Evaluating Large Language Models on Wikipedia-Style Survey Generation

Fan Gao[1,9], **Hang Jiang**[2], **Rui Yang**[3], **Qingcheng Zeng**[4], **Jinghui Lu**[6], **Moritz Blum**[5], **Dairui Liu**[7], **Tianwei She**[8], **Yuang Jiang**[6], **Irene Li**[1,6]

[1]University of Tokyo, [2]MIT Center for Constructive Communication, [3]Duke-NUS Medical School, [4]Northwestern University, [5]Bielefeld University, [6]Smartor.me, [7]University College Dublin, [8]Moveworks, [9]Tokyo Institute of Technology

fangao0802@g.ecc.u-tokyo.ac.jp, ireneli@ds.itc.u-tokyo.ac.jp

Abstract

Educational materials such as survey articles in specialized fields like computer science traditionally require tremendous expert inputs and are therefore expensive to create and update. Recently, Large Language Models (LLMs) have achieved significant success across various general tasks. However, their effectiveness and limitations in the education domain are yet to be fully explored. In this work, we examine the proficiency of LLMs in generating succinct survey articles specific to the niche field of NLP in computer science, focusing on a curated list of 99 topics. Automated benchmarks reveal that GPT-4 surpasses its predecessors, inluding GPT-3.5, PaLM2, and LLaMa2 by margins ranging from 2% to 20% in comparison to the established ground truth. We compare both human and GPT-based evaluation scores and provide in-depth analysis. While our findings suggest that GPT-created surveys are more contemporary and accessible than human-authored ones, certain limitations were observed. Notably, GPT-4, despite often delivering outstanding content, occasionally exhibited lapses like missing details or factual errors. At last, we compared the rating behavior between humans and GPT-4 and found systematic bias in using GPT evaluation.

1 Introduction

Recently, large language models (LLMs) have attracted significant attention due to their strong performance on general natural language processing (NLP) tasks (Shaib et al., 2023; Feng et al., 2022). Especially, the GPT family (Brown et al., 2020) shows great ability in various applications. While it has been demonstrated that they perform well in many general tasks, their effectiveness in domain-specific tasks continues to be under scrutiny (Tian et al., 2023). Specifically, the text produced by LLMs can sometimes exhibit issues like creating false information and hallucination (Zhao et al., 2023; Yang et al., 2024a).

In the context of scientific education, automatic survey generation aims to employ machine learning or NLP techniques to create a structured overview of a specific concept (Sun and Zhuge, 2022; Li et al., 2022; Yang et al., 2024b). Automating this process not only alleviates the manual effort but also ensures timely updates at a reduced cost. A common approach involves an initial information retrieval phase to select pertinent documents or sentences based on the query topic. This is followed by a summarization or simplification phase to produce the final survey (Jha et al., 2013; Li et al., 2022). While LLMs have the potential to be an alternative method for writing scientific surveys, their effectiveness and limitations are not yet thoroughly investigated.

Existing work focuses on applying LLMs to similar scenarios, including aiding scientific writing (Shen et al., 2023; Altmäe et al., 2023), question-answering with scientific papers (Tahri et al., 2022), writing paper reviews (Liang et al., 2023), and answering quiz or exam questions (Song et al., 2023; Wang et al., 2023). This study pushes the boundary of this research area as the first to evaluate the capability of LLMs in generating education surveys within the scientific domain of NLP (Li et al., 2022). Our primary objective is to understand whether LLMs can be used to explain concepts in a more structured manner. To this end, we aim to answer the following research questions (RQs):

- **RQ1**: How proficient are LLMs in generating survey articles on NLP concepts?
- **RQ2**: Can LLMs emulate human judgment when provided with specific criteria?
- **RQ3**: Do LLMs introduce a noticeable bias in evaluating machine-generated texts compared to human-written texts?

We empirically conduct experiments on LLaMa2 (Touvron et al., 2023), PaLM2 (Anil et al., 2023),

Findings of the Association for Computational Linguistics: ACL 2024, pages 5405–5418

Zero-Shot

Generate a survey about <Topic>. There should be five sub-sections: Introduction, History, Key Ideas, Variations and Applications. Each sub-section should contain 50-150 words.

One-Shot

Example survey:
<INTRODUCTION> Word2Vec is one of the most popular tools to learn word embeddings using shallow neural networks. It first constructs a vocabulary from the training text data and then learns word embeddings....
<HISTORY> Word2vec was developed by a group of researchers headed by Tomas Mikolov at Google. Machine learning models take vectors as input, ...
<KEY IDEAS> Word2Vec converts words into vector forms such that similar meaning words appear together and dissimilar words are located far away...
<USES/APPLICATIONS> Gensim provides the Word2Vec class for working with a Word2Vec model. Training your own word vectors can take a long time and uses lots of memory...
<VARIATIONS> Word embeddings is an active research area trying to figure out better word representations than the existing ones...
Generate a survey about <Topic>. There should be five sub-sections: Introduction, History, Key Ideas, Variations and Applications. Each sub-section should contain 50-150 words.

With Prompt

Generate a survey about <Topic>. There should be five sub-sections: Introduction, History, Key Ideas, Variations and Applications. Each sub-section should contain 50-150 words. The following is the guideline for each section:
SECTION 1: INTRODUCTION Describe what the topic is (a method, a model, a task, a dataset), which field/subfield it is part of, quick overview of applications and motivation behind concept and related ideas)
SECTION 2: HISTORY Describe when or by who the topic was introduced, in what context, what problems it addresses.
SECTION 3: KEY IDEAS Describe in greater depth (could provide some mathematical context or explain core concepts).
SECTION 4: USES/APPLICATIONS Describe for what tasks this model/data is used.
SECTION 5: VARIATIONS What variations or similar models, datasets, tasks exist and how does this topic fit into a bigger picture.

Figure 1: The three main prompt types we compared. We eliminated some text in the one-shot setting, which is the ground truth from the survey of `Word2Vec`.

GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) across four different settings. Furthermore, we engage human experts to provide a qualitative dimension, ensuring that our results not only reflect the technical performance but also incorporate subjective human perspectives. We release the LLMs-generated surveys of all these works[1].

2 Method

2.1 Dataset

We adopt the Surfer100 dataset (Li et al., 2022), which contains 100 manually written Wikipedia-style articles on NLP concepts. Each survey is structured into five sections: *Introduction*, *History*, *Key Ideas*, *Uses/Applications*, and *Variations*, with each section containing between 50 and 150 words.

2.2 Experiment Setup

We compare three settings: **zero-shot (ZS)**, **one-shot (OS)** and **description prompt (DP)**. For zero-shot, we directly ask the model to generate the article by providing the following prompt: *Generate a survey about <Topic>. There should be five sub-sections: Introduction, History, Key Ideas, Variations and Applications. Each subsection should contain 50-150 words.* For the one-shot setting, we add a ground-truth article of `Word2Vec` as the sample survey; for the description prompt setting, we add a detailed description to each section explaining what should be included. For example, *SECTION 1: INTRODUCTION Describe what the topic is (a method, a model, a task, a dataset), which field/subfield it is part of, quick overview of applications and motivation behind the concept and related ideas)*. To further enrich the provided information, we also introduce a combination of one-shot and description prompt **(OSP)**. The full prompt is shown in Figure 1. By employing a single ground truth for one-shot learning, we generate 99 surveys per setting. Moreover, we evaluate a special retrieval augmented generation (RAG) (Gao et al., 2023; Ram et al., 2023; Yang et al., 2023; Aksitov et al., 2023) setting (denoted as **OS+IR)** which enables GPT-4 link to Wikipedia pages and

[1]`https://github.com/astridesa/EDULLM/tree/master`

access to web data. We elaborate more experimental details in Appendix B

2.3 Evaluation Metrics

Automatic Evaluation We evaluate the generated surveys using a range of automatic metrics including **ROUGE**, **BERTScore** (Zhang* et al., 2020), **MoverScore** (Zhao et al., 2019), **UniEval** (Zhong et al., 2022) and **BARTScore** (Yuan et al., 2021). Tab. 1 provides an overview of results for the following LLMs: LLaMa2 (13B, 70B), PaLM2 (text-bison), GPT-3.5 (Turbo-0613) as well as GPT-4 (0613) across different prompt settings. We first notice that GPT-4 consistently outperforms other baselines, obtaining a significant improvement of around 2% to 20% when enhancing prompts. Specifically, `GPT-4 OSP` achieves the top spot under most situations. However, it is not to say that prompt enrichment always yields positive results. For instance, in the case of LLaMa2, one-shot and description prompts perform better than `OSP`. As for PaLM2, four types of prompts obtain similar results. When we add external knowledge (`GPT-4 OS+IR`), there is some improvement compared to `GPT-4 OS`. As our primary goal is to study the extent of knowledge LLMs possess in this task, we mainly focus on analyses in settings without external data. However, additional analysis about `GPT-4 OS+IR` setting can be found in Appendix B.

Human and GPTs Evaluation We employ two NLP experts, GPT-4 and G-Eval (Liu et al., 2023a) to evaluate surveys generated by the best `GPT-4 OSP` setting, focusing on 6 perspectives: **Readability**, **Relevancy**, **Hallucination**, **Completeness**, **Factuality**. Both GPT models and humans are required to score each aspect on a scale from 1 to 5, following the same guidelines. The detailed guidance can be found in Appendix A. It's important to note that we implement a pre-selection stage in the choice of human experts (Appendix A). Tab. 2 shows that both human experts and GPTs agree that the generated surveys perform well across most aspects, though the *completeness* exhibits marginally lowest scores. According to IAA, we can observe that human experts demonstrate a high consistent quality of the generated surveys whlie GPT-4 and G-Eval have more randomness. To better understand the degree of agreement between human experts and GPT-4 on ratings, we also calculate Kendall's τ and p-value as shown in Tab. 3. We can observe that the *Factuality* possesses the highest degree of correlation. In contrast, *Redundancy* displays the lowest correlation while the other aspects exhibit relatively lower correlation levels. This difference is largely because *Factuality* is based on objective ground truth, while *Redundancy* is more dependent on subjective judgment. Notably, we can conclude that in most scenarios, GPT-4 showcases similar evaluative opinions as humans, despite showing a higher degree of variability across different independent sessions. Regarding **RQ1** and **RQ2**, we find that 1) LLMs can produce high-quality survey articles, and 2) with specific guidance, there's a strong consistency between GPT outputs and human judgment.

3 Analysis

In this section, we provide an in-depth analysis of the LLMs' internal knowledge on survey writing ability, and compare the evaluation scores of human and LLM assessments.

Error Types We have shown that both automated and manual evaluations demonstrated that LLMs excel in crafting survey articles on scientific concepts. We analyze the best setting, `GPT-4 OSP`, assessing errors identified by two experts, and summarize error types and distributions in Fig. 2. We classify these errors into four categories: Verbose, Wrong Fact, Missing Information, and No Error (indicating flawless content). It shows that most errors are missing information, followed by verbosity and factual inaccuracies. Furthermore, the History and Introduction sections of the generated articles contained the highest number of errors, while the Application section exhibited the best.

Novel Entity Mention To further investigate how interesting the generated content is, we look at the mentions of novel entities following (Lee et al., 2022). Specifically, we examine the survey content by comparing the entities it contains with those in the ground truth. We employ Stanza (Qi et al., 2020) to identify all entities in both the LLM-generated text and the ground truth. Subsequently, we quantify the number of unique entities found in the LLM-generated content. For a fair comparison, we analyze the one-shot with prompt settings of LLaMa2-13b, PaLM2, and GPT-4, in addition to the ZS setting of GPT-3.5, as depicted in Fig. 3. Our findings reveal that PaLM2 exhibited the least variation in entity mentions, while LLaMa2-13b showcased the most. Despite GPT-4's outstanding performance in both automated and human evalu-

Method	ROUGE			BERTScore			MoverScore	UniEval	BARTScore
	R-1	R-2	R-L	P	R	F1			
LLaMa2-13B ZS	27.65	7.81	25.22	85.30	84.73	85.01	55.36	76.03	-4.78
LLaMa2-13B OS	26.53	7.01	24.39	84.86	84.43	84.65	54.94	71.98	-4.81
LLaMa2-13B DP	28.23	7.68	25.83	85.18	85.12	85.14	55.42	74.57	-4.65
LLaMa2-13B OSP	25.84	6.66	23.67	84.51	84.55	84.53	54.65	69.23	-4.74
LLaMa2-70B ZS	27.77	7.59	25.30	85.05	84.82	84.93	55.34	74.06	-4.73
LLaMa2-70B OS	29.69	8.49	27.39	85.72	85.49	85.60	55.63	71.46	-4.48
LLaMa2-70B DP	28.74	8.06	26.29	85.31	84.98	85.14	55.49	72.36	-4.67
LLaMa2-70B OSP	27.74	7.80	25.48	85.32	85.04	85.18	55.52	72.92	-4.68
PaLM2 ZS	27.95	8.95	25.99	85.28	84.61	84.94	55.21	72.69	-4.76
PaLM2 OS	28.81	9.05	26.90	85.16	84.71	84.93	55.35	72.73	-4.68
PaLM2 DP	28.77	9.13	26.65	85.27	84.66	84.96	55.31	72.41	-4.75
PaLM2 OSP	28.71	9.34	26.67	85.14	84.61	84.87	55.28	72.72	-4.74
GPT-3.5 ZS	26.60	6.30	24.36	85.57	84.68	85.12	55.47	**81.31**	-4.75
GPT-4 ZS	26.72	6.61	24.35	85.42	85.39	85.40	55.71	75.24	-4.66
GPT-4 OS	30.09	7.98	27.71	86.01	86.15	86.08	55.98	74.80	-4.38
GPT-4 OSP	**31.47**	**8.62**	**29.04**	**86.19**	**86.44**	**86.31**	**56.04**	75.55	**-4.28**
*GPT-4 OS+IR	31.96	9.43	29.60	86.27	85.88	86.07	56.44	78.56	-4.38

Table 1: Automatic evaluation scores: we compare ROUGE, BERTScore, MoverScore, UniEval, and BARTScore on different settings. The superior scores among the same models are underlined, while the highest scores across all models and settings are highlighted in bold. * We use plugins including A&B Web Search and Keymate.ai.

	Evaluator	Readability	Relevancy	Redundancy	Hallucination	Completeness	Factuality
Mean_{STD}	Human	$4.95_{0.30}$	$4.88_{0.47}$	$4.77_{0.53}$	$4.84_{0.48}$	$4.29_{0.68}$	$4.80_{0.55}$
	GPT-4	$4.84_{0.32}$	$4.67_{0.50}$	$4.85_{0.34}$	$4.86_{0.33}$	$3.93_{0.42}$	$4.56_{0.51}$
	G-Eval	$4.77_{0.64}$	$4.63_{0.68}$	$4.27_{0.74}$	$4.94_{0.51}$	$4.26_{0.76}$	$4.76_{0.66}$
$\text{IAA}_{\%}$	Human	$0.41_{96.96}$	$0.47_{87.87}$	$0.35_{68.68}$	$0.41_{82.82}$	$0.55_{66.66}$	$0.59_{82.82}$
	GPT-4	$0.09_{69.69}$	$0.35_{64.64}$	$0.003_{72.72}$	$0.08_{75.75}$	$0.32_{70.70}$	$0.45_{63.63}$
	G-Eval	$0.06_{49.49}$	$0.25_{38.38}$	$0.01_{32.32}$	$0.008_{95.95}$	$0.42_{33.33}$	$0.02_{58.58}$

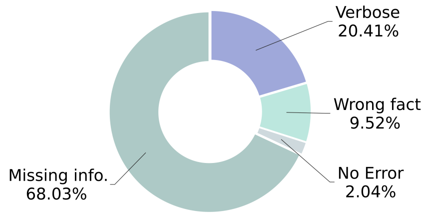
Table 2: Human and GPTs Evaluation Results. We report the mean and standard deviation. We also quantify the IAA (inter-annotator agreement) (Karpinska et al., 2021) between human experts and the GPT results, respectively, using Krippendorff's α coefficient and calculating the percentage (%) of scores that are identical.

	τ	p
Readability	0.16	0.09
Relevancy	0.18	0.05
Redundancy	0.07	0.46
Hallucination	0.11	0.22
Completeness	0.10	0.24
Factuality	0.24	0.01

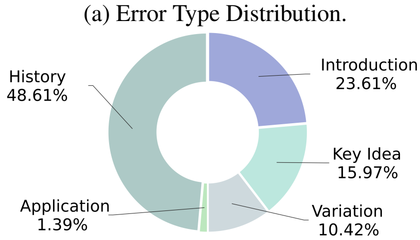
Table 3: The Kendall's τ correlation coefficient and p-value between human and GPT-4.

ations, we didn't discern a marked novelty in its entity mentions. We speculate that this might be an inherent compromise when generating high-fidelity content in relation to the ground truth. So far, regarding **RQ1**, although LLMs register commendable results based on predefined criteria, certain shortcomings are evident. Specifically, we observe some omitted details, particularly within the Introduction and History sections. While LLMs often introduce new entities, we don't find a significant correlation between this tendency and their performance. More case studies are in Appendix C.

LLM and Human Preference Previous studies have indicated that LLM-based evaluation methods tend to favor content generated by LLMs (Liu et al.,



(a) Error Type Distribution.



(b) Section Error Distribution.

Figure 2: Error Analysis by types and sections.

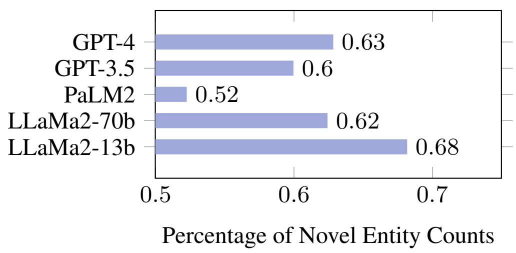


Figure 3: Comparison of novel entity mentions.

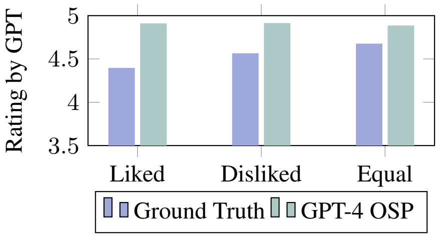


Figure 4: Evaluation comparison on ground truth and GPT-4 predictions, grouped by human preference.

2023b). To test the veracity of this assertion within the context of survey generation tasks, we took the opportunity to investigate whether a similar observation holds in the context of survey generation tasks. Hence, we recruited two human experts in a blind side-by-side comparison of both the ground truth survey articles and articles generated using the best GPT-4 settings, and they assessed the content based on 'Likeability'(Chiang and Lee, 2023). Subsequently, we categorized the survey articles into three groups: a) (human experts) Liked, b) (human experts) Disliked, and c) Equal (equally good). The experts reached a significant agreement, reflected in a Cohen's Kappa score of 0.68 (Cohen, 1960). In instances of disagreement, we randomly selected a score to reach a final consensus. We then apply the GPT-4 evaluation scores on the first four criteria except for *Factuality* and *Completeness* because both are impossible to do a blind test. We show the average ratings on all 99 concepts in Fig. 4. One main observation is the bias of GPT-4 towards texts generated by itself and consistently conferring high ratings – an observation consistent with other studies (Liu et al., 2023b). When evaluating the ground truth, GPT-4 consistently assigns marginally lower ratings across all three categories. Intriguingly, GPT-4 shows a preference for the *Disliked* group over the *Liked* group when considering the ground truth, a tendency that diverges from human inclinations. This suggests that when assessing human-composed text, such as ground truth survey articles, GPT-4 might not yet be an impeccable substitute for human discernment. Thus, in response to **RQ3**, we found that GPT-4 exhibits a notable preference for machine-generated texts with specific biases. Furthermore, we contend that the complete replacement of human experts by GPT-4 is a challenging prospect. For instance, human expertise remains indispensable for manual content fact checking.

4 Discussion and Conclusion

In this work, we evaluate the ability of LLMs to write surveys on NLP concepts. We find that LLMs, particularly GPT-4, can author surveys following specific guidelines that rival the quality of human experts, even though there are shortcomings such as incomplete information. Our findings also indicate that GPT-4 may not be a perfect replacement for human judgment when evaluating human-composed texts, and certain biases exist when asking it to rate machine-generated texts. Nevertheless, the results imply that these advanced generative LLMs could play a transformative role in the realm of education. They hold the promise of effectively structuring domain-specific knowledge tailored to general learners. This adaptability could potentially lead to a more interactive and personalized learning experience, enabling students to engage in query-driven studies that cater directly to their unique curiosities and learning objectives.

Acknowledgements

This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) through the priority program RATIO (SPP-1999), Grant No. 376059226. It was also supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 24K20832. We would like to thank these funding agencies for their generous support. Additionally, we would like to thank the anonymous reviewers for their helpful feedback.

Limitations and Ethical Considerations

GPT-4 can generate contemporary, accessible content, but sometimes compromises depth and detail, leading to potential information gaps and occasional factual inaccuracies. This requires extra verification. Comparing ratings between human experts and GPT-4 revealed a systematic bias in GPT evaluations, which can skew outcomes and mislead quality perception.

The primary objective of this work is to explore the potential applicability of LLMs in the field of education, with a specific emphasis on enhancing the understanding of LLMs' generative capabilities within computer science. Our focus is on assessing the efficacy and identifying the boundaries of the generated texts. It's important to note that the generated texts are devoid of any harmful content, and all data used and produced in this study contains no private information.

References

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models. *arXiv preprint arXiv:2302.05578*.

Signe Altmäe, Alberto Sola-Leyva, and Andres Salumets. 2023. Artificial intelligence in scientific writing: a friend or a foe? *Reproductive BioMedicine Online*.

Rohan Anil, Andrew M. Dai, Orhan Firat, and Melvin Johnson et al. 2023. Palm 2 technical report.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Aosong Feng, Irene Li, Yuang Jiang, and Rex Ying. 2022. Diffuser: Efficient transformers with multi-hop attention diffusion for long sequences. *ArXiv*, abs/2210.11794.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Rahul Jha, Amjad Abu-Jbara, and Dragomir R. Radev. 2013. A system for summarizing scientific topics starting from keywords. In *Annual Meeting of the Association for Computational Linguistics*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Irene Li, Alex Fabbri, Rina Kawamura, Yixin Liu, Xiangru Tang, Jaesung Tae, Chang Shen, Sally Ma, Tomoe Mizutani, and Dragomir Radev. 2022. Surfer100: Generating surveys from web resources, Wikipedia-style. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5388–5392, Marseille, France. European Language Resources Association.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel A McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.

OpenAI. 2023. Gpt-4 technical report.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Chantal Shaib, Millicent L. Li, Sebastian Joseph, Iain J. Marshall, Junyi Jessy Li, and Byron C. Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success).

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing ai support for real-world expository writing tasks. *arXiv preprint arXiv:2304.02623*.

Linxin Song, Jieyu Zhang, Lechao Cheng, Pengyuan Zhou, Tianyi Zhou, and Irene Li. 2023. Nlpbench: Evaluating large language models on solving nlp problems.

Xiaoping Sun and Hai Zhuge. 2022. Automatic survey generation based onposition closeness of key words. *SSRN Electronic Journal*.

Chyrine Tahri, Xavier Tannier, and Patrick Haouat. 2022. On the portability of extractive question-answering systems on scientific papers to real-life application scenarios. In *WIESP*.

Deshraj Yadav Taranjeet Singh. 2023. Embedchain: The open source rag framework. `https://github.com/embedchain/embedchain`.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *ArXiv*, abs/2307.10635.

Kaiyu Yang, Aidan M Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. Leandojo: Theorem proving with retrieval-augmented language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024a. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *ArXiv*, abs/2403.05881.

Rui Yang, Boming Yang, Sixun Ouyang, Tianwei She, Aosong Feng, Yuang Jiang, Freddy Lecue, Jinghui Lu, and Irene Li. 2024b. Leveraging large language models for concept graph recovery and question answering in nlp education. *ArXiv*, abs/2402.14293.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Theodore Zhao, Mu Wei, J. Samuel Preston, and Hoifung Poon. 2023. Llm calibration and automatic hallucination detection via pareto optimal self-supervision.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Human Evaluation Guidance

The detailed human evaluation guidance is listed in the following:

1. **Readability**:
   - **1 (bad):** The text is highly difficult to read, full of grammatical errors, and lacks coherence and clarity.
   - **5 (good):** The text is easy to read, well-structured, and flows naturally.
2. **Relevancy**:
   - **1 (bad):** The generated text is completely irrelevant to the given context or prompt.
   - **5 (good):** The generated text is highly relevant and directly addresses the given context or prompt.
3. **Redundancy**:
   - **1 (bad):** The text is excessively repetitive, containing unnecessary repetitions of the same information. For example, each section should have 50-150 tokens. If it is too long, we should give a low rating.
   - **5 (good):** The text is concise and free from redundancy, providing only essential information.
4. **Hallucination**:
   - **1 (bad):** The generated text includes false or misleading information that does not align with the context or is factually incorrect.
   - **5 (good):** The generated text is free from hallucinations and provides accurate and contextually appropriate information.
5. **Completeness/Accuracy**:
   - **1 (bad):** The generated text is incomplete (missing key information), leaving out crucial details or providing inaccurate information.
   - **5 (good):** The generated text is comprehensive, accurate, and includes all relevant information.
6. **Factuality**:
   - **1 (bad):** The text contains a significant number of factual inaccuracies or false statements, especially in History and Main Idea. For example, Year or people are wrong.
   - **5 (good):** The text is factually accurate, supported by evidence, and free from misinformation.

Pre-selection We initially engaged four NLP specialists to assess the surveys produced by GPT on 20 handpicked topics, as listed in Tab. 4. The evaluation scores across four model configurations are showcased in Tab. 5. Noting the considerable standard deviations among the evaluations of the four judges, we subsequently opted for two judges with a higher alignment in their scores to assess the entirety of the concepts.

B Comparisons with External Knowledge

We conduct further evaluations with the inclusion of links to Wikipedia articles (*GPT-4 OS+Wiki)* and information retrieval (`GPT OS+IR`). In the `GPT-4 OS+Wiki` set up, we apply Embedchain (Taranjeet Singh, 2023) which supports embedding open sources for LLM querying. We crawl Wikipedia articles to concepts in Surfer 100 datasets, yielding 87 effective links. We then prompt GPT-4 to generate survey articles for the respective 87 topics, providing corresponding Wikipedia links and a sample survey as references. As for the setting `GPT-4 OS+IR`, we ask GPT-4 to *'Search on the web for helpful information'*

[2] ZPS means zero-shot with description prompt.

BERT	Autoencoders	Clustering
Decision Trees	Ensemble Learning	Gaussian Mixture Model
Generative Adversarial Network	Gradient Boosting	Hidden Markov Models
Knowledge Graphs	Language Modeling	Long Short-Term Memory Network
Maximum Marginal Relevance	Meta Learning	Multilingual BERT
Perceptron	Relation Extraction	Residual Neural Network
RMSprop Optimizer	Sentiment Analysis	

Table 4: The 20 selected concepts in pre-selection stage.

Model	Readability Mean_{STD}	Relevancy Mean_{STD}	Redundancy Mean_{STD}	Hallucination Mean_{STD}	Completeness Mean_{STD}	Factuality Mean_{STD}
GPT-3.5 ZS	$4.01_{0.98}$	$3.66_{1.61}$	$3.62_{1.04}$	$3.82_{1.18}$	$2.77_{0.94}$	$3.56_{0.83}$
GPT-4 ZS	$4.56_{0.65}$	$4.25_{0.76}$	$4.20_{0.69}$	$4.52_{0.79}$	$3.50_{0.71}$	$3.91_{0.92}$
GPT-4 ZPS[2]	$4.58_{0.72}$	$4.41_{0.75}$	$4.03_{0.81}$	$4.56_{0.64}$	$3.93_{0.69}$	$4.07_{0.93}$
GPT-4 OPS	$4.60_{0.60}$	$4.35_{0.79}$	$4.20_{0.64}$	$4.45_{0.78}$	$3.90_{0.70}$	$4.96_{1.07}$

Table 5: Human evaluation scores on 20 topics of four human experts.

in the prompt and utilitze the web search APIs. Table 6 shows the comparison results between the GPT-4 with and without external knowledge. It's clear to see that both Wikipedia links and the information retrieval component significantly improve the Rouge scores. Notably, searching for web sources efficiently improve both the MoverScore and UniEval. In summary, external knowledge aids GPT-4 in generating higher-quality survey articles than using internal knowledge only. This suggests that LLMs possess limited proficiency when functioning as an academic search engine.

C More Case Study and Observations

C.1 Understanding of "Survey"

When we give the prompt to GPT models by asking them to write a "survey", they sometimes generate survey articles as desired, but they will write other types of content. For example, as indicated in Fig. 5, it appears that GPT would understand the term "Survey" as the questionnaire. Moreover, even if they are able to generate a survey article in the format, there is still the situation that the generated content is not a typical survey. As shown in Fig. 6, there are inconsequential sentences in an attempt to extend and explain the provided text. For example, it repeats saying *this section* and *participants*. But this is mostly observed in the GPT-3.5 zero-shot setting.

C.2 Incomplete Information

In the "History" section, GPT models occasionally produce incomplete evolutionary history, and thus, potentially result in misleading information. For instance, in Fig. 7, when discussing the Knowledge Graph topic, GPT-4 model simply asserts that the term was invented by Google, while the reality is that the concept of Knowledge Graph has a long history, and it is Google that popularized the term. Similarly, in the case of the topic on Decision Trees, although the GPT model yields accurate context, it ignores landmark events and consequently causes misunderstandings.

C.3 Nebulous Sentence Structure

We observe that GPT models frequently construct sentences, especially within the "Application" Section, that employ a rather vague sentence structure, which lacks specificity and can be used in different NLP topics. As shown in Fig. 8, it is evident that GPT models tend to generate similar sentences, such as "The Topic has a wide spectrum of applications" and "The Topic plays a vital role in Natural Language Processing and Natural Language Understanding"; These statements hold significant meaning when "The Topic" is substituted with any NLP topics.

	ROUGE			BERTScore			MoverScore	UniEval	BARTScore
	R-1	R-2	R-L	P	R	F1			
GPT-4 OS	30.09	7.98	27.71	86.01	**86.15**	**86.08**	55.98	74.80	**-4.38**
GPT-4 OS+Wiki	**31.99**	9.40	**29.63**	**86.31**	85.85	**86.08**	56.10	75.04	-4.43
GPT-4 OS+IR	31.96	**9.43**	29.60	86.27	85.88	86.07	**56.44**	**78.56**	**-4.38**

Table 6: Automatic evaluation scores for settings with external knowledge. Highest scores are highlted.

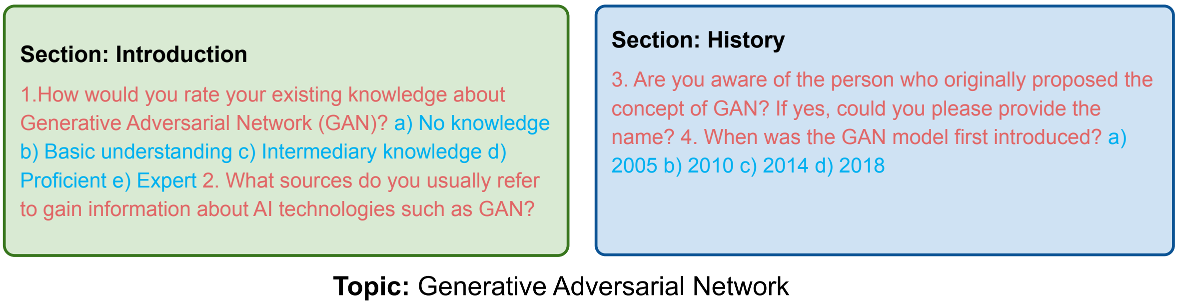


Figure 5: An example of generated questionnaire as a survey (part).

C.4 High-quality Survey

We also present a high-quality generated survey in Fig 9. It is designed to read and understand easily, providing readers with comprehensive and detailed information. The example survey on LSTM is well-structured, with a summary provided in the first sentence and followed by the detailed explanation in each section. Specially, when discussing applications, it demonstrates a high level of domain specificity. Most importantly, the generated information is both accurate and concise.

C.5 Going Beyond the Ground Truth

GPT-4 based methods maintain an overall high-quality response regarding all aspects. We show two examples by comparing the *GPT-4 one-shot with prompt* setting result with the ground truth in Fig. 10. In the first topic, multilingual BERT (mBERT), GPT successfully points out that the key idea behind mBERT is mapping words from distinct languages into a shared embedding space. However, the ground truth only mentions shared vocabulary, which is superficial. In the second example, Hidden Markov Models (HMMs), the GPT response is more precise and more complete than the ground truth. One can find that the content flow is present as algorithm category → features → applications → motivation (highlighted in bold and italicized words). In contrast, the ground truth texts spend a lot of words to which category HMM belongs, including many terminologies which is less informative.

Limitations of Ground Truth We refer to this previous work (Li et al., 2022) on how the ground truth was generated. In general, the human writer was asked to rely on web data when writing the survey article; while these data were collected in the year 2021, it may be hard to say if it is a fair ROUGE score comparison with GPT models in Tab. 1. While the ground truth may not be a perfect reference, in this work, we focus more on human evaluation and case studies.

D Potential Risks

Sole dependence on LLMs for educational content can lead to a homogenization of information and lack the nuanced understanding that human experts bring. If not properly vetted, the occasional factual errors made by GPT-4 could lead to the propagation of misinformation, especially detrimental in an educational context. Over time, heavy reliance on automated systems might diminish the role of human experts in content creation, leading to a potential loss of rich, experience-based insights. The observed systematic bias in GPT evaluations can lead researchers to draw incorrect conclusions about the quality of content, potentially impacting future research and educational endeavors.

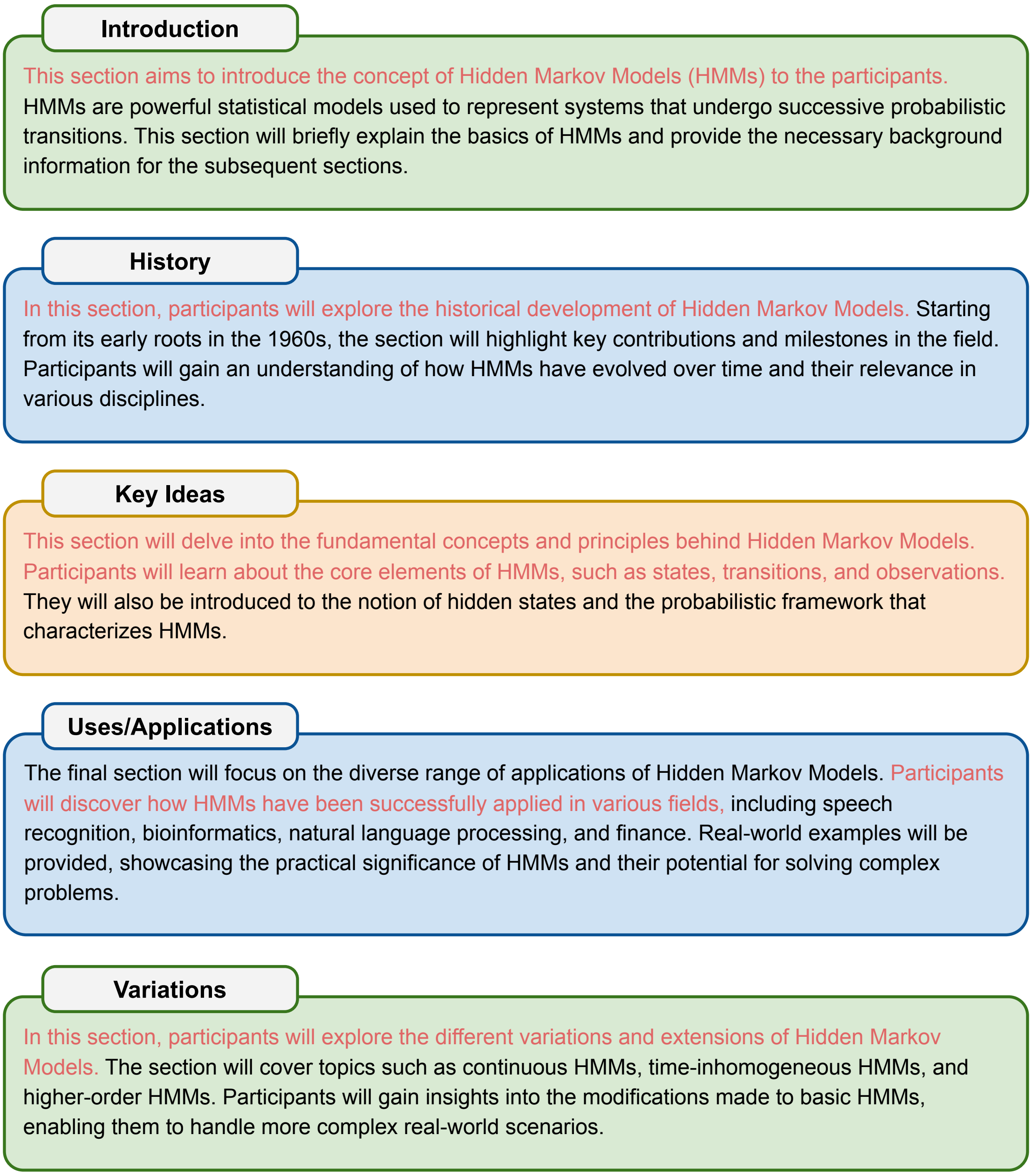


Figure 6: An example of an invalid generated survey.

E Experimental Details

In our work, we mainly utilize the paid GPT-4 model to generate Wiki-style survey articles and further explore its capabilities to score the generations. The total cost of these experiments is around 230 USD. During the automatic evaluation stage, we compute the ROUGE score and BERTScore using the officially provided APIs: `rouge`[3] and `bert_score`[4]. For calculations involving MoverScore, UniEval, and BARTScore, we directly download their source codes. All experiments were performed using the high-performance machine with 4 A100 40GB NVIDIA cards. As the experiments do not involve fine-tuning, for each setting, we were able to finish in a few hours. As for the human evaluation stage, we calculate the Krippendorff's and Kendall's scores with the authorized APIs `krippendorff`[5] and `scipy`[6].

[3] https://pypi.org/project/rouge/
[4] https://github.com/Tiiiger/bert_score
[5] https://pypi.org/project/krippendorff/
[6] https://scipy.org/

GPT-4 zero-shot with prompt

Debuted by Google in 2012, Knowledge Graph is a synthesis of graph theory and semantic searching that targets more precise search results by understanding intent and meaning of search queries. Google introduced it to refine their search engine, understanding user queries' context and delivering relevant results. The Knowledge Graph evolution saw ontology and machine learning playing significant roles, with schema verification, information extraction and data alignment offering more precise data interpretations and integration.

Ground Truth

In 1960, Semantic Networks were invented as a knowledge representation framework. The main idea of Semantic Networks was to capture a wide range of entities including real-world objects, events, situations or abstract concepts. In 2001, Tim Berners-Lee coined the term Semantic Web which is an application of Semantic Network combined with the Web. In 2012, Google popularized the term knowledge graph with the announcement of Knowledge Graph. there are very few technical details about its organization, coverage and size, but it is said that it contained 70 billion connected facts by 2016. Today, knowledge graphs are used by companies from a wide range of industries.

Topic: Knowledge Graph **Section:** History

GPT-4 one-shot with prompt

Decision Trees were first introduced by Ross Quinlan, a researcher in machine learning, who developed an algorithm known as ID3 (Iterative Dichotomiser 3) dating back to the 1980s. The algorithm was developed in order to create a Decision Tree that predicts the value of a target variable based on several input variables. This was a significant development in the field of Machine Learning as it addressed the problem of handling categorical data and missing values while predicting multi-output concepts.

Ground Truth

The first regression tree was invented in 1963 (AID project, Morgan and Sonquist). The first publication on decision trees was in 1966 (by Hunt). The first classification tree appeared in the THAID project (by Messenger and Mandell). In 1974, Statistics professors Leo Breiman and Charles Stone from Berkeley and Jerome Friedman and Richard Olshen from Stanford started developing the classification and regression tree (CART) algorithm. In 1977, Breiman, Stone, Friedman, and Olshen invented the first CART version.

Topic: Decision Trees **Section:** History

Figure 7: Two example surveys with incomplete information.

Model: GPT-4 zero-shot; **Topic:** Decision Trees; **Section:** Application

Knowledge Graphs have a broad spectrum of applications. They significantly improve the efficiency of search engines, personalized recommendations, innovations in AI, and semantic searches. They also play a vital role in Natural Language Processing tasks such as machine reading and language understanding. Other potential applications include healthcare for an accurate analysis of patient data, e-commerce for efficient product recommendation, social media analytics, and cultural heritage preservation. With tech giants continually enhancing their proprietary knowledge graphs, the use of Knowledge Graphs in various sectors is anticipated to grow exponentially in the future.

Model: GPT-4 one-shot with prompt; **Topic:** Meta-learning; **Section:** Application

Meta-learning has a wide spectrum of applications. Aside from its notable use in deep learning where neural networks are used, it's also used in reinforcement learning. These tasks include but aren't limited to, robotic control, game playing, and autonomous navigation. Additionally, meta-learning can be used for rapid adaptation in natural language processing and, in particular, natural language understanding. Several companies use meta-learning algorithms for designing personalized recommendations and adaptive chatbots.

Figure 8: Example surveys with nebulous sentence structure.

Introduction

Long Short-Term Memory Networks (LSTM) are an innovative type of Recurrent Neural Networks (RNNs) that have been specifically designed to effectively manage and maintain the essential elements extracted from time-varying or sequence data. LSTMs have been a remarkable advancement in the field of artificial intelligence because they address the shortcomings of traditional RNNs such as the undesirable effects of gradient vanishing and exploding, which impede RNNs' learning ability for long time-lag tasks.

History

Introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997, the LSTM addresses the vanishing gradient problem experienced by traditional RNNs. The main novelty of LSTM was the incorporation of gating units. These units permit or restrict information from flowing through the sequence chain, thereby effectively containing the exploding or vanishing gradient problem. Since its inception, LSTM has served as a prerequisite for various network designs addressing further issues and limitations.

Key Ideas

The core concept behind LSTM networks is the cell state, a controllable information pipeline that carries the required details from early input sequences to later ones, making provision for long-term dependencies. ***(Accurate and easy-to-understand)*** LSTM adjusts the cell state through carefully designed structures called gates, which are capable of removing or adding information to the cell state. There are three main types of gates: forget gate deciding what information should be discarded, input gate deciding what new information should be stored in the cell state, and output gate deciding what information should be utilized.

Uses/Applications

LSTM found a myriad of applications in fields where sequence or time series data are crucial. ***(demonstrate domain-specificity)*** In natural language processing, LSTM has proven effective in machine translation, sentiment analysis, named entity recognition, and text generation due to its ability to capture context dependencies. In computer vision, LSTM has been utilized for video processing and in action recognition tasks. Also, in time series analysis and forecasting including stock market prediction or weather forecasting, LSTM exhibits prime performance.

Variations

Over the years, several variations of LSTMs have been introduced to maximize efficiency and minimize computational resources. The Gated Recurrent Unit (GRU) is a popular variation, which merges the cell state and hidden state, and uses two gates. Another important variation is the Peephole LSTM, which lets the gate layers peep into the cell state. The Convolutional LSTM replaces simple multiplication operations with convolutional operations, making it especially effective for spatial temporal data.

Figure 9: An example of the high-quality survey.

GPT-4 one-shot with prompt

This specific type of training allows it to create shared representations for syntactically or semantically similar words across different languages. One core concept behind mBERT is that it does not require distinct models for each language; it works on the idea of shared multilingual embeddings, leading to representation where different languages' embeddings should be close to each other if they are translations.

Ground Truth

One important point to note is that while training on many languages, we keep a shared vocabulary for all languages rather than keeping distinct vocab for every language. This saves space and most importantly our model learns the root structure of language and learn the underlying structure rather than just learning of the vocab.

Topic: Multilingual BERT **Section:** Main Idea

GPT-4 one-shot with prompt

Hidden Markov Models (HMMs) are a statistical model used primarily in the field of machine learning and data analysis. Part of the broader category of Markov chain models ***(present the category)***, HMMs offer a way to describe the evolution of observable events that are driven by an internal state which is not directly accessible ***(present the features)***. The model's effectiveness has found it a important role in numerous applications, ranging from image recognition, speech recognition, genomics to finance ***(present the applications)***. The primary motivation behind HMMs is to provide a framework to infer the missing information from observable data, where every state from the hidden sequence generates an output according to a certain probability distribution related to that state ***(present the motivation)***.

Ground Truth

Hidden Markov models (HMMs) are a way of relating a sequence of observations to a sequence of hidden classes or hidden states that explain the observations. They are a class of a full probabilistic model—the model parameters and the overall sequence 'scores' are all probabilities. They form the foundation for creating probabilistic models of linear sequence 'labeling' problems. From just drawing an intuitive picture, HMMs offer fundamental concepts for building a complex model and are the core of many algorithms in computational sequence analysis, including genefinding, profile searches, multiple sequence alignment, and regulatory site identification.

Topic: Hidden Markov Models **Section:** Introduction

Figure 10: Two examples showing that the generated output is better than the ground truth.