

# Not too long do read: Evaluating LLM-generated extreme scientific summaries

Anonymous ACL submission

## Abstract

High-quality scientific extreme summary (TLDR) facilitates effective science communication. How do large language models (LLMs) perform in generating them? How are LLM-generated summaries different from those written by human experts? However, the lack of a comprehensive, high-quality scientific TLDR dataset hinders both the development and evaluation of LLMs' summarization ability. To address these, we propose a novel dataset, BiomedTLDR, containing a large sample of researcher-authored summaries from scientific papers, which leverages the common practice of including authors' comments alongside bibliography items. We then test popular open-weight LLMs for generating TLDRs based on abstracts. Our analysis reveals that, although some of them successfully produce humanoid summaries, LLMs generally exhibit a greater affinity for the original text's lexical choices and rhetorical structures, hence tend to be more extractive rather than abstractive in general, compared to humans. Our code and datasets are available at

## 1 Introduction

We, as researchers, are often called upon to condense our work into concise and informative summaries—whether it is a one-sentence teaser or a take-home message—across diverse contexts such as journal submissions and academic presentations. More broadly, high-quality extreme summaries, or TLDRs, facilitate the communication of scientific findings to diverse audiences, from domain experts to the lay public, which is essential for advancing research and promoting public engagement (Jamieson et al., 2017).

Recently, Large Language Models (LLMs) have begun to transform this landscape, promising to automate the generation of these TLDRs (Binz et al., 2025). However, the efficacy of LLMs in

The screenshot shows a web browser interface with three tabs: 'Sections', 'Figures', and 'References'. The 'References' tab is active, displaying a list of references. The first reference is highlighted in bold text: '19. Riquet, J. et al. Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. *Proc. Natl Acad. Sci. USA* **96**, 9252–9257 (1999). **This paper shows how the identity-by-descent approach can be applied to high-resolution mapping of quantitative trait loci in farm animals by using extensive pedigree information.**' Below the reference are links for 'Article', 'CAS', 'PubMed', 'PubMed Central', and 'Google Scholar'. Below the links is a section labeled 'B Abstract' which contains the abstract text: 'We previously mapped a quantitative trait locus (QTL) affecting milk production to bovine chromosome 14. ... In a second phase, we genotyped the seven selected sires for the newly developed high-density marker map and searched for a shared haplotype flanking an hypothetical, identical-by-descent QTL allele with large substitution effect. ... , providing additional support in favor of the location of the QTL within the corresponding interval.' Below the abstract is a 'Summary' section: 'This paper shows how the identity-by-descent approach can be applied to high-resolution mapping of quantitative trait loci in farm animals by using extensive pedigree information.'

Figure 1: (A) An example annotated bibliography entry cited by Andersson (2001). Bold text at the end is the annotation from the author, which is used as the summary in this research. (B) The corresponding abstract-summary pair to (A). The abstract is from Riquet et al. (1999).

this domain is bottlenecked by the scarcity of high-quality scientific summary corpora. Predominant large-scale datasets often rely on extracting citation contexts from the “Related Work” sections of citing papers. Such supervision signals are inherently fragmented, as they typically capture only specific facets of the cited work tailored to the citing author’s narrative, failing to provide a holistic overview. Furthermore, their reliance on full-text extraction introduces significant availability barriers

052 ers, as well as unnatural and imbalanced discipline  
053 distribution.

054 To overcome these obstacles, we introduce  
055 BiomedTLDR, a large-scale corpus derived from  
056 annotated bibliographies, which consists of cita-  
057 tions accompanied by a few brief, descriptive sen-  
058 tences in which the author explicitly summarizes  
059 the referenced work, as shown in Fig. 1A. This  
060 structural format provides a self-contained sum-  
061 mary of the cited paper’s core contributions, avoid-  
062 ing fragmentation and narrative bias. By systemati-  
063 cally collecting these entries, we construct a dataset  
064 of over 35 thousand abstract-summary pairs, with  
065 most of the source abstracts and target summaries  
066 originating from highly cited papers published in  
067 top-tier biology or medical journals.

068 Using BiomedTLDR, we then conduct a multi-  
069 dimensional evaluation of state-of-the-art LLMs,  
070 uncovering a distinct behavioral divergence be-  
071 tween LLM and human summarization. Our analy-  
072 sis reveals that while exhibiting a stronger extrac-  
073 tive tendency, some of LLMs have reached human  
074 parity in identifying salient information, but they  
075 generally differ fundamentally in how present and  
076 rearrange it.

077 Our contributions are summarized as follows:

- 078 1. We propose a novel data collection paradigm  
079 based on annotated bibliographies and release  
080 BiomedTLDR, a dataset comprising high-  
081 quality researcher-authored TLDRs, aiming  
082 to establish a new benchmark for extreme sci-  
083 entific summarization.
- 084 2. We provide a comprehensive evaluation frame-  
085 work that systematically quantifies the writ-  
086 ing styles and tendencies of state-of-the-art  
087 LLMs.

## 088 2 Related Work

089 Early research in automatic summarization primar-  
090 ily focused on newswire domains (e.g., CNN/Daily-  
091 Mail (Hermann et al., 2015)). However, the expo-  
092 nential growth of scientific literature has created a  
093 pressing need for automated tools to mitigate infor-  
094 mation overload and enable rapid screening. These  
095 tools have proven highly effective in real-world ap-  
096 plications, such as Semantic Scholar, which helps  
097 researchers speed up the process of targeting rele-  
098 vant papers by distilling complex research into a  
099 single, highly informative sentence. The scientific  
100 domain presents unique challenges due to its spe-  
101 cialized terminology, lengthy documents, and the

102 requirement for high-level domain expertise, which  
103 necessitates the construction of specialized scien-  
104 tific TLDR corpora distinct from traditional sum-  
105 marization datasets, for both the development and  
106 the evaluation of automatic summarization models.

107 Völske et al. (2017) pioneered this direction  
108 by mining user-generated summaries from Reddit,  
109 demonstrating the feasibility of learning extreme  
110 summarization from large-scale social media data.  
111 However, the informal and noisy nature of Reddit  
112 posts limits their direct applicability to rigorous  
113 scientific communication. Transposing this task  
114 to the scientific domain, Cachola et al. (2020) in-  
115 troduced SciTLDR, a benchmark dataset derived  
116 from OpenReview. By combining author-written  
117 summaries with those curated by graduate students,  
118 SciTLDR established a high-quality gold standard.  
119 Nevertheless, its reliance on manual curation re-  
120 sults in a relatively small scale and limited domain  
121 coverage.

122 To address the scalability bottleneck of manual  
123 annotation, recent efforts have turned to automated  
124 extraction methods. For example, Mao et al. (2022)  
125 proposed CiteSum, which extracts citation texts  
126 from the “Related Work” sections of papers as prox-  
127 ies for summaries. While this approach enables  
128 the construction of a large-scale dataset from the  
129 S2ORC corpus, it inherently relies on access to full-  
130 text articles, subjecting the dataset to copyright re-  
131 strictions and potential disciplinary biases (80.04%  
132 are computer science papers and only 2.13% are  
133 medical papers). Moreover, citation texts often em-  
134 phasize certain aspects pertinent to the citing work  
135 rather than providing a holistic summary of the  
136 cited paper.

137 Beyond single-document extreme summariza-  
138 tion, the research landscape has expanded into  
139 diverse input scenarios and objectives. Lu et al.  
140 (2020) introduced Multi-XScience with emphasis  
141 on multi-document summarization, using a whole  
142 paragraph from a paper’s “Related Work” section  
143 as the target text, and uses that paper’s abstract,  
144 together with abstracts of all papers cited in that  
145 paragraph, as the source text. Takeshita et al. (2022)  
146 and Atri et al. (2023) extended the boundaries of  
147 the task to cross-lingual and multimodal settings,  
148 respectively, incorporating multiple language trans-  
149 lations and audio-visual conference presentation  
150 data. A parallel line of research (Goldsack et al.,  
151 2022) and (Guo et al., 2024) focuses on focused  
152 “lay summarization”, aiming to make science acces-  
153 sible to the general public through expert-written

Table 1: Statistics of existing scientific TLDR datasets, where ours is in bold. Average values are reported for words and sentences, whereas the median is used for citation counts.

Dataset	Domain	# pairs	# words (source/target/ratio)	# sentences (source/target/ratio)	# citations (source/target)
SciTLDR	OpenReview	5,412	179.92/22.15/8.12	7.46/1.04/7.17	- / -
Multi-XScience	arXiv	40,528	772.69/120.50/6.41	30.72/4.85/6.33	192/5
CiteSum	S2ORC	92,946	283.69/25.16/11.28	11.34/1.15/9.86	89/ -
<b>BiomedTLDR</b>	Springer	35,626	191.58/25.29/7.58	7.37/1.14/6.46	308/435

simplifications and retrieval-augmented LLMs.

### 3 Our Dataset

We introduce a dataset comprising 35,626 paper abstracts paired with researcher-authored summaries. This resource aims to enrich academic-targeted training corpora and facilitate evaluations of LLMs’ summarization ability by establishing a human baseline.

To construct the dataset, we systematically crawl web pages for over 1.2 million academic papers from Springer’s websites, with an example shown in Fig. 1A. Then we parse retrieved HTML files to extract bold annotations and the DOI of corresponding citable items. For most pages, a Google Scholar link is provided for each reference, from which we obtain the embedded DOI; when the link is absent, we use CrossRef’s API to query the DOI based on the bibliography text. After collecting DOI-summary pairs, we use DOIs to match the abstracts of corresponding academic papers (as shown in Fig. 1B), resulting in a dataset consisting of 35,626 abstract-summary pairs involving 29,356 papers. Some samples from our dataset are available in § B.2.

Table 1 provides a quantitative comparison of several existing scientific TLDR datasets. Our BiomedTLDR dataset distinguishes itself from them in several qualitative dimensions. First, our use of annotated bibliographies—commonly found in highly cited review papers authored by experienced researchers—ensures a more authoritative and standalone overview of the referenced work. Moreover, by relying solely on universally accessible abstracts and bibliography metadata, our approach circumvents the copyright restrictions and parsing noise associated with full-text extraction, thereby enhancing reproducibility and disciplinary coverage, as shown in § A.2.

### 4 Generating Summary using LLMs

To measure LLMs’ summarization ability and to examine how LLM-generated summaries differ from those written by domain experts, we set up a local Ollama server which hosts 5 open-weights LLMs, namely gemma3:27b (Team et al., 2025), deepseek-r1:70b (DeepSeek-AI et al., 2025), gpt-oss:12b (OpenAI et al., 2025), llama4:16x17b (lla), and qwen3:235b (Yang et al.). We employ a few-shot learning framework that guides LLMs through summarization (see § B.1). Specifically, 5 randomly selected and then fixed abstract-summary pairs are provided for LLMs, along with the word count of each summary, as the training set. The other 35,621 pairs serve as the test set. Then, we feed LLMs with abstracts and word counts of human-written summaries as length reference and ask them to generate summarizations recursively on the test set. Samples in both the training and test sets are identical for all LLMs. We set the context window size to 4096 and the temperature to 1 for all LLMs. We disabled DeepSeek-R1, gpt-oss, and Qwen3’s thinking mode due to the complexity of the processing speed required for this task. Every abstract in the test set was summarized individually, and we got 178,105 (35,621 × 5) LLM-generated summaries in total. Some samples of them can be checked in § B.2.

### 5 Evaluation

We present multi-dimensional comparisons between the extreme scientific summaries in our dataset, their respective abstracts, and LLM-generated summaries. These analyses include summary statistics (§ 5.1), readability (§ 5.2), abstractiveness (§ 5.3), rhetorical structure (§ 5.4), and reference-based evaluation (§ 5.5), which collectively aim to understand the differences between different types of summaries.

Table 2: Statistics for summaries by LLMs and human researchers, compared with original abstracts. Reported are mean  $\pm$  standard deviation.

Model	words	Length		NEs per 100 words	Training set copy rate (%)		
		sentences	words/sent.		= 0%	> 10%	Max
Gemma 3	29.06 $\pm$ 11.03	1.03 $\pm$ 2.31	28.31 $\pm$ 9.59	4.49 $\pm$ 4.64	91.25	0.28	27.78
DeepSeek-R1	60.46 $\pm$ 158.34	2.31 $\pm$ 7.34	26.19 $\pm$ 8.30	4.13 $\pm$ 3.71	90.83	0.19	100.00
Gpt-oss	62.35 $\pm$ 22.84	1.64 $\pm$ 0.93	37.94 $\pm$ 15.51	<b>6.30 <math>\pm</math> 4.60</b>	<b>96.16</b>	<b>0.01</b>	<b>11.48</b>
Llama 4	43.11 $\pm$ 764.99	1.30 $\pm$ 3.32	33.24 $\pm$ 628.57	3.46 $\pm$ 4.28	86.63	1.05	46.15
Qwen3	153.29 $\pm$ 1799.10	6.22 $\pm$ 91.82	24.64 $\pm$ 140.67	4.90 $\pm$ 4.38	94.88	0.03	95.24
Human	25.29 $\pm$ 11.43	1.14 $\pm$ .41	22.22 $\pm$ 8.83	5.56 $\pm$ 5.93	-	-	-
Abstract	191.58 $\pm$ 82.52	7.37 $\pm$ 3.18	26.01 $\pm$ 11.71	5.11 $\pm$ 3.83	-	-	-

## 5.1 Descriptive Statistics

Table 2 presents some summary statistics of LLM-generated summaries, human-written summaries, as well as the original abstracts for reference. All LLMs, except Gemma 3, produce summaries whose lengths significantly exceed the intended length, despite providing a reference word count during training. Furthermore, LLMs tend to generate repetitive or verbose content to varying degrees (see § B.3), reflected in the extremely large standard deviations of Qwen3 and Llama 4’s summary lengths.

Identifying named entities using the pretrained `en_core_web_sm` model from spaCy (Honnibal et al., 2020), we find that gpt-oss exhibits the most distinguishable density of information content in its generated summary, reflecting a higher specificity. The other LLM-generated summaries show lower entity density compared to human-written ones in general.

We evaluate the extent LLM-generated summaries copy phrases from the training set examples, assessing possible overfitting or memorization. Specifically, we define the training set copy rate of a summary as the percentage of its 3-grams that overlap with those in the 5 training samples. Table 2 reports 3 statistics: the proportion of summaries with a 0% copy rate, the proportion with a copy rate exceeding 10%, and the maximum copy rate observed across all summaries. The results show that all LLMs exhibit a varying extent of copying from the training set.

## 5.2 Readability

Ideally, an effective TLDR should lower the cognitive barrier for readers, transforming complex technical content into plain language that is easier to digest. To assess the linguistic complexity

Table 3: Comparison of mean readability scores across different models and human baselines. For all metrics, lower values denote better readability.

Model	FKG $\downarrow$	GF $\downarrow$	SI $\downarrow$	ARI $\downarrow$	CLI $\downarrow$	LWF $\downarrow$	DCRS $\downarrow$
Gemma 3	19.20	23.46	19.52	21.68	19.55	21.12	14.48
DeepSeek-R1	19.69	23.85	19.76	22.47	20.80	21.02	14.94
Gpt-oss	23.79	27.82	22.61	28.45	21.55	29.90	15.77
Llama 4	18.13	21.80	18.25	20.22	17.58	19.86	13.54
Qwen3	22.94	27.27	22.05	26.52	23.15	25.70	15.94
Human	<b>15.25</b>	<b>18.79</b>	<b>16.13</b>	<b>16.58</b>	<b>16.08</b>	<b>15.70</b>	<b>13.05</b>
Abstract	16.17	19.75	17.18	17.91	16.61	16.38	13.65

and accessibility of the generated summaries, we employ 7 established readability metrics: Flesch-Kincaid Grade (FKG), Gunning Fog (GF), SMOG Index (SI), Automated Readability Index (ARI), Coleman-Liau Index (CLI), Linsear Write Formula (LWF), and Dale-Chall Readability Score (DCRS). These metrics estimate the US grade level required to comprehend a text by analyzing features such as sentence length, syllable count, and the proportion of complex or polysyllabic words. For all metrics, lower scores indicate higher readability and greater accessibility to a broader audience.

Table 3 reveals a distinct divide regarding readability between LLM-generated summaries and human-written ones. Surprisingly, we find that all LLM-generated summaries exhibit significantly higher scores across all metrics compared to not only human-written summaries but also the original abstracts. Among LLMs, Llama 4 proves to be the most readable model, despite its tendency to generate longer summaries. On the other hand, summaries from human researchers are consistently more readable than the corresponding abstracts, despite the small scale of the difference. We perform two-sided Mann-Whitney U tests and find that the differences between all summaries and the original abstract are statistically significant (all  $p$ -values

Table 4: Proportion of novel  $n$ -grams in LLM-generated summaries compared to human-written ones. Values are reported as mean  $\pm$  standard deviation. Higher values indicate more original vocabulary usage.

Model	% novel $n$ -grams		
	$n = 1$	$n = 2$	$n = 3$
Gemma 3	35.44 $\pm$ 11.35	77.12 $\pm$ 12.49	90.20 $\pm$ 9.90
DeepSeek-R1	33.48 $\pm$ 9.59	73.81 $\pm$ 10.13	87.90 $\pm$ 8.20
Gpt-oss	39.22 $\pm$ 9.36	80.94 $\pm$ 7.95	92.70 $\pm$ 5.64
Llama 4	39.85 $\pm$ 13.48	81.63 $\pm$ 13.23	92.67 $\pm$ 10.21
Qwen3	37.66 $\pm$ 10.39	80.74 $\pm$ 9.33	92.91 $\pm$ 6.78
<b>Human</b>	<b>46.13 <math>\pm</math> 14.23</b>	<b>86.33 <math>\pm</math> 13.16</b>	<b>95.09 <math>\pm</math> 9.86</b>

are smaller than 0.001).

We hypothesize that this discrepancy stems from the tendency of LLMs to excessively rely on high-density domain-specific jargon. In contrast, human researchers likely exercise greater audience awareness during the summarization process; they may intentionally adopt plain language to ensure their summaries remain accessible to reviewers or readers from outside the specific domain (Devaraj et al., 2021). We attribute this to its lower density of named entities (as shown in Table 2), a factor that inversely correlates with the poor readability of gpt-oss.

### 5.3 Abstractiveness

While extractive summarization primarily selects and rearranges existing sentences, an effective TLDR often requires abstractive capabilities to paraphrase complex concepts and synthesize information concisely. To evaluate this, following previous studies (Sharma et al., 2019; See et al., 2017), we calculate the proportion of  $n$ -grams in a summary that do not appear in the original source abstract (i.e., novel  $n$ -grams) as a measure of abstractiveness. A higher value indicates that the summary employs more original phrasing rather than simply copying chunks of text from the abstract, i.e., to be more abstractive rather than extractive, reflecting a higher level of lexical novelty and creativity exhibited by the writer.

Table 4 illustrates that human writers consistently exhibit a greater tendency to introduce new vocabulary not present in the source text compared to all LLMs, suggesting a significantly superior abstractiveness. Among the models, Llama 4 and gpt-oss demonstrate the highest level of abstractiveness. By contrast, smaller or more constrained models like Gemma 3 and DeepSeek-R1 show lower novelty scores, indicating a stronger reliance on ex-

tractive copying from the source abstracts. These results suggest that while advanced LLMs are becoming increasingly capable of paraphrasing, human researchers still maintain a distinct edge in synthesizing information with original language.

### 5.4 Rhetorical Structure

Understanding the rhetorical structure of summaries—how information is organized across functional roles such as background, objective, methods, results, and conclusions—provides insight into how writers prioritize and sequence content. This is particularly relevant for TLDR, where space constraints force selective emphasis on certain rhetorical elements.

To analyze rhetorical structure, we employ a DeepSeek-R1-based classifier to assign each sentence in the original abstracts to one of 5 rhetorical roles: BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS (see § C for details). For each word  $w$  appearing at a normalized position  $p$  in a summary, we compute the probability of its rhetorical role based on word occurrence statistics from the corresponding abstract. Specifically, let  $c_{w,r}$  denote the count of word  $w$  in sentences labeled with role  $r$ , and  $N_r$  the total word count in role  $r$ . We define the normalized probability as:

$$Pr(r | w) = \frac{c_{w,r}/N_r}{\sum_{r'} c_{w,r'}/N_{r'}}. \quad (1)$$

For words in the summary that do not appear in the corresponding abstract, we assign them to a special NOVEL category with  $Pr(\text{NOVEL} | w) = 1$ , indicating content introduced beyond the source abstract. We then aggregate these probabilities across all summaries by binning word positions into 20 equal intervals.

Figure 2 visualizes the resulting rhetorical role distributions, where darker colors indicate higher probabilities that words at a given position originate from a specific rhetorical section of the abstract. For example, if the BACKGROUND row shows dark blue at the beginning, it suggests that words in the opening of summaries are predominantly drawn from background-related sentences in the source abstract.

Several patterns emerge from this analysis. First, all text sources—LLM-generated summaries, human summaries, and original abstracts—exhibit a similar overall progression from BACKGROUND toward CONCLUSIONS, reflecting the conventional

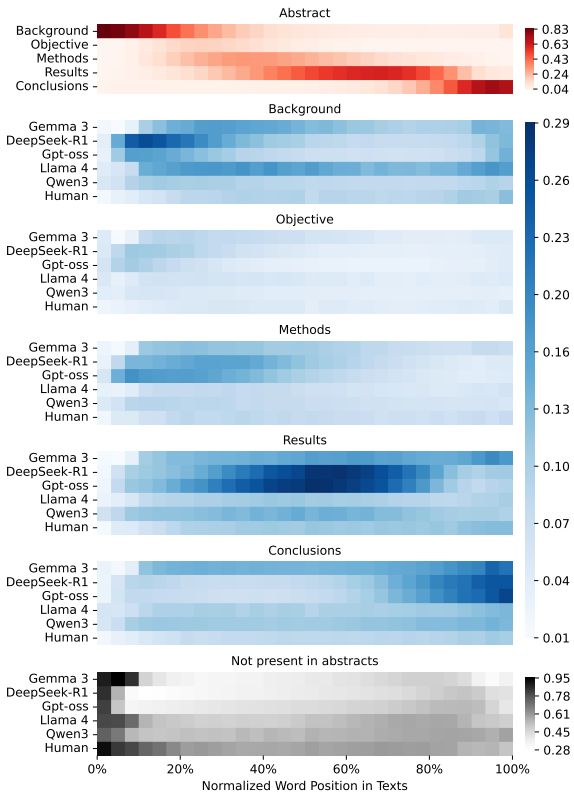


Figure 2: Word rhetorical role distribution across different types of summaries, and original abstracts as the reference. The  $x$ -axis represents normalized word position in texts, and color intensity indicates the average probability of belonging to each rhetorical role, based on word co-occurrence statistics against the original abstract.

structure of scientific discourse, while DeepSeek-R1 and gpt-oss display a closer mirroring of the rhetorical structure, when compared to the original abstracts. Second, across all summary types, words are predominantly drawn from the RESULTS sections of the original abstracts, with OBJECTIVE contributing the least, indicating a shared preference for reporting findings over restating research goals. Third, we observe that all types of summaries introduce a substantial number of novel words—words not present in the original abstracts—at the beginning positions. Upon inspection, these are predominantly meta-discourse markers such as “study,” “demonstrates,” “identifies,” and “shows” (see Figure 8), which serve to frame the research in line with research norms. Fourth, compared to the relatively concentrated distributions of BACKGROUND and CONCLUSIONS in original abstracts, summaries tend to spread these roles more broadly across positions, particularly for Gemma 3 and Llama 4, and human, indicating a potentially more

flexible functionality structure.

## 5.5 Reference-based Evaluation

We benchmark LLMs against human experts to assess summarization quality. Following the widely used evaluation methodology, we compare generated summaries to reference ones using lexical overlap-based measures, including BLEU (Papineni et al., 2002),  $F_1$  ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), as well as semantic similarity-based measures that capture deeper semantic correspondence, including BERTScore (Zhang et al., 2019) and MoverScore (Zhao et al., 2019).

Using researcher-written summaries as references, we find that Gemma 3 and DeepSeek-R1 perform best on semantic similarity (MoverScore) and lexical overlap (METEOR), respectively (upper half of Table 5). These metrics-based results are best interpreted when positioning human-authored summaries as a counterpart for comparison. Specifically, we leverage the fact that papers in our dataset can receive multiple summaries (typically 2-10; see Figure 6). To establish a human performance baseline, we conduct a within-paper, cross-referenced evaluation of human-authored summaries with 10 iterations. In each iteration, we randomly select one summary as the reference and calculate metrics for the remaining summaries. The reported values represent averages across these iterations. The results indicate that human authors achieve the highest scores on BLEU, ROUGE, and BERTScore metrics (the sixth row in Table 5), suggesting that human-written summaries of the same paper exhibit stronger  $n$ -gram exact matches and semantic consistency than those generated by LLMs. However, the high scores on METEOR and MoverScore metrics indicate that models like Gemma 3 and DeepSeek-R1 are highly effective at identifying the same salient information as human experts, by successfully capturing the gist and core technical entities that humans prioritize.

A contrasting pattern emerges when evaluating against the original abstracts. Human-authored summaries show both the lowest overlap and similarity scores (lower half of Table 5), suggesting that human annotators tend to substantially reformulate content and introduce novel phrasing. This might also stem from human authors drawing on full texts rather than abstracts alone for summarization. Among LLMs, Qwen3 and gpt-oss attain the highest BLEU and ROUGE scores, respec-

Table 5: Evaluation of summarization performance across LLMs and human authors using two references: researcher-written summaries and original paper abstracts.

Model	Overlap-based			Similarity-based	
	BLEU	ROUGE(1/2/L)	METEOR	BERTScore	MoverScore
<b>Human-referenced performance</b>					
Gemma 3	.0290	.2634/.0584/.2021	.2216	.8817	<b>.5660</b>
DeepSeek-R1	.0179	.2375/.0514/.1703	<b>.2486</b>	.8715	.5562
Gpt-oss	.0131	.2233/.0443/.1549	.2255	.8592	.5489
Llama 4	.0188	.2458/.0508/.1865	.2090	.8771	.5593
Qwen3	.0046	.2272/.0465/.1665	.2129	.8699	.5578
Human	<b>.0450</b>	<b>.2734/.0675/.2163</b>	.2155	<b>.8824</b>	.5642
<b>Abstract-referenced performance</b>					
Gemma 3	.0006	.1931/.0752/.1381	.1004	.8637	.5524
DeepSeek-R1	.0226	.3279/.1434/.2306	<b>.1974</b>	<b>.8838</b>	<b>.5836</b>
Gpt-oss	.0167	<b>.3436/.1323/.2347</b>	.1788	.8751	.5817
Llama 4	.0035	.1878/.0616/.1263	.0942	.8565	.5438
Qwen3	<b>.0307</b>	.2495/.0911/.1659	.1345	.8694	.5681
Human	.0001	.1409/.0382/.1005	.0638	.8410	.5302

tively, indicating their different inclinations in word choice and the trade-off between precision and recall. While DeepSeek-R1 achieves both the highest METEOR and similarity scores, surpassing gpt-oss and Qwen3, indicating that DeepSeek-R1 maintains the closest lexical and semantic fidelity to the source abstract, even with fewer parameters.

These contrasting results presented in Table 5 highlight an important, well-recognized tension in summarization: extractive versus abstractive (Durmus et al., 2020). Human researchers appear to prioritize the latter, whereas LLMs tend to emphasize the former by exhibiting stronger adherence to the source text, which aligns with Table 4.

## 6 Discussion

Our comprehensive evaluations reveals distinct behavioral profiles across LLMs, suggesting the need to align model selection for scientific summarization with user priorities. For approximating human behaviors, Gemma 3 emerges as a balanced candidate, uniquely adhering to length constraints while achieving high semantic similarity to expert-written references. If precise, high-fidelity condensation is the goal, DeepSeek-R1 and gpt-oss maintain strict faithfulness to the original vocabulary and rhetorical structure, though via distinct mechanisms: DeepSeek-R1 adopts a conservative, extractive approach. Its superior METEOR score

evidences a precise alignment with researchers in terminology preferences and content selection, yet its low lexical novelty suggests a minor level of paraphrasing. In contrast, gpt-oss functions as a high-fidelity rewriter; it achieves the highest content recall (ROUGE) alongside substantial lexical novelty, effectively rephrasing the source text without sacrificing factual detail. Llama 4 presents a different trade-off, offering the highest readability scores suitable for non-expert audiences, as well as the highest lexical novelty, yet it suffers from significant hallucinations in length control. Finally, Qwen3 demonstrates strong lexical precision (high BLEU) yet its practical utility is compromised by significant instability and a tendency toward verbosity.

Regarding model size, an interesting observation is that larger models do not necessarily outperform smaller ones in summarization. On the contrary, lighter models such as Gemma 3, DeepSeek-R1, and gpt-oss demonstrate stronger capabilities and greater similarity to human summaries. Notably, deepseek-r1:70b—a distilled model of llama3:70b fine-tuned on reasoning data generated by deepseek-r1:671b—even surpasses the base model’s next generation llama4:16x17b, which has more parameters. This result underscores the value of model distillation and data-centric fine-tuning (DeepSeek-AI et al., 2025),

prompting us to rethink the potential of parameter-efficient models and move beyond the conventional scaling law that favors increasing model size.

Beyond model performance, this study validates a novel and scalable pipeline for constructing high-quality scientific summarization datasets. By leveraging annotated bibliographies, where experts manually curate and summarize the literature, we circumvent the high demand for domain knowledge in large-scale annotation. This methodology is inherently generalizable since it can be readily adapted to any academic archive or indexing service that maintains bibliographic metadata. Moreover, by relying solely on abstracts and citation metadata, which are typically open-access or fall under fair use, our approach mitigates copyright concerns and lowers barriers for future research in NLP.

## 7 Conclusion

In this work, we present BiomedTLDR, a scalable summary corpus that provides a comprehensive disciplinary coverage aligned with the actual publication landscape, offering a high-quality, academic context tailored alternative to existing TLDR datasets. Using BiomedTLDR, we conduct a multi-dimensional evaluation of current LLMs and highlight that LLMs remain more extractive compared to human authors. Among the models, DeepSeek-R1 archives the highest fidelity to source texts, gpt-oss balances faithfulness with lexical novelty, and Gemma 3 uniquely prioritizes semantic alignment, mimicking the synthesis style of human researchers without relying on extensive lexical overlap.

## Limitations

We acknowledge several limitations of our work. First, although the dataset is high quality, its source—annotated bibliographies—is not a perfect proxy for TLDR. Annotation styles and intentions can vary, reflecting personal perspectives or critiques rather than purely objective summaries. Nevertheless, the dataset’s large scale enables a unique perspective into how LLMs perform extreme summarization relative to human authors.

Second, while LLMs were restricted to using only paper abstracts as input, human annotators may likely draw on abstracts and full texts, which could partly explain the observed low abstract-referenced similarity. Future work could explore whether providing LLMs with full-text would lead to further improvements in summary quality.

Third, although we explicitly constrained LLM output length in the prompt, most generated summaries were significantly longer than those by human researchers. This calls for dedicated experiments to rigorously decouple the relationship between summary quality and length, despite evaluation metrics (e.g., ROUGE, METEOR) claiming robustness to length variation.

Fourth, our dataset is not fully representative of the entire scientific literature in terms of citations, as all reported performance is based on highly cited works. Therefore, evaluating LLMs summarization on the general population of papers remains as an important direction for future research.

Finally, we evaluated only open-source models and selected their smaller variants, due to computational and cost constraints. Consequently, our study excludes proprietary models or the largest architectures (e.g., DeepSeek-r1:671b and llama4:128x17b), limiting a comprehensive performance comparison across the full spectrum of current LLM capabilities.

## Ethical considerations

Our experiments reveal significant instability in LLM outputs, posing challenges for their practical adoption in scientific workflows, as evidenced by the extreme standard deviations in summary length and the non-trivial rates of training set memorization. When coupled with our finding that model-generated summaries can be less readable than the source abstracts, these inconsistencies suggest that LLMs cannot yet be fully automated as unsupervised research assistants. Instead, they require rigorous human oversight and post-hoc verification to ensure the integrity and accessibility of scientific communication.

## Acknowledgments

## References

- The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.
- Leif Andersson. 2001. Genetic dissection of phenotypic diversity in farm animals. *Nature Reviews Genetics*, 2(2):130–138.
- Yash Kumar Atri, Vikram Goyal, and Tanmoy Chakraborty. 2023. Fusing multimodal signals on hyper-complex space for extreme abstractive text summarization (tl;dr) of scientific contents. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’23*,

608	page 3724–3736, New York, NY, USA. Association	Karl Moritz Hermann, Tomas Kocisky, Edward Grefen-	664
609	for Computing Machinery.	stette, Lasse Espeholt, Will Kay, Mustafa Suleyman,	665
610	Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An	and Phil Blunsom. 2015. Teaching Machines to Read	666
611	automatic metric for mt evaluation with improved cor-	and Comprehend. In <i>Advances in Neural Information</i>	667
612	relation with human judgments. In <i>Proceedings of</i>	<i>Processing Systems</i> , volume 28. Curran Associates,	668
613	<i>the acl workshop on intrinsic and extrinsic evaluation</i>	Inc.	669
614	<i>measures for machine translation and/or summariza-</i>	Matthew Honnibal, Ines Montani, Sofie Van Lan-	670
615	<i>tion</i> , pages 65–72.	degheem, and Adriane Boyd. 2020. spaCy: Industrial-	671
616	Marcel Binz, Stephan Alaniz, Adina Roskies, Bal-	strength Natural Language Processing in Python.	672
617	azs Aczel, Carl T Bergstrom, Colin Allen, Daniel	<i>Zenodo</i> .	673
618	Schad, Dirk Wulff, Jevin D West, Qiong Zhang,	Kathleen Hall Jamieson, Dan Kahan, and Dietram A	674
619	Richard M. Shiffrin, Samuel J. Gershman, Vencislav	Scheufele. 2017. <i>The Oxford handbook of the science</i>	675
620	Popov, Emily M. Bender, Marco Marelli, Matthew M.	<i>of science communication</i> . Oxford University Press.	676
621	Botvinick, Zeynep Akata, and Eric Schulz. 2025.	Chin-Yew Lin. 2004. Rouge: A package for automatic	677
622	How should the advancement of large language mod-	evaluation of summaries. In <i>Text summarization</i>	678
623	els affect the practice of science? <i>Proceedings of the</i>	<i>branches out</i> , pages 74–81.	679
624	<i>National Academy of Sciences</i> , 122(5):e2401227121.	R T Loftus, D E MacHugh, D G Bradley, P M Sharp, and	680
625	Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel	P Cunningham. 1994. Evidence for two independent	681
626	Weld. 2020. TLDR: Extreme summarization of sci-	domestications of cattle. <i>Proceedings of the National</i>	682
627	entific documents. In <i>Findings of the Association</i>	<i>Academy of Sciences</i> , 91(7):2757–2761.	683
628	<i>for Computational Linguistics: EMNLP 2020</i> , pages	Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-	684
629	4766–4777, Online. Association for Computational	XScience: A large-scale dataset for extreme multi-	685
630	Linguistics.	document summarization of scientific articles. In	686
631	DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang,	<i>Proceedings of the 2020 Conference on Empirical</i>	687
632	Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,	<i>Methods in Natural Language Processing (EMNLP)</i> ,	688
633	Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang,	pages 8068–8074, Online. Association for Computa-	689
634	Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhi-	tional Linguistics.	690
635	hong Shao, Zhuoshu Li, Ziyi Gao, and 181 others.	Yuning Mao, Ming Zhong, and Jiawei Han. 2022. Cite-	691
636	2025. DeepSeek-R1: Incentivizing Reasoning Capa-	Sum: Citation text-guided scientific extreme sum-	692
637	bility in LLMs via Reinforcement Learning. <i>Preprint</i> ,	marization and domain adaptation with limited su-	693
638	arXiv:2501.12948.	perision. In <i>Proceedings of the 2022 Conference</i>	694
639	Ashwin Devaraj, Iain Marshall, Byron Wallace, and	<i>on Empirical Methods in Natural Language Process-</i>	695
640	Junyi Jessy Li. 2021. Paragraph-level Simplifica-	<i>ing</i> , pages 10922–10935, Abu Dhabi, United Arab	696
641	tion of Medical Texts. In <i>Proceedings of the 2021</i>	Emirates. Association for Computational Linguistics.	697
642	<i>Conference of the North American Chapter of the</i>	OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai,	698
643	<i>Association for Computational Linguistics: Human</i>	Sam Altman, Andy Applebaum, Edwin Arbus,	699
644	<i>Language Technologies</i> , pages 4972–4984, Online.	Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao,	700
645	Association for Computational Linguistics.	Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita	701
646	Esin Durmus, He He, and Mona Diab. 2020. FEQA: A	Brett, Eugene Brevdo, Greg Brockman, Sebastien	702
647	question answering evaluation framework for faith-	Bubeck, Che Chang, and 107 others. 2025. Gpt-	703
648	fulness assessment in abstractive summarization. In	oss-120b & gpt-oss-20b Model Card. <i>Preprint</i> ,	704
649	<i>Proceedings of the 58th Annual Meeting of the Asso-</i>	arXiv:2508.10925.	705
650	<i>ciation for Computational Linguistics</i> , pages 5055–	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	706
651	5070, Online. Association for Computational Lin-	Jing Zhu. 2002. Bleu: a method for automatic evalu-	707
652	guistics.	ation of machine translation. In <i>Proceedings of the</i>	708
653	Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and	<i>40th annual meeting of the Association for Computa-</i>	709
654	Carolina Scarton. 2022. Making science simple: Cor-	<i>tional Linguistics</i> , pages 311–318.	710
655	pora for the lay summarisation of scientific literature.	Jason Priem, Heather Piwowar, and Richard Orr. 2022.	711
656	In <i>Proceedings of the 2022 Conference on Empiri-</i>	OpenAlex: A fully-open index of scholarly works,	712
657	<i>cal Methods in Natural Language Processing</i> , pages	authors, venues, institutions, and concepts. <i>Preprint</i> ,	713
658	10589–10604, Abu Dhabi, United Arab Emirates. As-	arXiv:2205.01833.	714
659	sociation for Computational Linguistics.	Juliette Riquet, Wouter Coppieters, Nadine Cambisano,	715
660	Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and	Juan-José Arranz, Paulette Berzi, Scott K. Davis,	716
661	Trevor Cohen. 2024. Retrieval augmentation of large	Bernard Grisart, Frédéric Farnir, Latifa Karim, Myr-	717
662	language models for lay language generation. <i>Jour-</i>	iam Mni, Patricia Simon, Jeremy F. Taylor, Pascal	718
663	<i>nal of Biomedical Informatics</i> , 149:104580.		

719 Vanmanshoven, Danny Wagenaar, James E. Womack, and Michel Georges. 1999. [Fine-mapping of quantitative trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle](#). *Proceedings of the National Academy of Sciences*, 96(16):9252–9257.

725 Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get To The Point: Summarization with Pointer-Generator Networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

732 Eva Sharma, Chen Li, and Lu Wang. 2019. [BIG-PATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

738 Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2022. [X-scitldr: cross-lingual extreme summarization of scholarly documents](#). In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, JCDL '22*, New York, NY, USA. Association for Computing Machinery.

745 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

750 Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. [TL;DR: Mining Reddit to learn automatic summarization](#). In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

756 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. [Qwen3 Technical Report](#). *Preprint, arXiv:2505.09388*.

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

767 Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. [Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance](#). *arXiv preprint arXiv:1909.02622*.

## A Dataset Statistics 772

### A.1 Details about Table 1 773

The numbers of words and sentences are counted using NLTK and spaCy (Honnibal et al., 2020), respectively. SciTLDR provides options for input space — “abstract-only” and “abstract, introduction, and conclusion sections (AIC)” — and the statistics reported are based on the former. We also noted that the dev set contains one more sample than claimed, resulting in a total pair count that exceeds the original paper (5,411) by one. The source texts in Multi-XScience are significantly longer as they comprise concatenated abstracts from both the citing and cited papers. It also uses whole paragraphs as target summaries, rather than using sentences like other datasets, which results in longer target summary lengths as well. 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788

The number of citations is obtained through CrossRef based on DOIs. For datasets using identifiers other than DOIs (e.g., CiteSum uses Semantic Scholar Corpus ID and Multi-XScience uses Microsoft Academic Graph ID), these identifiers are first transformed into DOIs. Due to this conversion, the subset of samples with successfully retrieved citation counts is reduced compared to ours; nevertheless, the values reported are computed based only on valid samples. For Multi-XScience, the source citation count represents the average number of citations for its constituent papers. The symbol “-” denotes that the citations cannot be obtained since the identifier is not provided by the dataset. 789 790 791 792 793 794 795 796 797 798 799 800 801 802

### A.2 Discipline distribution 803

Fig. 3 shows the sample-level discipline distribution of BiomedTLDR, i.e., discipline will be counted multiple times for multiple-annotated papers. The discipline is derived from level 0 concepts in OpenAlex (Priem et al., 2022). We take the concept with the highest score if there are multiple for one paper. 804 805 806 807 808 809 810

For comparison, we also plot the source text discipline distribution at the sample level and the paper level, for CiteSum in Fig. 4 and Multi-XScience in Fig. 5, respectively. The discipline in Fig. 4 is derived from the embedded MAG field of study information as in the original paper. 811 812 813 814 815 816

The discipline in Fig. 5 is derived using the same pipeline as in Fig. 3, but at the paper-level due to the one-to-many association in Multi-XScience. 817 818 819

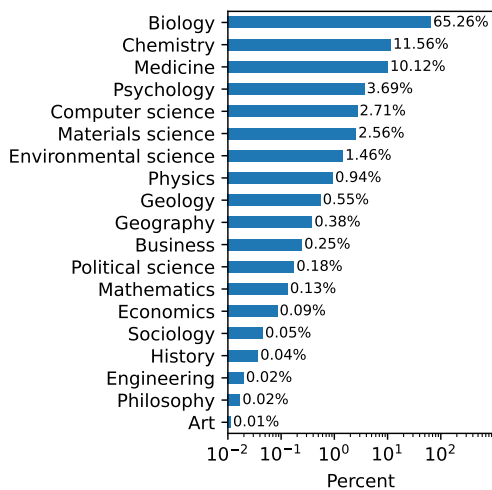


Figure 3: The distribution of source abstracts' discipline in BiomedTLDR.

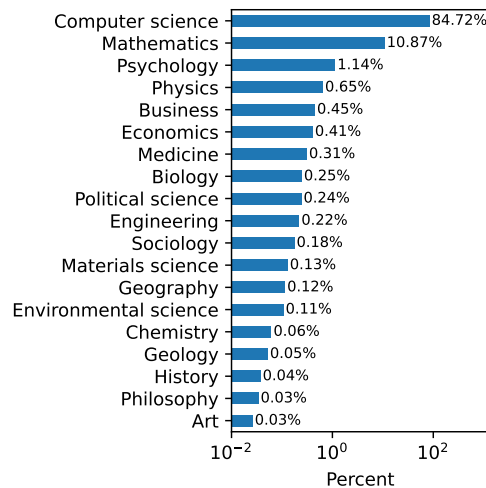


Figure 5: The distribution of source abstracts' discipline in Multi-XScience

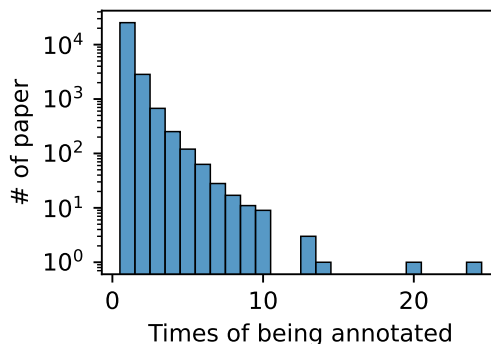


Figure 6: The paper-level distribution of times of being annotated.

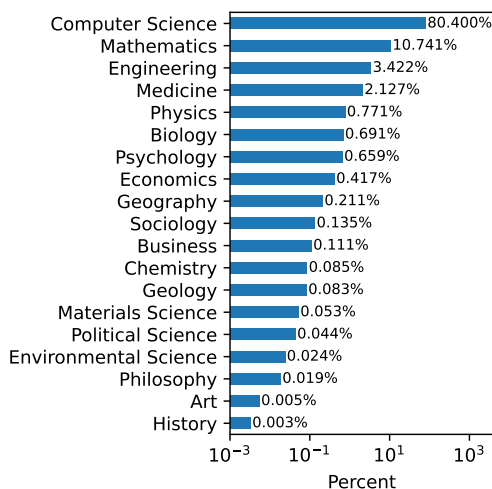


Figure 4: The distribution of source abstracts' discipline in CiteSum

### A.3 Multiple annotations

820

Fig. 6 shows the distribution of the times of annotation for 29,342 papers in our dataset, which lays the foundation for our methods of using multi-annotated papers as a human baseline in § 5.5.

821

822

823

824

## B Generating summaries using LLMs

825

### B.1 Instructions

826

We use the following SYSTEM and MESSAGE instructions to guide LLMs in summarizing papers.

827

828

**SYSTEM** You are an expert scientific assistant specialized in generating concise, accurate, and insightful bibliography annotations. Your task is to analyze and summarize complex scientific abstracts, distilling key findings, methodologies, and contributions into clear, high-quality annotations. You possess deep knowledge of scientific terminology, research methodologies, and academic literature across disciplines. Focus on highlighting novel

829

830

831

832

833

834

835

836

837

838 approaches, significant results, and broader impli- 890  
839 cations while maintaining precision and brevity. 891  
840 Output only the final annotation without additional 892  
841 commentary or formatting, and please make your 893  
842 annotation around the provided length. 894

843 **MESSAGE user** [Abstract] CD30 is up- 895  
844 regulated in several human diseases and viral in- 896  
845 fections but its role in immune regulation is poorly 897  
846 understood. Here, we report the expression of a 898  
847 functional soluble CD30 homologue, viral CD30 899  
848 (vCD30), encoded by ectromelia (mousepox) virus, 900  
849 a poxvirus that causes a severe disease related to 901  
850 human smallpox. We show that vCD30 is a 12-kD 902  
851 secreted protein that not only binds CD30L with 903  
852 high affinity and prevents its interaction with CD30, 904  
853 but it also induces reverse signaling in cells express- 905  
854 ing CD30L. vCD30 blocked the generation of inter- 906  
855 feron  $\gamma$ -producing cells in vitro and was a potent in- 907  
856 hibitor of T helper cell (Th)1- but not Th2-mediated 908  
857 inflammation in vivo. The finding of a CD30 ho- 909  
858 mologue encoded by ectromelia virus suggests a 910  
859 role for CD30 in antiviral defense. Characteriza- 911  
860 tion of the immunological properties of vCD30 has 912  
861 uncovered a role of CD30-CD30L interactions in 913  
862 the generation of inflammatory responses. [Word 914  
863 count: 56] 915

864 **MESSAGE assistant** References 38 and 39 re- 916  
865 port the identification of a CD30 homologue en- 917  
866 coded by poxviruses. Reference 39 also shows that 918  
867 viral CD30 is the first example of a viral cytokine 919  
868 receptor that induces reverse signalling when bind- 920  
869 ing its ligand, and provides evidence of a role for 921  
870 host CD30 in T H 1 inflammatory responses. 922

871 **MESSAGE user** [Abstract] Carbon monoxide 923  
872 occurs at relatively high concentrations ( $\geq 800$  parts 924  
873 per million) in Mars' atmosphere, where it repre- 925  
874 sents a potentially significant energy source that 926  
875 could fuel metabolism by a localized putative sur- 927  
876 face or near-surface microbiota. However, the plau- 928  
877 sibility of CO oxidation under conditions relevant 929  
878 for Mars in its past or at present has not been eval- 930  
879 uated. Results from diverse terrestrial brines and 931  
880 saline soils provide the first documentation, to our 932  
881 knowledge, of active CO uptake at water potentials 933  
882 (-41 MPa to -117 MPa) that might occur in putative 934  
883 brines at recurrent slope lineae (RSL) on Mars. 935  
884 Results from two extremely halophilic isolates 936  
885 complement the field observations. Halorubrum 937  
886 str. BV1, isolated from the Bonneville Salt Flats, 938  
887 Utah (to our knowledge, the first documented ex- 939  
888 tremely halophilic CO-oxidizing member of the 940  
889 Euryarchaeota), consumed CO in a salt-saturated 941

890 medium with a water potential of -39.6 MPa; ac- 891  
892 tivity was reduced by only 28% relative to activity 893  
894 at its optimum water potential of -11 MPa. A pro- 895  
896 teobacterial isolate from hypersaline Mono Lake, 896  
897 California, *Alkalilimnicola ehrlichii* MLHE-1, also 897  
898 oxidized CO at low water potentials (-19 MPa), at 898  
899 temperatures within ranges reported for RSL, and 899  
900 under oxic, suboxic (0.2% oxygen), and anoxic 900  
901 conditions (oxygen-free with nitrate). MLHE-1 901  
902 was unaffected by magnesium perchlorate or low 902  
903 atmospheric pressure (10 mbar). These results col- 903  
904 lectively establish the potential for microbial CO 904  
905 oxidation under conditions that might obtain at lo- 905  
906 cal scales (e.g., RSL) on contemporary Mars and at 906  
907 larger spatial scales earlier in Mars' history. [Word 907  
908 count: 23] 908

909 **MESSAGE assistant** This work is the first re- 909  
910 port of atmospheric trace gas oxidation by mem- 910  
911 bers of the archaea, including at astrobiologically 911  
912 relevant water potentials 912

913 **MESSAGE user** [Abstract] Abstract Inhibi- 913  
914 tion of class II trans-activator (CIITA) expres- 914  
915 sion prevents embryonic trophoblast cells from 915  
916 up-regulating MHC class II genes in response to 916  
917 IFN- $\gamma$ . This is thought to be one mechanism of 917  
918 maternal tolerance to the fetal allograft. The CI- 918  
919 ITA gene is regulated by four distinct promoters; 919  
920 promoter III directs constitutive (B cell) expres- 920  
921 sion, and promoter IV regulates IFN- $\gamma$ -inducible 921  
922 expression. Using in vivo genomic footprinting, 922  
923 promoter-reporter analysis, Southern blot analysis, 923  
924 and RT-PCR, we have examined the cause of CIITA 924  
925 silencing in a trophoblast-derived cell line. We re- 925  
926 port here that methylation of promoter IV DNA at 926  
927 CpG sites in Jar cells prevents promoter occupancy 927  
928 and IFN- $\gamma$ -inducible transcription. The inhibition 928  
929 of CpG methylation in Jar cells by treatment with 929  
930 5-aza-2'-deoxycytidine restores IFN- $\gamma$  inducibility 930  
931 to CIITA. This is the first description of an epige- 931  
932 netic mechanism involved in regulation of CIITA 932  
933 and MHC class II gene expression. [Word count: 933  
934 21] 934

935 **MESSAGE assistant** References 43 and 44 re- 935  
936 port that the gene encoding CIITA is irreversibly 936  
937 silenced in trophoblasts by DNA methylation of 937  
938 pIV. 938

939 **MESSAGE user** [Abstract] High-throughput di- 939  
940 rect sequencing techniques have recently opened 940  
941 the possibility to sequence genomes from Pleis- 941  
942 tocene organisms. Here we analyze DNA se- 942  
943 quences determined from a Neandertal, a mam- 943  
944 moth, and a cave bear. We show that purines are 944

942 overrepresented at positions adjacent to the breaks  
 943 in the ancient DNA, suggesting that depurination  
 944 has contributed to its degradation. We furthermore  
 945 show that substitutions resulting from miscoding  
 946 cytosine residues are vastly overrepresented in the  
 947 DNA sequences and drastically clustered in the  
 948 ends of the molecules, whereas other substitutions  
 949 are rare. We present a model where the observed  
 950 substitution patterns are used to estimate the rate  
 951 of deamination of cytosine residues in single- and  
 952 double-stranded portions of the DNA, the length  
 953 of single-stranded ends, and the frequency of nicks.  
 954 The results suggest that reliable genome sequences  
 955 can be obtained from Pleistocene organisms. [Word  
 956 count: 26]

957 **MESSAGE assistant** This study characterizes  
 958 typical nucleotide misincorporation and fragmenta-  
 959 tion patterns using HTS data from aDNA extracts,  
 960 which have been subsequently used as essential  
 961 authentication criteria.

962 **MESSAGE user** [Abstract] Summary: Annotating  
 963 genetic variants, especially non-coding variants,  
 964 for the purpose of identifying pathogenic variants  
 965 remains a challenge. Combined annotation-  
 966 dependent depletion (CADD) is an algorithm de-  
 967 signed to annotate both coding and non-coding  
 968 variants, and has been shown to outperform other  
 969 annotation algorithms. CADD trains a linear kernel  
 970 support vector machine (SVM) to differentiate evo-  
 971 lutionarily derived, likely benign, alleles from sim-  
 972 ulated, likely deleterious, variants. However, SVMs  
 973 cannot capture non-linear relationships among the  
 974 features, which can limit performance. To address  
 975 this issue, we have developed DANN. DANN uses  
 976 the same feature set and training data as CADD to  
 977 train a deep neural network (DNN). DNNs can cap-  
 978 ture non-linear relationships among features and  
 979 are better suited than SVMs for problems with a  
 980 large number of samples and features. We exploit  
 981 Compute Unified Device Architecture-compatible  
 982 graphics processing units and deep learning tech-  
 983 niques such as dropout and momentum training  
 984 to accelerate the DNN training. DANN achieves  
 985 about a 19% relative reduction in the error rate  
 986 and about a 14% relative increase in the area un-  
 987 der the curve (AUC) metric over CADD’s SVM  
 988 methodology. Availability and implementation: All  
 989 data and source code are available at [https://  
 990 cbcl.ics.uci.edu/public\\_data/DANN/](https://cbcl.ics.uci.edu/public_data/DANN/). Con-  
 991 tact: xhx@ics.uci.edu [Word count: 27]

992 **MESSAGE assistant** This paper presents a  
 993 pathogenicity scoring method, which is a deep

learning (CNN) version of CADD, for coding and 994  
 non-coding variant fitness impact 995

## B.2 Sample outputs 996

Here we provide two examples of papers summa- 997  
 rized by different LLMs and human authors. 998

1. DOI 10.1111/j.1474-9726.2012.00870.x 999

**Ttitle** Postmitotic neurons develop a p21- 1000  
 dependent senescence-like phenotype driven 1001  
 by a DNA damage response 1002

**Abstract** Summary In senescent cells, a DNA 1003  
 damage response drives not only irreversible 1004  
 loss of replicative capacity but also produc- 1005  
 tion and secretion of reactive oxygen species 1006  
 (ROS) and bioactive peptides including pro- 1007  
 inflammatory cytokines. This makes senes- 1008  
 cent cells a potential cause of tissue func- 1009  
 tional decline in aging. To our knowledge, we 1010  
 show here for the first time evidence suggest- 1011  
 ing that DNA damage induces a senescence- 1012  
 like state in mature postmitotic neurons in 1013  
 vivo . About 40-80% of Purkinje neurons 1014  
 and 20-40% of cortical, hippocampal and 1015  
 peripheral neurons in the myenteric plexus 1016  
 from old C57Bl/6 mice showed severe DNA 1017  
 damage, activated p38MAPkinase, high ROS 1018  
 production and oxidative damage, interleukin 1019  
 IL-6 production, heterochromatinization and 1020  
 senescence-associated  $\beta$ -galactosidase activ- 1021  
 ity. Frequencies of these senescence-like neu- 1022  
 rons increased with age. Short-term caloric 1023  
 restriction tended to decrease frequencies of 1024  
 positive cells. The phenotype was aggrava- 1025  
 ted in brains of late-generation TERC-/- 1026  
 mice with dysfunctional telomeres. It was 1027  
 fully rescued by loss of p21(CDKN1A) func- 1028  
 tion in late-generation TERC-/-CDKN1A-/- 1029  
 mice, indicating p21 as the necessary sig- 1030  
 nal transducer between DNA damage re- 1031  
 sponse and senescence-like phenotype in neu- 1032  
 rons, as in senescing fibroblasts and other 1033  
 proliferation-competent cells. We conclude 1034  
 that a senescence-like phenotype is possibly 1035  
 not restricted to proliferation-competent cells. 1036  
 Rather, dysfunctional telomeres and/or accu- 1037  
 mulated DNA damage can induce a DNA dam- 1038  
 age response leading to a phenotype in postmi- 1039  
 totic neurons that resembles cell senescence 1040  
 in multiple features. Senescence-like neurons 1041  
 might be a source of oxidative and inflamma- 1042  
 tory stress and a contributor to brain aging. 1043

1044	<b>Human</b> This study provides in vivo evidence	1094
1045	that postmitotic cells can develop a senescent-	
1046	like phenotype	
1047	<b>Gemma 3</b> This study establishes that postmi-	
1048	totic neurons can undergo a senescence-like	
1049	phenotype induced by DNA damage, charac-	
1050	terized by oxidative stress and inflammation,	
1051	potentially contributing to brain aging.	
1052	<b>DeepSeek-R1</b> This study provides evidence	
1053	that DNA damage can induce a senescence-	
1054	like state in mature, postmitotic neurons in	
1055	vivo, characterized by ROS production, in-	
1056	flammation, and senescence-associated mark-	
1057	ers. It shows that such neurons accumulate	
1058	with age and are exacerbated by telomere dys-	
1059	function, while caloric restriction or p21 loss	
1060	mitigates the phenotype.	
1061	<b>Gpt-oss</b> The authors demonstrate that DNA	
1062	damage triggers a senescence-like program in	
1063	post-mitotic neurons in vivo. In aged C57Bl/6	
1064	mice, 40-80 % of Purkinje cells and 20-40	
1065	% of cortical, hippocampal and enteric neu-	
1066	rons exhibit DNA lesions, p38MAPK activa-	
1067	tion, ROS, IL-6, heterochromatinization and	
1068	SA- $\beta$ -gal activity; these features increase with	
1069	age, are reduced by short-term caloric restric-	
1070	tion, worsen in telomere-deficient TERC-/-	
1071	mice, and are abolished by p21 (CDKN1A)	
1072	loss. The work extends cellular senescence	
1073	to non-proliferative neurons, implicating	
1074	DNA-damage-induced, p21-dependent	
1075	senescence-like neurons as contributors to ox-	
1076	idative, inflammatory stress and brain aging.	
1077	<b>Llama 4</b> These findings provide evidence that	
1078	neurons can undergo senescence-like changes,	
1079	characterized by DNA damage, oxidative	
1080	stress, and inflammatory responses, which can	
1081	contribute to brain aging and age-related dis-	
1082	eases. The study suggests that senescence-like	
1083	changes in neurons can be induced by DNA	
1084	damage and telomere dysfunction, and that	
1085	these changes can be rescued by loss of p21	
1086	function. The findings also highlight the po-	
1087	tential role of senescent neurons in promoting	
1088	oxidative stress and inflammation in the brain.	
1089	Overall, the study provides new insights into	
1090	the mechanisms underlying brain aging and	
1091	age-related diseases, and suggests that target-	
1092	ing senescent neurons may be a promising	
1093	therapeutic strategy for preventing or treating	
	these conditions.	1095
	<b>Qwen 3</b> This study demonstrates that DNA	1096
	damage induces a senescence-like phenotype	1097
	in postmitotic neurons, characterized by p21-	1098
	dependent activation of DNA damage re-	1099
	sponse, ROS production, and inflammation,	1100
	linking neuronal aging to systemic oxidative	1101
	and inflammatory stress.	
2.	<b>DOI</b> 10.1126/science.1080029	1102
	<b>Title</b> A Genomic View of the Human-	1103
	Bacteroides thetaiotaomicron Symbiosis	1104
	<b>Abstract</b> The human gut is colonized with	1105
	a vast community of indigenous microorgan-	1106
	isms that help shape our biology. Here, we	1107
	present the complete genome sequence of the	1108
	Gram-negative anaerobe Bacteroides thetaiota-	1109
	omicron , a dominant member of our normal	1110
	distal intestinal microbiota. Its 4779-	1111
	member proteome includes an elaborate ap-	1112
	paratus for acquiring and hydrolyzing oth-	1113
	erwise indigestible dietary polysaccharides	1114
	and an associated environment-sensing sys-	1115
	tem consisting of a large repertoire of extra-	1116
	cytoplasmic function sigma factors and one-	1117
	and two-component signal transduction sys-	1118
	tems. These and other expanded paralogous	1119
	groups shed light on the molecular mecha-	1120
	nisms underlying symbiotic host-bacterial re-	1121
	lationships in our intestine.	1122
	<b>Human</b> Description of the genome sequence	1123
	of B. thetaiotaomicron , which revealed 163	1124
	susCD paralogues and the largest number of	1125
	genes related to carbohydrate utilization of	1126
	any bacterium so far studied.	1127
	<b>Gemma 3</b> The study of Bacteroides thetaiota-	1128
	omicron is an example of a genomic anal-	1129
	ysis of a human symbiont and includes the	1130
	first complete genome sequence of a Gram-	1131
	negative anaerobe.	1132
	<b>DeepSeek-R1</b> This paper reports the complete	1133
	genome sequence of *Bacteroides thetaiota-	1134
	omicron*, a key gut commensal, revealing	1135
	its extensive proteome specialized for polysac-	1136
	charide metabolism and environmental sens-	1137
	ing.	1138
	<b>Gpt-oss</b> The study reports the finished	1139
	genome of Bacteroides thetaiotaomi-	1140
	cron, a key human gut symbiont, reveal-	1141
	ing a 4,779-protein repertoire rich in	1142

1143	polysaccharide-degrading enzymes and an	paralysis. These data support the hypothesis	1192
1144	extensive suite of extracytoplasmic-function	that a conventional kinesin plays a role in the	1193
1145	sigma factors and one-/two-component	microtubule-dependent slow axonal transport	1194
1146	systems that together elucidate the molecular	of at least one cargo, the NF proteins.	1195
1147	basis of host-microbe mutualism in the		
1148	intestine.	<b>Human:</b> This study analyses the conse-	1196
1149		quences of kinesin disruption in mice and pro-	1197
1150	<b>Llama 4</b> Reference 51 details the complete	vides evidence for a role of kinesin in neuro-	1198
1151	genome sequence of *Bacteroides thetaio-	filament transport.	1199
1152	taomicron*, revealing extensive genomic ca-		
1153	capacity for polysaccharide degradation and en-	<b>Llama 4:</b>	1200
1154	vironment sensing, shedding light on host-	This study reveals that tau, a MAP, differen-	1201
1155	microbe symbiosis.	tially inhibits the movement of the two major	1202
1156		motor proteins, dynein and kinesin. ... (repeti-	1203
1157	<b>Qwen3</b> This study reports the complete	tions) The study reveals that tau differentially	1204
1158	genome sequence of *Bacteroides thetaio-	inhibits the movement of dynein and kinesin.	1205
1159	taomicron*, a key gut symbiont, revealing	The study reveals that tau differentially in-	1206
1160	expanded polysaccharide utilization systems,	hibits the movement of dynein and kinesin.	1207
1161	extracytoplasmic function sigma factors, and	The study reveals that tau differentially in-	1208
1162	two-component signal transduction systems	hibits the movement of dynein and kinesin.	1209
1163	that underpin its adaptation to the intestinal	The study reveals that tau differentially in-	1210
1164	environment.	hibits the movement of dynein and kinase ac-	1211
1165		tivity. The study reveals that tau differentially	1212
1166	<b>B.3 Samples of Nonsense output from LLMs</b>	inhibits the movement of dys, a and kin). The	1213
1167	Here we provide two samples of meaningless out-	study reveals that tau differentially inhibits the	1214
1168	put from LLMs. We noticed that Llama 4 tend	movement of the kinase. The study reveals	1215
1169	to repeat medium-sized phrases, while Qwen 3	that tau differentially inhibits the movement	1216
1170	directly output its thinking, though we explicitly	of the kinase. The study reveals that tau differ-	1217
1171	disable this feature in Ollama configuration.	ently, and efficiently. The study reveals that,	1218
1172		due to the different level of efficiency and,	1219
1173	1. <b>DOI:</b> 10.1083/jcb.200301026	the impact that the differential inhibition of	1220
1174	<b>Abstract:</b> To test the hypothesis that fast an-	the efficiency of the movement of the differ-	1221
1175	terograde molecular motor proteins power the	ent and specific kinemics. The study reveals	1222
1176	slow axonal transport of neurofilaments (NFs),	that the level of inhibition and the level of the	1223
1177	we used homologous recombination to gener-	movement of the kinesi and the level of the	1224
1178	ate mice lacking the neuronal-specific conven-	kinetic activity of the kinesis of the human	1225
1179	tional kinesin heavy chain, KIF5A. Because	brain, which includes and/or the human being.	1226
1180	null KIF5A mutants die immediately after	The study reveals that the level of inhibition of	1227
1181	birth, a synapsin-promoted Cre-recombinase	the movement of the kineses. The study reveals	1228
1182	transgene was used to direct inactivation of	that the level of the inhibition of the level of	1229
1183	KIF5A in neurons postnatally. Three fourths	the chemical reaction. The level of chemical	1230
1184	of such mutant mice exhibited seizures and	kinetic The level of chemical reactions and	1231
1185	death at around 3 wk of age; the remaining ani-	the level of kinetic chemical reaction of the	1232
1186	imals survived to 3 mo or longer. In young mu-	reaction and the level of the reaction, or the	1233
1187	tant animals, fast axonal transport appeared	level of the reaction. The level of the level	1234
1188	to be intact, but NF-H, as well as NF-M and	of chemical and kinetic efficiency of the ki-	1235
1189	NF-L, accumulated in the cell bodies of pe-	netic activity of the level of the activity of the	1236
1190	ripheral sensory neurons accompanied by a	level of the activity of the level of the chem-	1237
1191	reduction in sensory axon caliber. Older ani-	ical and kinetic. The level of the level of ...	1238
	imals also developed age-dependent sensory	(repetitions)	1239
	neuron degeneration, an accumulation of NF		
	subunits in cell bodies and a reduction in ax-	2. <b>DOI:</b> 10.1073/pnas.0509720103	1240
	ons, loss of large caliber axons, and hind limb	<b>Abstract:</b> The diversity of antigen receptors	1241

1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292

in the adaptive immune system of jawed vertebrates is generated by a unique process of somatic gene rearrangement known as V(D)J recombination. The Rag1 and Rag2 proteins are the key mediators of this process. They are encoded by a compact gene cluster that has exclusively been identified in animal species displaying V(D)J-mediated immunity, and no homologous gene pair has been identified in other organisms. This distinctly restricted phylogenetic distribution has led to the hypothesis that one or both of the Rag genes were coopted after horizontal gene transfer and assembled into a Rag1/2 gene cluster in a common jawed vertebrate ancestor. Here, we identify and characterize a closely linked pair of genes, SpRag1L and SpRag2L, from an invertebrate, the purple sea urchin (*Strongylocentrotus purpuratus*) with similarity in both sequence and genomic organization to the vertebrate Rag1 and Rag2 genes. They are co-expressed during development and in adult tissues, and recombinant versions of the proteins form a stable complex with each other as well as with Rag1 and Rag2 proteins from several vertebrate species. We thus conclude that SpRag1L and SpRag2L represent homologs of vertebrate Rag1 and Rag2. In combination with the apparent absence of V(D)J recombination in echinoderms, this finding strongly suggests that linked Rag1 - and Rag2 -like genes were already present and functioning in a different capacity in the common ancestor of living deuterostomes, and that their specific role in the adaptive immune system was acquired much later in an early jawed vertebrate.

**Human:** This paper is the first to report a RAG1L–RAG2L gene pair in any organism other than a jawed vertebrate (in this case, the purple sea urchin *S. purpuratus*), providing evidence for an early evolutionary origin for the RAG recombinase

**Qwen3:**

<think> Okay, let me start by reading the abstract carefully. The study is about tropical cyclones (TCs) causing power outages and how these outages combined with heatwaves could become a bigger risk in the future, especially under climate change. They used Harris County, Texas as a case study. The key points from the abstract are: TCs lead to power out-

ages, which when followed by heatwaves create a compound hazard. Under RCP8.5, the scenario with high greenhouse gas emissions, the frequency of long-duration heatwaves after strong TCs is expected to increase. The study found that the percentage of residents in Harris County facing these compound hazards could rise from 0.8% to 18.2% over 20 years by the end of the century. They also suggest that improving the power grid, like undergrounding lines and using distributed energy sources, can mitigate this risk. Now, the user wants an annotation that's concise, highlighting the main findings, methods, and implications. The previous examples focused on the main contribution, sometimes mentioning the methods briefly and the broader impact. I need to make sure to mention the compound hazard (TC-blackout-heatwave), the use of projections coupled with power outage analysis, the location (Harris County), the projected increase in risk, and the mitigation strategies proposed. Also, note that this is under RCP8.5, which is the high-emissions scenario. The word count here is 11, so the annotation needs to be very tight. Looking at previous responses, they often start with "This study..." or "This paper...", then state the main finding, methods, and implications. For example, the DANN paper started with the method (deep learning version of CADD), the problem it addresses, and the improvement. Similarly, the Mars CO oxidation study focused on the first report of archaea using CO under those conditions. In this case, the main takeaway is the increased compound hazard risk due to climate change and the mitigation strategies. The coupling of TC and heatwave projections with power outage analysis is the methodology. The annotation should state that the study projects a significant increase in compound hazards in Harris County under RCP8.5 and that grid enhancements can reduce this risk. The user might want to emphasize the urgency of climate adaptation actions. I need to avoid technical jargon where possible but still be precise. Words like "project" or "coupled TC and heatwave projections" could be in the annotation. Also, mention Harris County as an example. The result is a 23-fold increase in risk, but the annotation should state it clearly without the exact factor if word count is tight.

1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344

1345 Putting it all together: "This study couples  
1346 TC and heatwave projections with power out-  
1347 age modeling to show that climate change  
1348 could sharply increase the risk of TC-blackout-  
1349 heatwave compound hazards, particularly in  
1350 coastal regions like Harris County, Texas. The  
1351 findings highlight the urgency of enhancing  
1352 grid resilience through measures such as un-  
1353 dergrounding power lines and distributed en-  
1354 ergy systems to mitigate escalating risks from  
1355 compound climate extremes." That's around  
1356 40 words. The user's previous examples were  
1357 shorter (e.g., "This work is the first report..."  
1358 at 15 words). Wait, the user's word count here  
1359 is 11, but the instruction might have meant  
1360 the length of the annotation. Wait, looking  
1361 back at the history, the user's examples vary  
1362 in length but are generally concise. The latest  
1363 example was about the DANN paper with an  
1364 annotation of about 20 words. Maybe the user  
1365 wants a very short annotation. Let me check  
1366 the latest query: the user provided a word  
1367 count of 11 for the abstract and the previous  
1368 examples for similar word counts had annota-  
1369 tions around 20-30 words. Hmm, perhaps the  
1370 user wants the annotation to be roughly pro-  
1371 portional. Let me check the initial instruction:  
1372 "please make your annotation around the pro-  
1373 vided length." The provided length here is the  
1374 word count of the abstract, which is 11. So the  
1375 annotation should be around 11 words. Wait,  
1376 but looking back, for example, the abstract  
1377 with word count 56 had a reference of about  
1378 28 words. The DANN paper abstract had 27  
1379 words, and the annotation was 20 words. The  
1380 user's instruction says to make the annotation  
1381 "around the provided length". So in this case,  
1382 the abstract's word count is 11, so the anno-  
1383 tation should be around 11 words. However,  
1384 the example annotations provided by the user  
1385 sometimes exceed the word count. For in-  
1386 stance, the Mars abstract had a word count of  
1387 23, and the annotation was 20 words. Maybe  
1388 "around" allows for some flexibility. In this  
1389 case, to fit into 11 words: "This study projects  
1390 a 23-fold increase in TC-blackout-heatwave  
1391 compound hazards in Harris County by 2100  
1392 under RCP8.5, emphasizing urgent grid re-  
1393 siliance needs." That's 24 words. Too long.  
1394 Alternatively: "Climate change may increase  
1395 TC-blackout-heatwave hazards; power grid  
1396 upgrades can reduce risk." That's 11 words.

1397 But maybe it's better to include the location  
1398 and the method. Hmm. Let me rephrase. "Pro-  
1399 jections show rising TC-blackout-heatwave  
1400 risks in Harris County under RCP8.5, neces-  
1401 sitating grid resilience measures." 17 words.  
1402 Still over. Let's try: "Under RCP8.5, Har-  
1403 ris County faces rising TC-blackout-heatwave  
1404 hazards; grid enhancements mitigate risk."  
1405 15 words. Still over. Maybe: "TC-blackout-  
1406 heatwave hazards to surge in Harris County  
1407 under RCP8.5; grid upgrades reduce risk."  
1408 13 words. Closer. Alternatively: "Projec-  
1409 tions indicate sharp increases in TC-blackout-  
1410 heatwave compound hazards with climate  
1411 change." 11 words. But misses the grid so-  
1412 lution. Maybe the most critical part is the  
1413 finding and the method. The user's exam-  
1414 ples sometimes include implications. For in-  
1415 stance, the Mars study included "including  
1416 at astrobiologically relevant water potentials".  
1417 The DANN paper included the method (deep  
1418 learning). In this case, the key elements are:  
1419 coupling of TC and heatwave projections, the  
1420 increased risk, the location, and the mitiga-  
1421 tion. However, given the word limit, perhaps  
1422 focus on the main finding and implication.  
1423 So: "Coupled projections indicate sharply in-  
1424 creased TC-blackout-heatwave hazards under  
1425 climate change, urging grid resilience strate-  
1426 gies." That's 16 words. Still a bit over, but  
1427 captures the main points. Alternatively: "Cli-  
1428 mate change heightens TC-blackout-heatwave  
1429 risks; enhanced power grids mitigate hazard  
1430 exposure." 13 words. Given the previous ex-  
1431 amples, sometimes they're concise without  
1432 specific details. For instance, the CADD  
1433 paper's annotation: "This paper presents a  
1434 pathogenicity scoring method, which is a deep  
1435 learning (CNN) version of CADD, for cod-  
1436 ing and non-coding variant fitness impact."  
1437 That's about 25 words. The user's instruc-  
1438 tion might not strictly enforce the word count  
1439 but to be around. The Mars abstract's anno-  
1440 tation was "This work is the first report..."  
1441 at 15 words. So, to balance, perhaps: "This  
1442 study projects sharp increases in TC-blackout-  
1443 heatwave compound hazards under RCP8.5,  
1444 emphasizing urgent power grid adaptations."  
1445 18 words. Alternatively, including the loca-  
1446 tion: "Projections show Harris County's ris-  
1447 ing TC-blackout-heatwave risk under RCP8.5;  
1448 grid upgrades mitigate." 14 words. Alterna-

1449	tively, considering the original answer given	better. But the user's example annotations are	1501
1450	by the assistant in the example, which for the	often even shorter. For instance, the CADD	1502
1451	CADD paper's abstract (word count 21) had	paper's reference was two sentences, but the	1503
1452	a reference of "References 43 and 44 report	user example for that was a single sentence.	1504
1453	that the gene encoding CIITA is irreversibly	The previous answers vary, but aim for one	1505
1454	silenced in trophoblasts by DNA methylation	sentence, capturing key elements. Final at-	1506
1455	of pIV." (22 words). So around the same as the	tempt: "This study projects a 23-fold increase	1507
1456	abstract's word count. Thus, for an 11-word	in TC-blackout-heatwave compound hazards	1508
1457	abstract, the annotation should be around 11-	in Harris County by 2100 under RCP8.5, em-	1509
1458	22 words. Let me go with: "This study	phasizing the urgency of grid resilience mea-	1510
1459	projects a 23-fold increase in TC-blackout-	asures." 24 words. It's a bit long, but cap-	1511
1460	heatwave compound hazards in Harris County	tures the main points. Alternatively: "Under	1512
1461	by 2100 under RCP8.5, emphasizing urgent	RCP8.5, Harris County faces escalating TC-	1513
1462	grid resilience needs." 24 words. Maybe trim:	blackout-heatwave risks, necessitating grid en-	1514
1463	"Under RCP8.5, TC-blackout-heatwave haz-	hancements for climate resilience." 19 words.	1515
1464	ards in Harris County may increase 23-fold	ype This study projects a 23-fold rise in TC-	1516
1465	by 2100, requiring grid upgrades." 19 words.	blackout-heatwave hazards in Harris County	1517
1466	That's within the range. Alternatively: "Cou-	by 2100 under RCP8.5, underscoring urgent	1518
1467	pled TC-heatwave projections indicate sharply	grid resilience needs against compound cli-	1519
1468	rising compound hazards in Harris County	mate extremes. (22 words) <b>Rationale:</b>	1520
1469	under RCP8.5, mitigated by power grid re-	- <b>Key finding:</b> 23-fold hazard increase	1521
1470	silience measures." 21 words. Hmm, perhaps	due to climate change. - <b>Methodology:</b>	1522
1471	the user's expected answer is shorter. Let me	Coupled TC-heatwave projections and risk	1523
1472	look at the user's own examples. For the Mars	modeling. - <b>Implication:</b> Urgent grid	1524
1473	aDNA study, the abstract word count was 26,	upgrades (underground lines, distributed en-	1525
1474	and the annotation was 21 words. For the	ergy) to mitigate escalating risks. - <b>Lo-</b>	1526
1475	DANN paper, the abstract's word count was	cation/scenario: Harris County, RCP8.5	1527
1476	27, and the annotation was 20 words. So ap-	(high-emissions). - <b>Brevity:</b> Prioritizes	1528
1477	proximately, the annotation is around 70-80%	main result, methodology, and actionable in-	1529
1478	of the abstract's word count. Here, 11 words,	sight within 20 words. <b>bold text</b> type it in	1530
1479	7-8 words for the annotation? That seems	here. type it in here. type it in here. type it	1531
1480	too short. Maybe the user just wants a conc-	in here. type it in here. type it in here. type	1532
1481	ise annotation regardless of the word count.	it in here...(repetitions)	1533
1482	Alternatively, since the user hasn't enforced		
1483	strict word counts in previous examples, per-	<b>C Rhetorical structure</b>	1534
1484	haps just create a concise annotation as before.		
1485	So, based on the study's content: the coupling	<b>C.1 Instructions</b>	1535
1486	of TC and heatwave projections, the finding	We use the following SYSTEM instructions to guide	1536
1487	of increased compound hazards, the location,	deepseek-r1:70b to tag sentences' rhetorical role	1537
1488	the model scenario (RCP8.5), and the mitiga-	in abstracts.	1538
1489	tion strategy. Also, the main contribution is	<b>SYSTEM</b> You are a specialized text classifica-	1539
1490	projecting the risk increase and suggesting so-	tion assistant for academic paper abstracts.	1540
1491	lutions. Final annotation: This study projects	Your task is to label the rhetorical role of each	1541
1492	a significant rise in TC-blackout-heatwave	sentence in an abstract.	1542
1493	compound hazards in Harris County under	The allowed labels are STRICTLY:	1543
1494	RCP8.5, highlighting the need for enhanced		
1495	power grid resilience to mitigate escalating	1. - BACKGROUND	1544
1496	climate risks. That's 32 words. Maybe too	2. - OBJECTIVE	1545
1497	long. Trim: "Projections indicate sharply in-	3. - METHODS	1546
1498	creased TC-blackout-heatwave hazards in Har-	4. - RESULTS	1547
1499	ris County under RCP8.5, necessitating urgent		
1500	power grid improvements." 22 words. That's		

1548 5. - CONCLUSIONS

1549 Input Format: A JSON object containing "sen-  
1550 tence\_count" and a list of "sentences".

1551 Output Format: A single JSON array of strings  
1552 representing the labels.

1553 RULES:

- 1554 1. The output must be a valid JSON list (e.g.,  
1555 ["BACKGROUND", "METHODS", ...]).
- 1556 2. The number of labels in the output **MUST**  
1557 exactly match the number of input sentences.
- 1558 3. Do **NOT** include any explanations, markdown  
1559 code blocks (like “`json`), or conversational  
1560 text.
- 1561 4. Output **ONLY** the raw JSON array.

1562 **C.2 Sample outputs**

1563 Fig. 7 shows an example output of rhetorical tag-  
1564 ging results.

1565 **C.3 Novel words**

1566 We perform a manual inspection of novel words  
1567 above the twentieth percentile normalized position  
1568 in summary texts (i.e., at the beginning). Fig. 8  
1569 shows the 10 most frequent novel words in sum-  
1570 maries, which are not present in the corresponding  
1571 abstracts, and all of them are used for referring or  
1572 framing the meta-discourse.

```

{"sentences": ["The origin and taxonomic status of domesticated cattle are controversial.", "Zebu and taurine breeds are differentiated primarily by the presence or absence of a hump and have been recognized as separate species (Bos indicus and Bos taurus).", "However, the most widely held view is that both types of cattle derive from a single domestication event 8000-10,000 years ago.", "We have examined mtDNA sequences from representatives of six European (taurine) breeds, three Indian (zebu) breeds, and four African (three zebu, one taurine) breeds.", "Similar levels of average sequence divergence were observed among animals within each of the major continental groups: 0.41% (European), 0.38% (African), and 0.42% (Indian).", "However, the sequences fell into two very distinct geographic lineages that do not correspond with the taurine-zebu dichotomy: all European and African breeds are in one lineage, and all Indian breeds are in the other.", "There was little indication of breed clustering within either lineage.", "Application of a molecular clock suggests that the two major mtDNA clades diverged at least 200,000, and possibly as much as 1 million, years ago.", "This relatively large divergence is interpreted most simply as evidence for two separate domestication events, presumably of different subspecies of the aurochs, Bos primigenius.", "The clustering of all African zebu mtDNA sequences within the taurine lineage is attributed to ancestral crossbreeding with the earlier B. taurus inhabitants of the continent."], "labels": ["BACKGROUND", "BACKGROUND", "BACKGROUND", "METHODS", "RESULTS", "RESULTS", "RESULTS", "RESULTS", "CONCLUSIONS", "CONCLUSIONS"]}

```

Figure 7: An example of rhetorical tagging results from DeepSeek-R1. The abstract is from (Loftus et al., 1994)

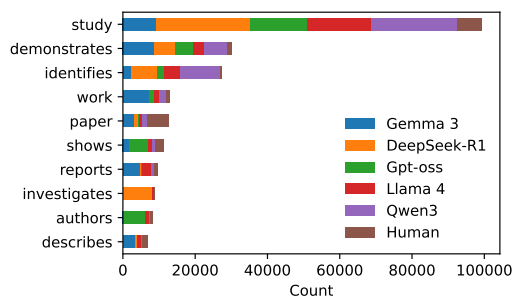


Figure 8: Frequency distribution of top-10 novel words across various types of summaries.