experimental results indicate that the DPO-based approach outperforms large LLMs (GPT-3.5, Gemini and GPT-4) in 1-shot evaluations for spotlight generation.
- Also, to validate the GPT-4 critique scores against robust human judgment, we conducted an additional human annotation study similar to (Mullick et al., 2024c) to evaluate the alignment between human judgment and automated assessments for 20 documents for each dataset. In this study, annotators were presented with the model-generated spotlight and asked to categorize them into one of two options - 'good' or 'bad'. Then, we calculate the Point-biserial correlation coefficient between GPT-4 critique scores and human judgment. We obtain an overall average of 0.751 point-biserial correlation coefficient between them, which validates GPT-4 critique scores are robust and reliable against human judgment studies.

I.3 Reference-less metrics for Spotlight Generation

In Table 19, we focused on the best-performing models—specifically, the DPO-based Llama3-8b and GPT-4 (1-shot) approaches—and assessed their

outputs using reference-less metrics, including UniEval (Coherence, Consistency, Fluency) and G-Eval(Relevance, Coherence, Factuality, Conciseness). The results indicate that while GPT-4 (1-shot) generates spotlights that are notably readable and contextually relevant, the DPO-based model consistently excels in semantic coherence and consistency across diverse datasets. These findings suggest that the targeted fine-tuning strategy inherent in the DPO approach effectively enhances the model's capacity to capture the nuanced characteristics of high-quality spotlights, underscoring its robustness for applications demanding high semantic fidelity.

I.4 Different Variations of DPO with Held-out Data

Table 6 demonstrates that DPO applied to the larger, fine-tuned Llama3-8b model consistently outperforms other LLM-DPO combinations across all datasets and metrics. This suggests that model size and fine-tuning play a crucial role in DPO's effectiveness. We have explored different DPO configurations using data and fine-tuning as in Table 29:

- *DPO (vnl)*: all data for DPO directly on vanilla (pre-trained) LLMs without fine-tuning.
- *DPO (n/2)*: 50% data for the supervised fine-tuning model and rest 50% data for DPO

Model	UniEval			G-Eval			
	Coherence	Consistency	Fluency	Relevance	Coherence	Factuality	Conciseness
News							
L3-8b-VA	75.3	85.0	80.1	68.6	73.2	75.4	72.5
GPT-4 (1-shot)	83.8	90.2	88.2	82.2	85.8	88.5	83.4
L3-8b-FT	82.8	93.0	89.9	81.5	84.4	87.1	83.1
DPO (L3-8b)	83.4	94.8	90.1	83.6	85.2	88.4	84.4
CSPubSum							
L3-8b-VA	62.5	70.5	72.3	60.3	63.4	64.4	62.8
GPT-4 (1-shot)	66.2	77.8	87.2	66.9	69.1	72.5	68.2
L3-8b-FT	69.8	74.4	88.3	64.8	68.6	70.1	67.4
DPO (L3-8b)	69.5	75.3	89.7	67.4	70.6	73.0	69.5
WikiPedia							
L3-8b-VA	65.2	73.8	80.2	62.1	64.5	66.7	63.5
GPT-4 (1-shot)	73.1	81.9	88.6	71.3	75.5	79.1	73.2
L3-8b-FT	73.7	80.4	90.6	70.3	74.2	78.7	72.8
DPO (L3-8b)	75.9	79.2	91.2	69.0	73.6	77.5	71.2
Research Presentation							
L3-8b-VA	80.4	69.6	85.3	69.2	79.1	73.4	70.5
GPT-4 (1-shot)	88.2	78.9	86.3	74.3	79.4	78.5	75.5
L3-8b-FT	86.9	72.3	88.3	75.8	80.2	78.1	76.4
DPO (L3-8b)	84.8	76.6	89.7	77.4	81.9	80.2	77.5

Table 19: UniEval and G-Eval reference-free evaluation criteria Results (The best outcomes are in **Bold**) for each dataset, Additional Results of Table 6

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1		
L2-7b-FT	34.9	14.6	27.9	31.2	5.5	81.0	80.6	89.6
Mst-7b-FT	32.1	12.3	25.0	28.3	3.8	80.4	79.9	89.4
L3-8b-FT	36.5	15.8	28.1	32.8	6.5	81.5	82.2	89.4
LF	14.9	6.1	10.9	22.2	1.9	71.9	75.7	48.8
T5-FT	18.0	10.6	18.0	25.0	4.0	76.0	77.5	69.2
BART	19.6	8.3	14.2	28.4	3.1	75.8	81.9	71.2
Pegasus	20.1	7.1	14.4	24.7	2.4	75.7	77.2	70.7
Falcon	15.0	2.7	10.0	18.0	0.3	72.0	76.3	48.0
TLDR	29.9	11.1	23.1	24.0	1.5	79.0	77.3	69.3

Table 20: Traditional, UniEval and GPT-4 Critique evaluations on News Dataset - comparison of baselines with SFT, Additional Results of Table 5

Our cutting-edge AI model redefines workplace safety by eliminating tedious individual calibration. Powered by ANFIS, this breakthrough technology delivers fast, accurate, and hassle-free workload assessments—because efficiency shouldn't come at the cost of precision. **The future of ergonomic excellence starts now!**

Figure 8: Spotlight Example of failure case

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1		
L2-7b-VA	15.5	3.8	11.4	17.2	0.5	70.6	62.3	72.6
L2-7b-FT	34.9	14.6	27.9	31.2	5.5	81.0	80.6	89.6
DPO (L2-7b)	35.5	15.2	28.5	31.6	6.1	81.2	82.3	89.7
KTO (L2-7b)	31.3	11.9	23.6	24.3	2.8	79.5	81.9	82.3
Mst-7b-VA	20.1	5.6	14.2	22.9	0.7	74.2	70.1	76.5
Mst-7b-FT	32.1	12.3	25.0	28.3	3.8	80.4	79.9	89.4
DPO (Mst-7b)	34.4	12.9	25.8	29.8	4.3	82.5	82.6	91.4
KTO (Mst-7b)	26.3	9.5	19.7	24.2	2.1	70.6	80.3	89.3
L3-8b-VA	15.9	6.8	11.8	25.2	2.0	73.8	70.6	78.0
L3-8b-FT	36.5	15.8	28.1	32.8	6.5	81.5	82.2	89.4
DPO (L3-8b)	36.8	16.0	28.6	33.5	6.9	83.6	84.7	91.3
KTO (L3-8b)	28.8	11.1	21.7	28.9	2.7	77.9	81.4	89.3
GPT-3.5 (1-shot)	28.3	11.3	22.0	23.7	2.3	77.1	82.5	80.4
GPT-4 (1-shot)	31.3	12.1	22.3	32.8	3.4	79.5	83.1	90.1
Gemini (1-shot)	29.3	14.1	23.4	23.9	3.4	80.6	82.7	90.0

Table 21: Traditional, UniEval and GPT-4 Critique evaluations on News Dataset, (The best outcomes are marked in **Bold**) Additional Results of Table 6

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1		
L2-7b-FT	28.1	8.6	20.4	24.8	2.7	81.3	66.6	77.8
Mst-7b-FT	25.7	6.9	17.7	17.6	1.7	77.4	65.1	77.0
L3-8b-FT	31.9	9.7	22.6	23.6	3.1	69.0	67.1	79.5
LF	16.0	4.8	11.5	16.7	1.7	67.4	55.2	68.7
T5-FT	20.0	5.0	16.0	15.0	2.0	71.0	63.2	75.9
BART	19.1	3.3	8.3	19.3	2.4	71.9	64.7	70.2
Pegasus	17.8	4.1	9.2	16.9	1.7	70.8	62.5	60.9
Falcon	17.0	3.0	11.0	15.0	0.8	74.0	55.2	38.4
TLDR	19.5	5.6	14.8	11.6	0.6	77.6	62.9	72.3

Table 22: Traditional, UniEval and GPT-4 Critique evaluations on CSPubSum Dataset - comparison of baselines with SFT, Additional Results of Table 5

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1		
L2-7b-VA	14.3	3.3	9.7	10.9	0.9	45.9	58.6	79.0
L2-7b-FT	28.1	8.6	20.4	24.8	2.7	81.3	64.8	77.8
DPO (L2-7b)	30.4	9.2	26.8	25.1	2.8	81.6	65.1	83.1
KTO (L2-7b)	27.7	8.1	19.0	22.9	2.5	79.7	64.6	78.6
Mst-7b-VA	13.8	2.9	9.1	10.6	0.7	46.4	59.2	79.7
Mst-7b-FT	25.7	6.9	17.7	17.6	1.7	77.4	64.0	77.0
DPO (Mst-7b)	29.9	7.6	20.1	25.8	2.9	81.4	65.4	85.0
KTO (Mst-7b)	27.9	7.4	18.2	25.1	2.6	80.0	64.8	78.1
L3-8b-VA	15.5	3.8	10.2	12.8	1.0	47.2	60.1	73.6
L3-8b-FT	31.9	9.7	22.6	23.6	3.1	81.8	67.1	79.5
DPO (L3-8b)	32.3	9.9	23.0	24.9	3.5	82.3	67.8	89.0
KTO (L3-8b)	29.1	5.4	13.2	15.1	1.6	75.0	64.6	81.0
GPT-3.5 (1-shot)	27.5	6.9	17.9	19.9	2.6	79.5	65.2	76.8
GPT-4 (1-shot)	28.7	7.4	17.9	23.3	3.2	80.8	66.8	82.1
Gemini (1-shot)	26.6	5.7	17.2	18.0	1.6	78.5	67.1	71.5

Table 23: Traditional, UniEval and GPT-4 Critique evaluations on CSPubSum Dataset, (The best outcomes are marked in **Bold**) Additional Results of Table 6

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1	Rel	
L2-7b-FT	29.3	13.6	26.2	22.8	5.9	70.0	67.2	51.4
Mst-7b-FT	25.5	14.5	21.8	20.3	8.4	78.8	66.9	52.8
L3-8b-FT	42.4	26.0	38.2	38.2	16.6	81.3	69.2	64.8
LF	16.2	4.6	12.0	20.3	1.0	70.0	63.2	43.5
T5-FT	25.2	8.2	21.1	19.5	7.1	68.7	65.2	54.6
BART	23.9	8.5	17.6	27.5	3.3	74.6	69.2	41.3
Pegasus	19.8	5.5	14.2	21.9	1.9	71.9	64.2	47.8
Falcon	17.0	5.0	12.0	22.0	1.2	71.0	63.4	45.8
TLDR	25.8	9.8	21.4	20.3	3.14	74.4	63.9	61.0

Table 24: Traditional, UniEval and GPT-4 Critique evaluations on Wikipedia Dataset - comparison of baselines with SFT, Additional Results of Table 5

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1	Rel	
L2-7b-VA	17.0	4.1	12.7	18.3	1.2	67.3	62.1	42.4
L2-7b-FT	29.3	13.6	26.2	22.8	5.9	70.0	67.2	51.4
DPO (L2-7b)	38.6	23.5	35.5	30.9	15.7	80.1	67.5	55.6
KTO (L2-7b)	31.5	18.6	27.6	29.8	13.4	81.1	66.1	55.6
Mst-7b-VA	18.9	4.9	13.7	20.4	1.4	70.2	63.4	36.4
Mst-7b-FT	25.5	14.5	21.8	20.3	8.4	78.8	66.9	52.8
DPO (Mst-7b)	37.2	20.8	28.1	27.1	13.9	82.9	67.1	65.2
KTO (Mst-7b)	34.1	19.9	22.7	23.1	11.7	78.6	66.2	63.0
L3-8b-VA	20.3	6.5	15.0	23.6	2.2	71.7	65.7	52.1
L3-8b-FT	42.4	26.0	38.2	38.2	16.6	81.3	69.2	64.8
DPO (L3-8b)	42.5	26.3	38.4	38.7	16.7	83.4	67.3	67.3
KTO (L3-8b)	32.0	14.9	22.4	29.1	11.5	82.0	65.2	60.0
GPT-3.5 (1-shot)	29.8	11.5	23.6	22.3	4.2	77.7	68.3	60.2
GPT-4 (1-shot)	34.6	13.8	26.1	30.8	7.1	80.4	70.2	73.5
Gemini (1-shot)	29.7	15.3	26.1	28.9	5.1	79.0	68.0	61.0

Table 25: Traditional, UniEval and GPT-4 Critique evaluations on Wikipedia Dataset, (The best outcomes are marked in **Bold**) Additional Results of Table 6

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1	Rel	
L2-7b-FT	29.9	7.1	16.5	14.9	2.1	77.2	74.7	80.1
Mst-7b-FT	36.6	8.9	17.0	18.1	2.2	79.2	75.2	80.5
L3-8b-FT	28.3	6.9	14.7	13.7	1.6	76.5	76.7	81.2
LF	13.2	2.2	5.5	7.9	0.5	61.5	69.7	68.7
T5-FT	18.3	2.3	9.6	6.6	1.1	67.0	70.7	71.4
BART	16.8	1.5	3.2	11.4	1.5	68.9	73.7	61.0
Pegasus	15.5	1.1	3.5	10.2	1.2	68.8	69.7	52.9
Falcon	15.2	1.6	6.2	7.0	0.6	71.0	68.7	39.6
TLDR	17.1	2.1	8.9	8.6	0.4	73.6	69.7	67.3

Table 26: Traditional, UniEval and GPT-4 Critique evaluations on Research Presentation Dataset - comparison of baselines with SFT, (The best outcomes are marked in **Bold**) Additional Results of Table 5

Model	Traditional Metrics						UniEval	GPT-4 Critique
	R1	R2	RL	Mt	Bl	BeF1		
L2-7b-VA	12.5	3.1	9.6	8.1	0.7	46.3	68.5	78.5
L2-7b-FT	29.9	7.1	16.5	14.9	2.1	77.2	74.7	80.1
DPO (L2-7b)	30.5	6.9	16.6	15.4	1.9	77.5	75.1	85.2
KTO (L2-7b)	28.5	6.6	14.9	14.5	2.0	75.3	74.4	83.6
Mst-7b-VA	13.6	2.6	9.0	10.2	0.6	47.2	69.3	79.7
Mst-7b-FT	36.6	8.9	17.0	18.1	2.2	79.2	75.2	80.5
DPO (Mst-7b)	37.3	8.8	17.5	18.6	2.6	79.3	75.8	85.6
KTO (Mst-7b)	34.8	7.3	16.5	18.0	2.2	78.5	74.9	82.1
L3-8b-VA	15.6	3.2	10.1	12.8	1.0	48.5	70.2	73.6
L3-8b-FT	28.3	6.9	14.7	13.7	1.6	76.5	76.7	81.2
DPO (L3-8b)	38.2	8.9	18.7	19.9	3.4	79.6	76.8	88.8
KTO (L3-8b)	28.1	6.3	14.6	14.2	1.4	77.3	74.2	86.4
GPT-3.5 (1-shot)	26.1	6.5	13.6	14.1	1.3	75.3	74.8	78.3
GPT-4 (1-shot)	32.3	7.1	16.8	17.9	2.1	77.2	75.2	81.6
Gemini (1-shot)	27.3	6.9	15.1	16.4	1.6	76.4	75.8	80.5

Table 27: Traditional, UniEval and GPT-4 Critique evaluations on Research Presentation Dataset, (The best outcomes are marked in **Bold**) Additional Results of Table 6

Dataset	Model	Rat	
		Model	Rat
News	DPO (L2-7b)	91.6	
News	DPO (Mst-7b)	94.6	
News	DPO (L3-8b)	93.5	
CSPubSum	DPO (L2-7b)	85.3	
CSPubSum	DPO (Mst-7b)	84.6	
CSPubSum	DPO (L3-8b)	87.1	
Wikipedia	DPO (L2-7b)	58.3	
Wikipedia	DPO (Mst-7b)	69.2	
Wikipedia	DPO (L3-8b)	71.4	
Research Presentation	DPO (L2-7b)	83.9	
Research Presentation	DPO (Mst-7b)	84.2	
Research Presentation	DPO (L3-8b)	87.3	

Table 28: Gemini Critique on DPO variants

Dataset	Model	Traditional Metrics						UniEval	GPT-4 Critique
		R1	R2	RL	Mt	Bl	BeF1		
DPO using Vanilla Model									
News Corpus	DPO (L3-8b)	16.2	7.1	16.3	26.3	2.6	74.9	74.0	82.3
CSPubSum	DPO (L3-8b)	17.5	4.6	11.1	13.3	1.7	52.6	57.3	75.1
Wikipedia	DPO (L3-8b)	23.6	8.2	16.7	25.2	3.8	73.9	60.6	57.1
Research Presentation	DPO (L3-8b)	16.8	6.9	14.8	22.5	3.1	63.1	61.4	68.4
DPO using SFT Model on full data									
News Corpus	DPO (L3-8b)	18.6	8.6	19.3	27.3	2.8	75.1	75.2	85.2
CSPubSum	DPO (L3-8b)	18.4	6.5	15.1	22.3	2.7	56.8	56.6	76.9
Wikipedia	DPO (L3-8b)	23.8	9.3	18.1	25.7	4.6	73.1	59.2	59.3
Research Presentation	DPO (L3-8b)	17.5	7.2	16.3	24.5	3.7	64.5	58.2	69.4
DPO using 50% data on SFT model and 50% on DPO									
News Corpus	DPO (L3-8b)	36.8	16.0	28.6	33.5	6.9	83.6	84.7	91.3
CSPubSum	DPO (L3-8b)	32.3	9.9	23.0	24.9	3.5	82.3	67.8	89.0
Wikipedia	DPO (L3-8b)	42.5	26.3	38.4	38.7	16.7	83.4	67.3	67.3
Research Presentation	DPO (L3-8b)	38.2	8.9	18.7	19.9	3.4	79.6	76.8	88.8

Table 29: DPO variant - DPO using vanilla model, SFT using full data and SFT using 50% of data

Training set	Test set	Traditional Metrics						GPT-4 Critique
		R1	R2	RL	Mt	Bl	BeF1	
News	CSPubSum	28.7	9.4	21.3	23.2	3.0	80.2	71.5
News	Wikipedia	34.2	21.6	32.7	33.2	11.2	76.2	63.1
CSPubSum	News	28.3	9.3	24.2	27.1	3.4	79.3	75.0
CSPubSum	Wikipedia	35.1	19.7	28.9	29.1	10.3	75.4	65.8
Wikipedia	News	30.1	11.3	25.3	26.1	4.9	81.5	74.0
Wikipedia	CSPubSum	28.3	9.2	20.4	22.1	3.1	78.3	72.0

Table 30: Cross Dataset Validation Assessment on DPO (Llama3-8b) model