Can Large Language Model Summarizers Adapt to Diverse Scientific Communication Goals?

Marcio Fonseca   **Shay B. Cohen**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB
m.fonseca@ed.ac.uk, scohen@inf.ed.ac.uk

Abstract

In this work, we investigate the controllability of large language models (LLMs) on scientific summarization tasks. We identify key stylistic and content coverage factors that characterize different types of summaries such as paper reviews, abstracts, and lay summaries. By controlling stylistic features, we find that non-fine-tuned LLMs outperform humans in the MuP review generation task, both in terms of similarity to reference summaries and human preferences. Also, we show that we can improve the controllability of LLMs with keyword-based classifier-free guidance (CFG) while achieving lexical overlap comparable to strong fine-tuned baselines on arXiv and PubMed. However, our results also indicate that LLMs cannot consistently generate long summaries with more than 8 sentences. Furthermore, these models exhibit limited capacity to produce highly abstractive lay summaries. Although LLMs demonstrate strong generic summarization competency, sophisticated content control without costly fine-tuning remains an open problem for domain-specific applications.[1]

1 Introduction

Recent work on the evaluation of large language models (LLMs) has shown unprecedented performance on diverse language generation tasks, even in zero-shot settings (Clark et al., 2021). Specifically in text summarization, Goyal et al. (2022) found that human evaluators prefer summaries produced by GPT-3 over strong supervised baselines. In a similar experiment testing ten large language models, Zhang et al. (2023) concluded that instruction-tuned models performed on par with human freelance writers. Additionally, their results suggest that the evaluation of news summarization is hindered by the low quality of reference summaries. Hence, the conventional summarization
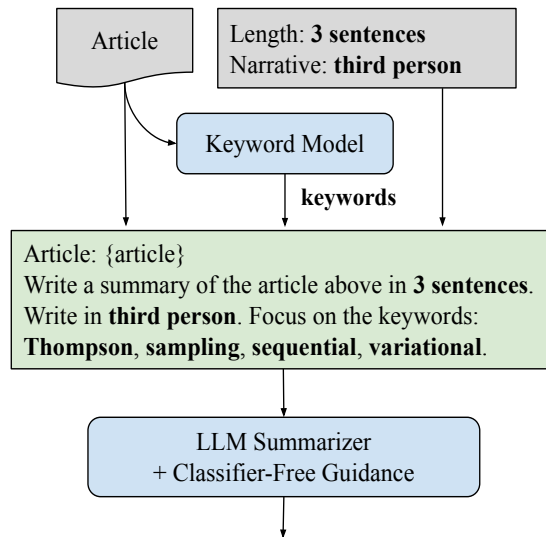
Figure 1: An overview of our controllability experiments. We expose LLM summarizers to prompts conveying communicative intentions related to *conciseness*, *narrative perspective*, and *keywords* inferred by a *keyword model*. Then, we measure how generated summaries adhere to those intentional targets.

paradigm based on fine-tuning and evaluating on reference datasets is questionable in the face of increasingly competent language models.

However, most of those studies highlighting the success of LLM summarizers are based on news articles and the validity of those claims for long technical documents deserves further investigation. In this work, we aim to fill this gap by posing the following research questions: 1) How do LLM sum-

---

[1] Code is available at https://github.com/thefonseca/scisum

maries compare to human-generated paper summaries? 2) To what extent LLM summarizers can be controlled to fulfill different goals of scientific communication? In both cases, we focus on the zero-shot setting, since the cost to fine-tune such large models is often prohibitive.

To address the first question, we design an experiment to compare human and machine summarizers, in which human summaries are judged in terms of their similarity with respect to another set of human-written reference summaries. By evaluating the lexical overlap with reference summaries from the multi-perspective scientific summarization dataset - MuP (Cohan et al., 2022), we find that human reviewers achieve lower ROUGE scores (Lin, 2004) compared to LLMs. The high quality of LLM summaries is confirmed by a human judgment experiment, where machine-generated summaries are preferred in 83% of the instances.

The strong preference for LLM summaries hints that the variability in communicative intentions (Giulianelli et al., 2023; Andreas, 2022) outweighs the usual quality criteria such as coherence, fluency, and relevance (Fabbri et al., 2021). In fact, an inspection of human summaries for the same article reveals arbitrary decisions related to conciseness and coverage of scientific aspects. We argue that those subjective decisions affect the perceived quality of summaries and that evaluation protocols should consider the *adaptability* of summarizers to diverse contexts (Jones, 1998).

Motivated by the experiment on the MuP dataset and previous literature on controllable summarization (He et al., 2022), we elect three intentional aspects: conciseness, narrative perspective (first or third person), and keyword coverage. For each of those aspects, we define two elements: *intention prompts*, and *intention control metrics*. Intentional prompts are designed to specify the task and to elicit a given intentional behavior from LLMs (e.g., *summarize this article in **5 sentences***). Finally, the control metrics assess the intentional alignment of prompts and the generated summaries.

To investigate the second research question, we evaluate our intentional prompts on the tasks of abstract generation (Cohan et al., 2018) and lay summarization (Goldsack et al., 2022), using GPT-3.5 and LLAMA-2 (Touvron et al., 2023). We find that both models can follow conciseness and narrative perspective intentions accurately to generate short abstracts for arXiv and PubMed. Furthermore, by using keyword-based intentional prompts

and classifier-free guidance (Sanchez et al., 2023), we can direct to the lexical content of summaries leading to ROUGE scores comparable to strong supervised baselines.

However, the eLife lay summarization benchmark presents a harder challenge for LLMs. Besides being highly abstractive, these lay summaries are much longer (18 sentences on average), and have a particular concept distribution emphasizing research background. Our controllability results show that intentional prompts achieve limited success in replicating these characteristics.

Overall, our experiments spanning four benchmarks indicate that LLMs are effective summarizers for long scientific documents, both in terms of lexical alignment and human preferences. However, when summarization tasks deviate from the training distribution (e.g., lay summarization), LLMs cannot consistently match the features of human summaries. Thus, LLMs do not usher the "death" of summarization (Pu et al., 2023) but the transition from reference-based evaluation to more nuanced, domain-specific evaluation protocols, as envisioned by Jones (1998) 25 years ago.

2 Guiding Summarizer Intentions

As uncertainty in intentionality is intrinsic in language generation (Giulianelli et al., 2023), we argue that summarization performance relates to the capacity of a system to adapt its behavior given not only the source documents but also the target communicative intentions. We investigate language model intent adaptability by changing the summarization context via prompting, and measuring how it affects the perceived intention in the summaries (see Figure 1 for an example). Let P_θ be a language model parameterized by θ. Then, we define summarization as a conditional sequence generation:

$$S \sim P_\theta(\cdot \mid D, p_{I_1}, \ldots, p_{I_N}),$$

where D is an input article and $p_{I_1}, \ldots, p_{I_N}$ are prompts inducing the intentions $I_1, \ldots, I_N$. In this work, we consider three types of intentions that are important to adapt scientific summaries to different goals, namely *conciseness*, *narrative perspective*, and *keyword coverage*:

- **Conciseness**: as one of the most important intentional factors in summarization, conciseness defines the compression rate between the source document and the summary. We are

interested in the ability of LLMs to follow specific conciseness instructions ranging from short abstracts (6-8 sentences) to longer lay summaries (14 or more sentences).

- **Narrative Perspective**: depending on the perspective of the summarizer, the summary is written using first or third-person narrative. In our experiments, we consider that paper abstracts from arXiv and PubMed use first-person narrative, and other summaries such as paper reviews from MuP (Cohan et al., 2022) and lay summaries from eLife (Goldsack et al., 2022) use third person narrative.

- **Keyword Coverage**: by guiding the coverage of keywords, we can indirectly manipulate the level of abstractiveness of a summary and favor simplified language, which is particularly relevant for lay summarization. Furthermore, when reference summaries are available, keyword guidance provides a mechanism to guide LLM generation using relatively small *keyword models* (Section 2.1.2) that demand fewer resources to fine-tune.

Since we specify intention guidance via prompts, we have no control over how the language models trade off their unconditional summarization behavior and the intention instructions. To address this issue, we modify the decoding process in Eq. 2 to include classifier-free guidance (CFG) weighting (Ho, 2022; Sanchez et al., 2023):

$$\hat{P}_\theta(S|D, p_I, p_\epsilon) \propto \frac{P_\theta(S|D, p_I)^\gamma}{P_\theta(S|D, p_\epsilon)^{\gamma-1}},$$

where $\gamma \geq 1$ is the guidance strength and p_ϵ is a "non-intentional" summarization prompt:

> Write a summary of the article above.

When $\gamma > 1$, the next-token probabilities are changed so that the generated summary is closer to the target intentions than its default summarization behavior elicited by the prompt p_ϵ.

To measure the adherence of summaries to intention prompts, we define reference-free *intention control* metrics $k_I(y_I, S)$ that gauge the intentional alignment of S with respect to a target value y_I. In the next sections, we detail the intention prompts p_I and control metrics k_I used in our experiments.

2.1 Intention Prompt Templates

The next component of our methods refers to the way we prompt language models to induce the in-

tentions described above. The general prompt template is defined as follows:

> Article: {text}
> $\{p_{I_1} \ldots p_{I_N}\}$

where the placeholder {text} denotes the input article and $\{p_{I_1} \ldots p_{I_N}\}$ refers to the concatenation of intention prompts, namely conciseness, narrative perspective, and keyword coverage.

2.1.1 Style Intention Prompts

Conciseness prompt We define an intention prompt for a target number of sentences $y_{\text{conciseness}}$:

> Write a summary of the article above in $\{y_{\text{conciseness}}\}$ sentences.

We choose to specify the target in sentences as it was found to be effective in previous work (Goyal et al., 2022) and in our own experiments.

Narrative perspective prompt For paper abstract generation tasks, we use the following prompt to instruct the model to write in the usual first-person plural perspective:

> Write in first person "we" when applicable.

For other tasks, we do not have to prompt the model to use third person voice, as we observe this is the default behavior of the language models we evaluate (see Tables 4 and 6).

2.1.2 Keyword Coverage Prompt

The last prompt type instructs the model to focus on a collection of keywords to generate the summary:

> Focus on the following keywords: $\{y_{\text{keywords}}\}$.

where the $\{y_{\text{keywords}}\}$ placeholder indicates a comma-separated list of terms, which are provided by a *keyword model*.

Keyword model In our experiments, we use pretrained encoder-decoder summarizers as keyword generators. Specifically, we use FACTORSUM (Fonseca et al., 2022) and BART (Lewis et al., 2020) for abstract generation and lay summarization respectively. Using these models, we generate summaries for each document in the evaluation dataset. Then, we extract part-of-speech tags for tokens in summaries using the NLTK library (Bird et al., 2009) and keep as keywords only nouns, verbs, adjectives, foreign words, and symbols. The full list of keyword POS tags is presented in Appendix B.

2.2 Intention Control Metrics

In this section, we identify the controllability metrics k_I for stylistic features and keyword coverage. Although we present summary-level definitions, in our experiments we report dataset-level metrics, that is, their average over all evaluation samples.

Conciseness In this work, we measure the conciseness of a summary S by counting the number of sentences $|S|$. Then, the conciseness controllability k_I for a summary S and a target number of sentences y_I is defined by:

$$k_{\text{conciseness}}(y_I, S) = \text{abs}(y_I - |S|),$$

where abs is the absolute difference function.

Narrative perspective We define the narrative perspective metric based on a text classifier $f_{\text{narrative}}(s)$ that maps each summary sentence $s \in S$ to the label $\hat{y}_I \in \{\text{first}, \text{third}\}$. Then, we defined the summary-level perspective controllability as the percentage of sentences written in the target perspective y_I:

$$k_{\text{narrative}}(y_I, S) = \frac{100}{|S|} \sum_{s_i \in S} \mathbb{1}[f_I(s_i) = y_I],$$

where $|S|$ is the number of sentences in S. For the classifier $f_{\text{narrative}}(s)$, we use a simple heuristic where the sentence is classified as first-person perspective if 1) it starts with a first-person pronoun FPP $\in \{$"we", "our"$\}$ or 2) if it contains the pattern ", FPP", i.e., a comma followed by a FPP.

Keyword Coverage We also measure the coverage of a target set of keywords y_I in summaries using the ROUGE-1 recall metric (Lin, 2004):

$$k_{\text{keywords}}(y_I, S) = \text{ROUGE-1}_{\text{recall}}(y_I, S).$$

3 Experiments

In this section, we present the experimental settings spanning three styles of scientific summaries: abstracts, reviews, and lay summaries.

3.1 Datasets

Multi-perspective scientific summarization The MuP dataset (Cohan et al., 2022) is a corpus of summaries extracted from publicly available scientific peer reviews[2] capturing multiple summarization perspectives for a single document. In our experiments, we use 1,060 review summaries (and their corresponding papers) from the validation set, covering topics primarily from the AI, Machine Learning, and Natural Language Processing fields.

Abstract generation We use arXiv and PubMed (Cohan et al., 2018), which are two large-scale benchmarks for abstract generation. In our experiments, we report results for 1,000 random samples from each dataset test set.

Lay summarization We also evaluate LLMs on the task of lay summarization using the eLife dataset (Goldsack et al., 2022). In contrast to reviews and abstracts, these summaries are much longer (around 18 sentences on average), more abstractive, and exhibit a strong bias towards background content, which makes them easier to parse by non-specialists. We report results on the 241 samples from the eLife test set.

3.2 Models

Supervised baselines For the abstract generation task, we use BIGBIRD (Zaheer et al., 2020), a transformer-based model that uses a sparse attention mechanism to handle long input sequences. Also, we include FACTORSUM (Fonseca et al., 2022), a model that breaks the summarization task into shorter subtasks, achieving strong performance on the arXiv and PubMed datasets. For lay summarization, we train our own version of the BART model (Lewis et al., 2020) on the eLife dataset (refer to Appendix A for training details).

LLAMA-2 A collection of pre-trained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters (Touvron et al., 2023). The intention prompts described in Section 2.1 are wrapped in a model-specific instruction-tuning prompt (see Appendix A for details). Unless otherwise stated, we report results for the *chat* variant with 7B parameters (16-bit floating point), with 4,096 maximum context tokens using nucleus sampling (Holtzman et al., 2020) with temperature 0.8 and $p = 0.95$.

GPT-3.5 A proprietary model based on INSTRUCTGPT (Ouyang et al., 2022). We use the model version gpt-3.5-turbo-0301 via the chat completion API endpoint[3]. Although we do not have access to the implementation details behind

---

[2] https://openreview.net/

[3] https://platform.openai.com/docs/api-reference/chat

commercial APIs, they were extensively studied in recent work (Zhang et al., 2023; Goyal et al., 2022). This model version is trained on data up to September 2021, and we generate summaries in December 2023.

3.3 Evaluation Metrics

In addition to the intention control metrics defined in Section 2.2, we report other relevant metrics for summarization, namely ROUGE (Lin, 2004) and summary abstractiveness (percentage of n-grams in summary that are absent in the source document). For the lay summarization task, we also report the Flesch-Kincaid Grade Level (FKGL) readability score (Kincaid et al., 1975).

4 Results and Discussion

In this section, we discuss our experimental results comparing human and LLM-generated summaries (Section 4.1), and the controllability experiments on the abstract generation and lay summarization tasks (Section 4.2).

4.1 Comparing Human and LLM Summaries: Reviewer Perspectives

We leverage the MuP dataset (Cohan et al., 2022) for a controlled comparison between human and machine summaries. First, we select documents from the validation set with more than one human-written summary and randomly choose one of the summaries to serve as a reference and a second human-written summary as a fictitious system summary. This setting puts humans in a similar condition as usual reference-based summarization benchmarks, that is, humans are evaluated on their ability to *guess* reference summaries.

As machine summarizers, we evaluate LLAMA-2 and GPT-3.5. To minimize the confounding factors related to summary length, we employ conciseness prompts (Section 2.1.1) to instruct the LLMs to generate summaries with a number of sentences such that the number of generated tokens approximates the human summaries[4]. In Table 1, we report metrics comparing LLM and human summaries for 1,060 samples from the MuP validation set.

In addition to automatic evaluation, we perform human evaluation based on 30 random samples of human and LLM-generated summaries. Similarly

---

[4]We take the number of tokens in each human summary and divide by the average number of tokens per sentence generated by LLAMA-2 and GPT-3.5, resulting in a different conciseness target for each document.

Metric	Ref	Human	LLAMA	GPT-3.5
ROUGE-1	100	35.55	**37.35**	36.39
ROUGE-2	100	8.12	**9.20**	8.51
ROUGE-L	100	20.55	**24.47**	24.02
Avg. Tokens	115	113	112	113
Third Person	99.8	99.6	99.3	100.0
Novel bigram	66.4	**67.0**	39.9	34.1

Table 1: Metrics comparing human-written, LLAMA-2-7B, and GPT-3.5 summaries to reference reviews (Ref) from the MuP validation set. We report average tokens per summary and the percentage of sentences using *third person* perspective. *Novel bigram* measures the percentage of novel bigrams in summaries.

to Goyal et al. (2022), we ask evaluators to perform blind A/B judgments based on the paper abstract and a triplet consisting of one human and two LLM summaries. Evaluators are tasked to elect from the triplet which summary (or summaries) they judge is the best/worst alternative to the paper abstract. For each answer, the evaluators need to provide a short justification. Each summary triplet is annotated by two graduate students specializing in NLP. We provide detailed task instructions in Appendix C.

ROUGE penalizes abstractiveness in human summaries The results in Table 1 show that LLMs achieve comparable albeit higher ROUGE scores compared to human reviewers, which by standard reference-based protocols, would be an indicator that LLM summaries have "superior quality." However, we observe quantitative differences between human and machine reviews across other metrics. Notably, human summaries are significantly more abstractive as measured by the percentage of novel bigrams in the summaries. Given the more creative nature of human reviews, it is expected that they offer more possibilities for lexical disagreement than a conservative summary that prioritizes content from the source document.

Humans strongly prefer LLM summaries The results in Table 2 reveal the marked preference of evaluators towards LLM summaries, which account for 83% of votes for best summary. The preference for the worst summary is more balanced, with humans and GPT-3.5 taking the majority of votes. Also, we note that Cohen's κ scores (Artstein and Poesio, 2008) are low, which indicates the high subjectivity of the task.

To understand the factors underlying human pref-

Vote	Human	LLAMA	GPT-3.5	Cohen's κ
Best ↑	16.9	40.0	**43.1**	0.06
Worst ↓	36.1	**27.9**	36.1	0.17

Table 2: Human preferences (percentage) for best/worst MuP summaries. To account for multiple summary choices, we use the weighted Cohen's κ agreement statistic (Artstein and Poesio, 2008).

Best	%	Worst	%
Coverage	**64.1**	**Coverage**	**42.9**
Coherence	12.8	Relevance	11.1
Fluency	6.4	Factuality	9.5
Conciseness	3.8	Informativeness	7.9

Table 3: Frequency of top criteria for human preferences on best and worst MuP summaries. Cohen's κ score is -0.16 and 0.1 for best and worst votes respectively.

erences, we categorize the evaluators' comments into quality criteria, as shown in Table 3. For both best and worst summary choices, the dominant reason is *coverage* of research aspects such as background or experimental results. Interestingly, *factuality* only appears as a third factor for choosing a bad summary, which is an indicator of the high quality of LLM summaries. However, we note that factuality issues might be underrepresented in our results, as the annotators do not use the original article in their judgments.

Takeaways Despite the advantage of LLMs on reference-based metrics and human evaluation, *our aim with this experiment is not to claim that LLM summaries are superior*. Instead, we emphasize the limitations of evaluation based on reference summaries, as bias in communicative intentions impacts ROUGE scores and even preferences of human annotators. Since LLM summaries exhibit human-level quality in terms of coherence, fluency, and informativeness, we argue that the real challenge for LLMs lies in the adaptability to different tasks, which we explore in the next experiment.

4.2 Controlling Style and Keyword Coverage

In this section, we gauge the performance of LLMs on benchmarks for abstract generation and lay summarization. Our goal is to compare the performance of zero-shot inference in LLMs to supervised baselines, and most importantly, to assess the flexibility of LLMs, via intention prompts, to replicate preferences such as narrative perspective

and conciseness. In addition to LLAMA-2 and GPT-3.5, we evaluate strong fine-tuned encoder-decoder baselines: BIGBIRD and FACTORSUM (Zaheer et al., 2020; Fonseca et al., 2022) for abstract generation and BART (Lewis et al., 2020) for lay summarization. Also, we report results for LLAMA-CFG, denoting LLAMA-2 with classifier-free guidance (refer to Section 2 for details).

Model	R-1	R-2	R-L	Control Metrics k_I		
				Con↓	Nar↑	Key↑
Reference	100	100	100	1.8	29.7	42.5
BIGBIRD	45.0	18.3	39.8	1.9	46.2	42.3
FACTORSUM	**48.9**	**20.1**	**43.8**	1.9	50.9	100
Prompt: *Write an abstract of the article above.*						
LLAMA	43.1	14.6	38.0	2.0	23.2	42.9
GPT-3.5	43.7	15.0	38.5	2.0	0	42.5
"Non-intentional" baseline prompt p_ϵ						
LLAMA	42.0	14.3	37.3	1.1	0.2	49.3
GPT-3.5	42.9	14.4	37.8	1.1	0.1	51.2
$p_{\text{conciseness}}$						
LLAMA	42.6	14.3	37.7	0.4	1.3	47.7
GPT-3.5	42.5	13.8	37.4	**0.3**	0.1	50.5
LLAMA-CFG	42.4	14.0	37.4	0.3	2.2	39.3
$p_{\text{conciseness}} + p_{\text{narrative}}$						
LLAMA	43.9	14.7	38.7	0.3	83.6	48.4
GPT-3.5	44.0	15.0	38.9	0.3	42.2	53.2
LLAMA-CFG	43.4	14.4	38.2	0.3	**84.5**	40.9
$p_{\text{conciseness}} + p_{\text{keywords}} + p_{\text{narrative}}$						
LLAMA	45.2	15.7	39.8	0.4	84.0	64.4
GPT-3.5	47.0	17.6	41.4	0.6	61.3	68.6
LLAMA-CFG	44.5	14.9	39.1	0.5	81.5	**74.2**

Table 4: Summarization results on the arXiv test set (1,000 samples) using different intention prompts p_I (defined in Section 2.1). We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness, narrative perspective (first person), and keyword recall. The target conciseness in $p_{\text{conciseness}}$ is 6 sentences. The list of keywords in p_{keywords} is derived from FACTORSUM summaries as described in Section 2.1.2. Results for the PubMed dataset show similar trends and are reported in Appendix D.

LLMs capture style intents We analyze the effects of style prompts compared to two baseline prompts: the non-intentional prompt p_ϵ and a more direct prompt that specifies the type of summary

(e.g., *write a lay summary of the article above*). For abstract generation, LLMs are able to follow conciseness instructions ($p_{\text{conciseness}}$) consistently, as indicated by the significant reduction in the average deviation from the target number of sentences ($k_{\text{conciseness}}$ in Table 4). In this aspect, GPT-3.5 exhibits superior performance compared to LLAMA-2, achieving a deviation of 0.28 and 0.53 sentences for arXiv and PubMed, respectively.

By adding narrative perspective prompts ($p_{\text{conciseness}} + p_{\text{narrative}}$), we observe a similar level of controllability, with the percentage of sentences in first person voice increasing from zero to 84% for LLAMA-2 ($k_{\text{narrative}}$ in Table 4). This change of perspective also results in higher similarity to reference summaries (as measured by ROUGE), which validates the effectiveness of narrative guidance. Interestingly, GPT-3.5 generates fewer sentences in first person perspective, balancing first and third voices to achieve better fluency compared to LLAMA-2 summaries.
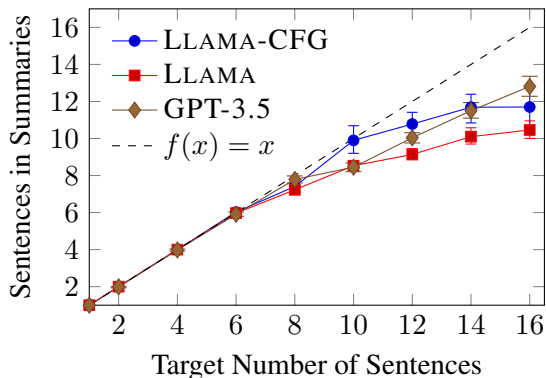

Figure 2: Number of sentences in generated summaries subject to varying conciseness targets (100 samples from eLife validation set).

Long summaries remain challenging While conciseness prompts are effective for abstract generation, LLMs cannot achieve low conciseness deviation for long lay summaries (Table 6). To evaluate the limits of conciseness guidance, we compare LLAMA-2 and GPT-3.5 summaries (100 samples from the eLife validation split) with target conciseness varying from 1 to 16 sentences (Figure 2). We find that both LLAMA-2 and GPT-3.5 reproduce conciseness intentions almost perfectly up to 6 sentences when it starts to degrade its instruction adherence. We hypothesize that longer generation targets conflict with the notion of "summary" that LLMs learn from pre-training data.

Model	R-1	R-2	R-L	Control Metrics k_I	
				Concise↓	Narrative↑
BIGBIRD	21.18	2.47	19.28	4.96	57.8
"Non-intentional" baseline prompt p_ϵ					
LLAMA	45.28	14.23	42.55	0.95	0.3
GPT-3.5	46.58	15.51	43.87	1.06	0.1
$p_{\text{conciseness}} + p_{\text{narrative}}$					
LLAMA	46.42	14.62	43.30	**0.32**	**89.6**
GPT-3.5	**47.91**	**15.94**	**45.09**	0.68	18.5

Table 5: Summarization results for 500 arXiv samples published between December 2023 and January 2024. We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness and narrative perspective (first person). The prompts p_ϵ, $p_{\text{conciseness}}$ and $p_{\text{narrative}}$ are defined in Section 2.1.

Keyword prompts improve lexical alignment We introduce the keyword coverage prompt p_{keywords} with a keyword model derived from FAC-TORSUM summaries for arXiv and PubMed. The results in Table 4 demonstrate that the keyword recall for LLAMA-2 and GPT-3.5 increases significantly with keyword guidance, leading to ROUGE scores higher than BIGBIRD summaries. Similarly, keyword prompt improves lexical alignment in lay summarization (Table 6). In this case, in addition to keyword recall, we also observe that keyword guidance results in better readability scores and higher abstractiveness (Table 7), which are desirable properties for lay summaries.

Classifier-free guidance improves intention control We report results for LLAMA-2 with classifier-free guidance (CFG), using $\gamma = 1.5$ for all prompt variants[5] (Tables 4 and 6). In all cases, we observe an improvement in intention control metrics versus LLAMA-2 with regular decoding. In Figure 2, we can clearly observe how LLAMA-CFG can follow conciseness instructions for up to 10 sentences but fail to produce longer summaries.

LLMs perform well out-of-distribution One important concern is that LLAMA-based models might simply be memorizing their training data, which likely include scientific papers that overlap with the arXiv and PubMed summarization datasets. To check this possibility, we collected 500 arti-

---

[5]We set $\gamma = 1.5$ based on the results reported by Sanchez et al. (2023) and our experiments on the validation set.

Model	R-1	R-2	R-L	Control Metrics k_I		
				Con↓	Nar↑	Key↑
Reference	100	100	100	10.8	99.5	44.1
BART	**48.1**	**13.9**	**30.5**	**2.0**	99.9	100
Write a lay summary of the article above.						
LLAMA	39.5	10.0	23.4	11.1	**100.0**	33.3
GPT-3.5	32.9	7.8	21.3	12.2	99..9	33.0
"Non-intentional" baseline prompt p_ϵ						
LLAMA	34.5	7.7	21.0	8.1	99.3	28.4
GPT-3.5	29.3	6.7	19.3	9.2	**100.0**	27.4
$p_{\textbf{conciseness}}$						
LLAMA	39.4	8.5	23.7	3.9	99.4	32.6
GPT-3.5	40.6	8.7	25.5	3.1	99.2	38.0
LLAMA-CFG	37.4	7.9	23.2	3.0	98.7	32.3
$p_{\textbf{conciseness}} + p_{\textbf{keywords}}$						
LLAMA	44.0	10.3	25.5	2.3	99.6	60.2
GPT-3.5	42.4	9.9	26.4	4.7	99.7	61.5
LLAMA-CFG	42.9	9.6	25.1	3.1	99.7	**81.5**

Table 6: Summarization results on the eLife test set (241 samples) using different intention prompts p_I (defined in Section 2.1). We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness, narrative perspective (third person), and keyword recall. The target conciseness in $p_{\text{conciseness}}$ is 14 sentences. The list of keywords in $p_{\textbf{keywords}}$ is derived from BART summaries as described in Section 2.1.2.

Model	FKGL↓	Novel Bigram↑
Reference	10.8	**66.6**
BART	**10.2**	57.8
Write a lay summary of the article above.		
LLAMA	10.9	47.6
"Non-intentional" baseline prompt p_ϵ		
LLAMA	15.69	31.5
$p_{\textbf{conciseness}}$		
LLAMA	14.8	28.4
LLAMA-CFG	14.1	27.0
$p_{\textbf{conciseness}} + p_{\textbf{keywords}}$		
LLAMA	10.9	47.2
LLAMA-CFG	12.8	55.8

Table 7: Effects of intention prompts and classifier-free guidance (CFG) on readability (FKGL) and abstractiveness (novel bigrams) on the eLife test set (241 samples).

cles (category cs.CL) submitted to arXiv between December 2023 and January 2024 and removed all the information before the introduction section (including title, abstract, and other information). Then, we evaluate the same models without further fine-tuning, using the narrative perspective and conciseness prompts. The results in Table 5 show that instruction-tuned models achieve even higher ROUGE scores (compared to Table 4), suggesting that memorization is not a cause for their good summarization performance. In contrast, we observe a pronounced loss of performance for BIGBIRD, which tends to generate repeated sentences.

5 Related Work

Our work fits in the context of recent research of large language models applied to summarization, mostly for news articles (Goyal et al., 2022; Zhang et al., 2023). In this work, we explore summarization in the scientific domain and its specific challenges related to document length (Beltagy et al., 2020; Fonseca et al., 2022) and technical writing style. Furthermore, we consider our work as a contribution towards understanding production variability by humans and language models, which was explored on tasks such as translation, story generation, and open-domain dialogue by Giulianelli et al. (2023).

Previous work has explored summarization controllability, including length control and entity coverage (He et al., 2022; Fan et al., 2018). In this work, we explore similar summary features and expand our analysis to the challenging setting of lay summarization. This task requires the production of long, highly abstractive summaries that cover more background information. Several other challenging scientific summarization tasks (e.g., focusing on paper contributions and related literature) would lead to promising research directions.

Finally, our work is related to attempts to improve the prompt-adherence in generation tasks. Pu and Demberg (2023) apply prompt-based methods to control summarization according to target audience (expert vs. layman) and style (formal vs. informal). Kumar et al. (2022) propose a non-autoregressive generation method to introduce soft and hard constraints, including adherence to keywords. In a different direction, (Sanchez et al., 2023) showed that Context-Free Guidance (CFG)

(Ho and Salimans, 2022), a technique originally used in text-to-image models, improves the performance of language models across several tasks. While Sanchez et al. (2023) applied CFG to tasks such as question answering, code generation and translation, our work is the first to explore this technique for summarization controllability.

6 Conclusion

Our experimental results on four scientific summarization benchmarks confirm that large language models are effective summarizers, as measured by conventional lexical overlap metrics and human preferences. Moreover, we find that LLMs can follow intentional prompts for style and keyword coverage, especially for short summaries. This summarization controllability can be further improved using simple decoding changes such as classifier-free guidance. However, longer summaries that require a higher level of abstractiveness are not easily achievable with our prompting techniques. Our findings suggest that analyzing summaries as an expression of communicative intentions leads to informative and actionable insights for future model improvements, where traditional reference-based evaluation shows its limits.

Limitations

Model scale and proprietary APIs   Our experiments operate at model scales up to 7 billion parameters (LLAMA-2), and we expect to observe more performance gains for larger models (Wei et al., 2022). By including results from the OpenAI API, we can get an estimation of the level of performance of larger models, although we do not have access to information about the number of parameters of the underlying model and whether it includes extra machinery on top of language modeling. Lastly, we limit our experiments to zero-shot summarization as it represents a relevant application of LLMs, given the high costs involved in obtaining high-quality reference summaries.

Excluded models and reproducibility issues Despite our best prompting efforts, we could not make encoder-decoder instruction-tuned models T0 (Sanh et al., 2022) and FLAN-T5 (Chung et al., 2022) generate long summaries (at least six sentences), which makes them unsuitable for comparison. We also tried other models that claim high ROUGE performance but do not provide public code (Pang et al., 2023), fine-tuned checkpoints for

the scientific papers (Guo et al., 2022; Xiong et al., 2022), or reproducible results with the provided resources (Phang et al., 2022).

Human evaluation scope   Our human evaluation experiment uses a limited number of samples from a dataset of computer science papers, and our conclusions might not generalize to other domains. Furthermore, we limit our analysis of variability in communicative intents from the point of view of summarizers, but we believe human evaluators are also subject to these factors. Future investigation could explore how different annotation conditions and guidelines potentially bias human judgments and to what extent human evaluation is still valid as the gold standard.

Other summarization domains   Related work in news summarization explores the generation of summaries based on guidelines for topic and entity coverage (Ahuja et al., 2021; Maddela et al., 2022). We believe that the methods we presented in this paper could be applied to these types of coverage requirements. However, given that news summarization has been extensively investigated in previous work, we choose to perform a more comprehensive evaluation in the scientific domain.

Ethics Statement

The strong zero-shot performance of large language models dramatically reduces the costs of deploying high-quality summarization services. As a consequence, we expect that much of the information available on social media and other sources will be abridged versions generated by commoditized summarization systems. In this context, special attention is needed regarding existing issues of generative language models such as hallucination (McKenna et al., 2023). Crucially, we have shown that language models can adapt their outputs according to specific communication intentions, which might bias the summary contents towards the (potentially harmful) beliefs and desires of the actors behind those systems.

Acknowledgements

References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2021. Aspectnews: Aspect-oriented summarization of news documents. *ArXiv preprint*, abs/2110.08296.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Arman Cohan, Guy Feigenblat, Tirthankar Ghosal, and Michal Shmueli-Scheuer. 2022. Overview of the first shared task on multi perspective scientific document summarization (MuP). In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 263–267, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Marcio Fonseca, Yftah Ziser, and Shay B. Cohen. 2022. Factorizing content and budget decisions in abstractive summarization of long documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6341–6364, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan Ho. 2022. Classifier-free diffusion guidance. *ArXiv preprint*, abs/2207.12598.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Karen Sparck Jones. 1998. Automatic summarising: factors and directions. *arXiv preprint cmp-lg/9805011*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of

new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training, University of Central Florida.*

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bo Pang, Erik Nijkamp, Wojciech Kryscinski, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2023. Long document summarization with top-down and bottom-up inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1267–1284, Dubrovnik, Croatia. Association for Computational Linguistics.

Jason Phang, Yao Zhao, and Peter J Liu. 2022. Investigating efficiently extending transformers for long input summarization. *ArXiv preprint*, abs/2208.04347.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *ArXiv preprint*, abs/2309.09558.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. 2023. Stay on topic with classifier-free guidance. *ArXiv preprint*, abs/2306.17806.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multi-task prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *ArXiv preprint*, abs/2206.07682.

Wenhan Xiong, Anchit Gupta, Shubham Toshniwal, Yashar Mehdad, and Wen-tau Yih. 2022. Adapting pretrained text-to-text models for long text sequences. *ArXiv preprint*, abs/2209.10052.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *ArXiv preprint*, abs/2301.13848.

A Fine-tuning and Generation Details

In this section, we provide details on the summary generation implementation. The main generation parameters are provided in Table 8. LLAMA-2 and BIGBIRD checkpoints are downloaded from the HuggingFace model hub[6].

For abstract generation, we use the FACTORSUM checkpoints provided by Fonseca et al. (2022)[7]. We perform inference with a fixed budget of 6 and 8 sentences for arXiv and PubMed respectively, and content guidance from BIGBIRD summaries.

For lay summarization, we fine-tune our own BART-base model[8] on the eLife dataset[9]. We use the Hugging Face summarization training script (commit 5c67682[10]) with the hyperparameters listed in Table 9. We choose the checkpoint with the best ROUGE-2 (F1) score on the 241 of the validation set. The fine-tuning process takes about 8 GPU hours (2 Nvidia GTX 1080 12GB GPUs).

A.1 Prompt Details

In addition to the prompt specification provided in Section 2.1, we provide model-specific prompt details in the section. For LLAMA-2, we use the following prompt:

```
[INST] {instruction} [/INST]
```

where {instruction} is an intention prompt from Section 2.1.

B Keyword Model Details

As described in Section 2.1.2, we perform part-of-speech tagging on tokenized summaries by FACTORSUM and BART. Specifically, we use the averaged_perceptron_tagger provided by NLTK (Bird et al., 2009)[11]. Then, we select only tokens tagged as adjectives, nouns, verbs, symbols, and foreign words. The complete set of keyword POS tags is listed in Table 10.

C Human Evaluation Task Description

Human evaluators are presented with the following short description of the tasks:

---

[6]https://huggingface.co/models
[7]https://github.com/thefonseca/factorsum
[8]https://huggingface.co/facebook/bart-base
[9]https://huggingface.co/datasets/tomasg25/scientific_lay_summarisation
[10]https://github.com/huggingface/transformers/commits/5c67682b169576c4859700d551090ff79d450a9a/examples/pytorch/summarization/run_summarization.py
[11]www.nltk.org/api/nltk.tag.perceptron.html

BIGBIRD	
Checkpoint (arXiv)	bigbird-pegasus-large-arxiv
Checkpoint (PubMed)	bigbird-pegasus-large-pubmed
Number of parameters	576.9M
Max source length	3072
Generation beams	5
Length penalty	0.8
FACTORSUM	
Checkpoint	bart-base
Number of parameters	139.4M
Max source length	1024
Max source length	128
Generation beams	4
Length penalty	1.0
LLAMA-2	
Checkpoint	Llama-2-7b-chat-hf
# parameters	7B
Max context length	2048
Parameter type	float16
Nucleus temperature	0.8
Nucleus top-p	0.95
GPT-3.5	
Model	gpt-3.5-turbo-0301
temperature	1
top_p	1
presence_penalty	0
frequency_penalty	0
All models (except FACTORSUM)	
Max target length (arXiv, MuP)	256
Max target length (PubMed, eLife)	512

Table 8: Summary generation details and parameters.

Checkpoint	bart-base
Epochs	30
Batch size	4
Optimizer	Adam (β_1=0.9; β_2=0.999)
Learning rate	5×10^{-5}
Weight decay	linear
Max. target length	1024
Validation metric	ROUGE-2 (F1)

Table 9: Fine-tuning parameters for BART-base on the eLife summarization dataset.

POS Tag	Description
FW	Foreign word
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
SYM	Symbol
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Table 10: Part-of-speech tags used to filter keywords for summary guidance. Tag descriptions are taken from the Penn Treebank Project (Marcus et al., 1993).

The goal of this study is to evaluate machine-generated summaries of scientific articles. Each reference article abstract will be presented with 3 alternative summaries. Your task is to identify which of the alternatives is the best according to your personal preferences and experience reading papers. The criteria may include good coverage of the abstract content, factuality issues, and linguistic fluency, among others. It is expected that in some cases the alternative summaries cover details that cannot be verified based on the reference abstract. You can disregard the factuality of those details if you consider it appropriate.

A sample task with a reference abstract, the 3 candidate summaries, and the questions is shown in Table 11.

D PubMed Summarization Results

We report summarization and intention control metrics for the PubMed dataset in Table 12.

E Sample Summaries

In Tables 13 to 17, we provide sample summaries for the MuP, arXiv, PubMed, and eLife datasets.

Reference Abstract ([Zhou et al., 2020](#))

Learning disentangled representations is regarded as a fundamental task for improving the generalization, robustness, and interpretability of generative models. However, measuring disentanglement has been challenging and inconsistent, often dependent on an ad-hoc external model or specific to a certain dataset. To address this, we present a method for quantifying disentanglement that only uses the generative model, by measuring the topological similarity of conditional submanifolds in the learned representation. This method showcases both unsupervised and supervised variants. To illustrate the effectiveness and applicability of our method, we empirically evaluate several state-of-the-art models across multiple datasets. We find that our method ranks models similarly to existing methods. We make our code publicly available at https://github.com/stanfordmlgroup/disentanglement.

Summary A

The article proposes a new evaluation metric for disentanglement in generative models, which is based on the concept of persistent homology. The metric measures the topological similarity between submanifolds conditioned on different latent dimensions in the generative model, and can be used to evaluate the disentanglement of different models and datasets. The proposed metric provides a more reliable and robust way of evaluating disentanglement compared to existing methods, and can be applied across different model architectures and datasets without requiring additional training or hyperparameter tuning.

Summary B

The paper proposes a novel metric for evaluating disentanglement by taking a manifold-topological perspective on the representations learnt. The key insight is that for a disentangled representation, when we fix a certain factor of variation at different values the topology of the conditional sub-manifolds should be similar. Using this insight the paper proposes a metric for disentangling which does not require annotations of the factors of variation and is more general than previous such tests.

Summary C

Researchers have proposed an unsupervised method for evaluating the level of disentanglement in deep generative models. This evaluation metric uses topology and density analysis of samples to measure the degree of topological similarity across submanifolds conditioned on disentangled latent dimensions versus entangled ones. The approach shows consistent results across 10 models and three datasets, without relying on training additional models or introducing a dataset-specific preprocessing step.

Questions

1. Given the reference abstract above, which of the alternative summaries do your prefer the most? You can select multiple if there are equally good summaries.

2. Justify your selection in the text box below. A possible reason could be "Summary A better represents the motivation of the paper and is more coherent."

3. Which summary is the worst? (Like the previous question, you can choose multiple if no summary is clearly worse than the others)

4. Justify your choice in the text box below. A possible reason could be "Summary B presents non-factual research results."

Table 11: Sample annotation task with a reference abstract, 3 candidate summaries, and questions for the evaluators.

Model	PubMed					
	R-1	**R-2**	**R-L**	Control Metrics k_I		
				Conciseness↓	Narrative↑	Keywords↑
Reference	100	100	100	2.51	8.8	43.97
BIGBIRD	44.73	19.13	40.55	3.55	10.1	48.82
FACTORSUM	**47.25**	**19.99**	**43.25**	1.26	10.5	100
Baseline prompt: *Write an abstract of the article above.*						
LLAMA	44.00	15.86	39.56	3.09	4.0	44.54
GPT-3.5	43.88	16.08	39.60	2.83	0	42.50
"Non-intentional" baseline prompt p_ϵ: *Write a summary of the article above.*						
LLAMA	43.43	15.48	39.01	2.27	0	44.06
GPT-3.5	41.75	14.73	37.58	2.68	0	43.25
$p_{\text{conciseness}}$: *Write a summary of the article above in {$y_{conciseness}$} sentences.*						
LLAMA	43.41	15.18	39.17	1.03	0	44.27
GPT-3.5	43.85	15.44	39.61	**0.53**	0	49.73
LLAMA-CFG	43.42	15.12	39.16	0.72	0.2	41.50
$p_{\text{conciseness}} + p_{\text{narrative}}$: *Write a summary of the article above in {$y_{conciseness}$} sentences. Write in **first person "we"** when applicable.*						
LLAMA	44.05	15.83	39.64	1.02	62.7	45.14
GPT-3.5	44.10	15.61	39.83	0.68	22.3	50.40
LLAMA-CFG	43.69	15.68	39.27	0.79	**70.6**	43.76
$p_{\text{conciseness}} + p_{\text{keywords}} + p_{\text{narrative}}$: *Write a summary of the article above in {$y_{conciseness}$} sentences. Focus on the following keywords: {$y_{keywords}$}. Write in **first person "we"** when applicable.*						
LLAMA	44.64	16.50	40.15	1.37	64.5	66.16
GPT-3.5	46.94	18.50	42.41	1.15	30.6	69.60
LLAMA-CFG	44.11	15.94	39.67	1.43	67.9	**76.36**

Table 12: Summarization results on the PubMed test sets (1,000 samples) using different intention prompts p_I. We report ROUGE (F1) and intention control metrics $k_I(y_I, S)$ for conciseness, narrative perspective (first person), and keyword recall. The target conciseness $y_{\text{conciseness}}$ is 8 sentences. The list of keywords y_{keywords} is derived from FACTORSUM summaries as described in Section 2.1.2.

Human Summary 1 (Reference)

This paper presents a fast method for solving certain classes of optimal control problems, by implementing a value iteration (VI) algorithm that takes advantage of convex duality. The basic idea is to take advantage of the fact that for two functions f_1, f_2, we have $(f_1 \square f_2)^* = f_1^* + f_2^*$, where f^* denotes the Legendre-Fenchel transform of f, and $(f_1 \square f_2)(x) = \inf\{f_1(x_1) + f_2(x_2) : x_1 + x_2 = x\}$ denotes the infimal convolution of f_1, f_2. Thus, the slow $\inf$ operation can be replaced with a fast addition operation. The paper provides analyses of the convergence, time complexity, and error of the algorithm. When compared to a "naive" VI algorithm in which each iteration takes $O(XU)$ (where X, U are the grid sizes of the state and input spaces, respectively), the algorithm presented in this paper can achieve $O(X + U)$. Finally, the paper provides a numerical example implemented in MATLAB.

Human Summary 2

The paper develops an approximate value iteration method for an infinite-horizon, discounted-cost Markov Decision Processes (MDPs) that satisfy a given set of regulatory assumptions. The idea is to work on a dual space that replaces the DP value function by a reformulation written in terms of biconjugate operations. The authors present an algorithm that applies the conjugate operators iteratively, and present convergence and other structural results associated with the resulting errors. Numerical results compare the proposed approach against a traditional value iteration and a variant that generated dynamic discretization grids.

Prompt: *Write a summary of the article above in 4 sentences.*

LLAMA-2 Summary

The article discusses the problem of optimal control of continuous-state systems with discrete-time dynamics and stochastic cost. The authors propose a new algorithm, called Conjugate Value Iteration (ConjVI), which combines the power of the DP operator with the efficiency of discrete-time methods to solve this problem. ConjVI is based on a modified version of the Discrete Dual Domain (d-CDP) operator, which allows for a more efficient computation of the value function. The authors provide theoretical guarantees on the convergence of ConjVI and analyze its error, showing that it can achieve a linear time complexity of O(X+U) in each iteration, where X and U are the cardinalities of the discrete state and input spaces, respectively.

GPT-3.5 Summary

Value iteration (VI) is an algorithm used for problems in reinforcement learning (RL) and optimal control formulated as Markov decision processes (MDPs). However, VI suffers from a high computational cost for large-scale finite state spaces. A solution to this problem is to incorporate function approximation techniques and compute the output of the DP operator for a finite sample of the underlying continuous state space. The article proposes the conjugate VI (ConjVI) algorithm based on a modified version of the d-CDP operator that extends existing results in three directions for solving optimal control problems of discrete-time systems with a continuous state-input space.

Table 13: Sample summaries for Kolarijani et al., 2021 (MuP validation set).

Reference Abstract

We extend the scalar sector of the neutrinophilic two-Higgs-doublet model, where small masses of Dirac neutrinos are obtained via a small vacuum expectation value of the neutrinophilic -doublet scalar field which has a Yukawa interaction with only right-handed neutrinos. A global symmetry is used for the neutrinophilic nature of the second -doublet scalar field and also for eliminating Majorana mass terms of neutrinos. By virtue of an appropriate assignment of the -charges to new particles, our model has an unbroken symmetry, under which the lightest -odd scalar boson can be a dark matter candidate. In our model, is generated by the one-loop diagram to which -odd particles contribute. We briefly discuss a possible signature of our model at the LHC.

Prompt: *Write a summary of the article above in 6 sentences.*
Write in first person "we" when applicable.

LLAMA-2 Summary

We present a new model for the origin of neutrino masses, which is based on the neutrino Yukawa coupling and the Higgs mechanism. We extend the Standard Model with a new scalar field, which is a singlet under the SM gauge group and has a non-zero yukawa coupling to neutrinos. This new scalar field generates neutrino masses through the Higgs mechanism, and its vev generates a small neutrino mass matrix. We show that the scenario of spontaneous breaking of the @xmath2 symmetry is not allowed by the measurement of the invisible decay of the @xmath30 boson. We also discuss the possibility of dark matter candidates being involved in the loop diagram, which can generate a small vev for the new scalar field.

GPT-3.5 Summary

The article discusses the possibility that neutrinos may not be Majorana fermions but instead Dirac fermions. The neutrinophilic two Higgs doublet model is introduced as a new physics model where neutrinos are regarded as Dirac fermions. This model explains the smallness of neutrino masses by a tiny vacuum expectation value of the neutrinophilic scalar without extremely small yukawa coupling constants for neutrinos. The article extends the neutrinophilic two Higgs doublet model to introduce a dark matter candidate, which may suppress the soft-breaking parameter by extending the model. The new particle discovered at the LHC is likely to be the SM Higgs boson, and may provide information on the origin of particle masses. Collider experiments can test the predictions of this model, which could not be spoiled by certain assumptions.

Table 14: Sample summaries for Kanemura et al., 2013 (arXiv test set).

Reference Abstract

This study examines the composition and activity of the planktonic community during the polar night in the high Arctic Kongsfjord, Svalbard. Our results are the first published evidence of bioluminescence among zooplankton during the Arctic polar night. The observations were collected by a bathyphotometer detecting bioluminescence, integrated into an autonomous underwater vehicle, to determine the concentration and intensity of bioluminescent flashes as a function of time of day and depth. To further understand community dynamics and composition, plankton nets were used to collect organisms passing through the bathyphotometer along with traditional vertical net tows. Additionally, using a moored bathyphotometer closed to the sampling site, the bioluminescence potential itself was shown not to have a diurnal or circadian rhythm. Rather, our results provide evidence for a diel vertical migration of bioluminescent zooplankton that does not correspond to any externally detectable changes in illumination.

Prompt: *Write a abstract for the article above in 8 sentences.*
Write in first person "we" when applicable.

LLAMA-2 Summary

We conducted a study on plankton distribution and bioluminescence in the polar night at Kongsfjorden, Svalbard, during January 2010. We used a remotely operated vehicle (ROV) equipped with bioluminescence, depth, and temperature sensors to collect data on plankton abundance and distribution. Our results showed that bioluminescence potential was detected throughout the water column, with higher intensity at depth during the day and increased surface bioluminescence at night. We found significant differences in acoustic backscatter coefficients between day and night, indicating the presence of zooplankton migration. Our plankton enumerations showed an increase in the majority of the most abundant zooplankton taxa at night, including Pseudocalanus spp., which are known to be present throughout the year in this region. We also found that Metridia lucens, Metridia longa, Oncaea borealis, Thysanoessa inermis, and Thysanoessa longicaudata likely contributed to the increase in high-intensity bioluminescent flashes at 15 and 45 m during the night. Our results suggest that there is a coordinated movement of biomass indicative of vertical migration in this region.

GPT-3.5 Summary

The article discusses the importance of bioluminescence and other adaptations in marine environments. Bioluminescence has evolved independently over 40 times and has various functions including defensive functions such as counter-illumination and burglar alarm and offensive mechanisms such as prey attraction and intra-specific communication. Some adaptations involve entering a dormant state and overwintering at depth, such as seen in the copepods calanus glacialis and c. hyperboreus. The study aimed to characterize plankton abundance and distribution patterns during a time of year rarely studied, using vertical net tows and autonomous underwater vehicle (AUV) surveys. The AUV was fitted with upward and downward facing RD instruments, a CTD, and a bioluminescence bathyphotometer. The data collected off the coast of Svalbard showed bioluminescence was detected throughout the water column both day and night, and higher bioluminescence was present at depth during the day and increased surface bioluminescence at night. Vertical net hauls showed that there was an increase above 60 meters in the majority of the most abundant zooplankton taxa at night, indicating dvm. Lastly, plankton enumerated from the > 20m net collection of the bp exhaust suggests that during the day, the greatest biomass occurred at 45m and was dominated by copepod nauplii, copepod eggs, and the tintinnid acantostomella norvegica.

Table 15: Sample summaries for Berge et al., 2012 (PubMed test set).

Reference Lay Summary

Imagine a gymnastics competition in which participants take turns to cartwheel and somersault across the floor. The routines on display comprise sequences of precisely timed movements learned through practice. This is also true for many of the actions we perform every day, such as reaching for a cup of coffee. A region of the brain called the cerebellum helps us learn sequences of movements. But how does it do this? To find out, Khilkevich et al. came up with a new version of an old experiment. Rabbits were first trained to blink their eye in response to a specific external cue. This type of learning, called associative learning, has been shown before in the cerebellum. But Khilkevich et al. wondered whether the cerebellum could also use internal feedback signals from the eyeblink as a cue to learn the next movement? If so, this might explain how the cerebellum can chain movements together in a sequence. As predicted, Khilkevich et al. found that rabbits could learn to blink their eye in response to an initial signal, and then blink again in response to the first blink. Control experiments confirmed that the second eyeblink was coupled to the first, and not to the original cue. Moreover, on many trials the rabbits showed a third and even fourth eyeblink. This is because feedback signals from the first, second or third blink were the same. Thus, the feedback signals from the first blink triggered the second blink, feedback from the second triggered the third, and so forth. Rabbits could also learn to use a blink of the left eye as a cue for a blink of the right eye. Similar patterns of neuronal activity accompanied each blink, suggesting that the same mechanism generated them all. The cerebellum can thus use feedback from one movement as a cue to learn the proper timing of the next movement in a sequence. A key question is whether this mechanism of sequence learning extends beyond movement. The cerebellum has extensive connections to the brain's outer layer, the cortex, including many areas involved in cognition. Future experiments should test whether the cerebellum might help guide sequences of cortical activity during cognitive tasks.

GPT-3.5 Summary
Prompt: *Write a summary of the article above in 14 sentences.*

The article focuses on how the cerebellum, which is responsible for accurate movement execution and learning, can learn and execute movement sequences. Previous studies have shown that cerebellar pathology leads to severe deficits in the execution and learning of sequences. To test the sufficiency of feedback signals from a movement as a cue for the next component in a sequence, the researchers trained rabbits using electrical stimulation of mossy fibers as a cue. Once subjects acquired a sequence of conditioned eyelid responses, the researchers tested whether the feedback information about the first component is sufficient for the cerebellum to use it as a new "CS" to learn the subsequent CR. The results show that the cerebellum can learn to chain together a sequence of inter-connected movement components by using feedback signals from one component to serve as a cue for the next component in the sequence. In vivo recordings from cerebellar cortex showed that Purkinje cell activity relates as strongly to sequences as it relates to single-component eyelid responses. The experiments suggest that the feedback signals are strong enough to serve as cues for the cerebellum to learn subsequent movements. The authors suggest that cerebellar mechanisms of learning single movement components can be extended to learning of movement sequences, providing a general framework for how the cerebellum can use feedback signals to learn to chain together correctly timed responses to produce a movement sequence. The researchers emphasize that most human movements are comprised of sequences, and movement sequences as well as multi-joint movements are sensitive to cerebellar dysfunction. The article provides insights into how researchers can improve the treatment of movement disorders involving the cerebellum.

Table 16: GPT-3.5 summary for Khilkevich et al., 2018 (eLife test set).

Prompt: *Write a summary of the article above in 14 sentences.*
LLAMA-2 Summary

Most movements are composed of sequences, and the cerebellum is involved in learning and executing accurate movements. The cerebellum has been linked to learning and execution of movement sequences, and studies have shown that cerebellar dysfunction can lead to deficits in sequence learning or the inability to learn sequences at all. However, most of what we know about cerebellar mechanisms of learning comes from studies using single-component movements. Three possible ways of implementing cerebellar learning of movement sequences have been proposed, and one of these ways involves the use of feedback signals from one movement component to learn the next component. To test the hypothesis that the cerebellum can use feedback signals from a movement to learn the next movement in a sequence, researchers trained rabbits using electrical stimulation of mossy fibers as a cue. The cerebellum can learn to chain together a sequence of interconnected movement components by using feedback signals from one component to serve as a cue for the next component in the sequence. The design of most experiments does not permit distinguishing between different possibilities of sequence learning, but the use of eyelid conditioning to test the sufficiency of feedback signals (FS) for sequence learning provided direct evidence that the cerebellum can learn to use FS to chain together movement components. In the experiment, the first movement was a conditioned eyelid response (CR) elicited by a sensory stimulus, and the second movement was a sequence of movements produced by the same muscle or different muscle groups used for the different components of the sequence. The results showed robust learning of later components, including sequences of movements produced by the same muscle and sequences with different muscle groups used for the different components. The in vivo recordings from cerebellar cortex showed that Purkinje cell activity related as strongly to the latter components of sequences as it has been shown to relate to single-component eyelid responses. These findings provide a general framework for how the cerebellum can use feedback signals to learn to chain together appropriately timed responses to produce a movement sequence. The results demonstrate the importance of cerebellar learning mechanisms in mediating the acquisition of movement sequences and suggest that the cerebellum can use feedback signals from a movement to learn the next movement in a sequence.

Table 17: LLAMA-2 summary for Khilkevich et al., 2018 (eLife test set).