

Document Version
Accepted version

Citation for published versión:

Martínez-Cruz, R., López-López, A.J. & Portela, J. ChatGPT vs state-of-the-art models: a benchmarking study in keyphrase generation task. *Appl Intell* **55**, 50 (2025). https://doi.org/10.1007/s10489-024-05901-4

Please note that the full-text provided on Comillas' Research Portal is the Author Accepted Manuscript or Post-Print version.

General rights
This manuscript version is made available under the CC-BY-NC-ND 4.0 licence (https://web.upcomillas.es/webcorporativo/RegulacionRepositorioInstitucionalComillas.pdf).

Take down policy
If you believe that this document breaches copyright please contact Universidad Pontificia Comillas providing details, and we will remove access to the work immediately and investigate your claim

Manuscript Click here to view linked References

ChatGPT vs State-of-the-Art Models: A Benchmarking Study in Keyphrase Generation Task

Roberto Martínez-Cruz[1,2*], Alvaro J. López-López[2] and José Portela[2]

[1*]Moody's Data Solutions, Moody's Analytics, Prague, Czech Republic.
[2]Institute for Research in Technology, ICAI School of Engineering, Comillas Pontifical University, Madrid, Spain.

*Corresponding author(s). E-mail(s): roberto.martinezcruz@moodys.com;
Contributing authors: allopez@comillas.edu; jportela@comillas.edu;

Abstract

Transformer-based language models, including ChatGPT, have demonstrated exceptional performance in various natural language generation tasks. However, there has been limited research evaluating ChatGPT's keyphrase generation ability, which involves identifying informative phrases that accurately reflect a document's content. This study seeks to address this gap by comparing ChatGPT's keyphrase generation performance with state-of-the-art models, while also testing its potential as a solution for two significant challenges in the field: domain adaptation and keyphrase generation from long documents. We conducted experiments on eight publicly available datasets spanning scientific, news, and biomedical domains, analyzing performance across both short and long documents. Our results show that ChatGPT outperforms current state-of-the-art models in all tested datasets and environments, generating high-quality keyphrases that adapt well to diverse domains and document lengths.

Keywords: ChatGPT, Text Generation, Keyphrase Generation, Natural Language Processing, Deep Learning, Domain Adaptation, Long Documents, Large Language Models

1 Introduction

Keyphrase generation (KPG) is the process of automatically identifying or creating a set of phrases that effectively capture the essence of a document or text. These keyphrases provide a succinct summary of the main topics or themes discussed in the text and can be utilized for a range of downstream tasks such as document classification [1], clustering [2], summarization [3], recommendation [4], and information retrieval [5].

There are two types of keyphrases: extractive (found in the document) and abstractive (not found in the document). Historically, extractive methods, known as Keyphrase Extraction (KPE), based on sequence tagging models [6–10] have demonstrated the highest accuracy, although they are limited in their ability to predict abstractive keyphrases. KPG offers two primary advantages over KPE: it can predict both extractive and abstractive keyphrases and It can leverage prompt-based learning to more effectively benefit from multitask learning.

The majority of KPG models follow the text-to-text generation training paradigm, which has greatly benefited from the development of transformer models and pre-trained language models (PLMs). These PLMs are capable of acquiring a comprehensive contextualized representation of the text by undergoing pre-training with a vast corpus of data via diverse self-supervised learning tasks. The current state-of-the-art (SotA) model, KeyBART [11], is based on this training paradigm and employs the transformer-based architecture of BART [12], a PLM, as its foundation. KeyBART has demonstrated promising performance on various keyphrase generation benchmarks, surpassing previous state-of-the-art models. However, there is still ample room for improvement in the KPG domain, particularly in the generation of abstractive keyphrases and the effective integration of external knowledge sources.

Text-to-text generation has proven to be an effective approach to facilitate multitask learning. The incorporation of prompt-based learning, as demonstrated in models such as T5 [13], has significantly enhanced the performance of text-to-text generation models, particularly in few-shot and zero-shot learning scenarios.

Generative Pre-trained Transformer (GPT) models have garnered significant attention for their ability to generate coherent and context-aware text [14]. The latest GPT model, ChatGPT [15], based on GPT-3.5 [16], has demonstrated its impressive ability to generate human-like responses to text-based conversations across various domains and topics. Given the adaptability of ChatGPT to new domains and the transfer learning capabilities from other tasks learned by the model, we believe that it has the potential to achieve state-of-the-art results in the KPG task. Additionally, ChatGPT may also help address a long-standing challenge in the field of KPG - domain adaptation. By leveraging the power of transfer learning and its ability to adapt to new domains, ChatGPT can potentially overcome domain-specific challenges in KPG and enhance the performance and efficiency of the task.

Keyphrase generation from long documents is a latent problem in the field. Current approaches often rely on summarizing texts, such as using abstracts, to identify important phrases. However, this method has limitations. Real-world situations may not always provide summaries, leading to reduced algorithm performance on longer texts. Additionally, crucial keyphrases may be missing from the summaries, and contextual information in the original text may not be reflected in them, greatly

reducing the algorithm's effectiveness. While SotA approaches utilize contextualized text representations from PLMs, these representations are limited to a maximum number of words, preventing the embedding of long-term word relationships. The greater maximum token limit of ChatGPT, which is four times larger than that of KeyBART, may lead to better performance on long documents.

Our goal is to assess the potential of ChatGPT for KPG by conducting performance tests across diverse topics, such as scientific, news and biomedical domains, as well as varying document lengths. We intend to determine whether ChatGPT's performance remains consistent across different document lengths and topics. To do so, we compare its results with the SotA algorithms, specifically KeyBART, on six publicly available datasets. Additionally, we provide real-world examples to demonstrate how ChatGPT generates keyphrases from input texts.

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, it's the first attempt to employ ChatGPT for the KPG task in both the news and long document domains. To assess its performance, we benchmark it against state-of-the-art models across eight widely used datasets, which include a mix of short and long documents from scientific, biomedical, and news sources.
- To the extend of our knoledge, no other study has benchmarked keyphrase generation from long documents. Previous studies have relied solely on the title and abstract to generate keyphrases, which may not provide an accurate representation of real-world scenarios where the entire document needs to be processed due to the absence of a summary.
- Our results are comprehensively analyzed, and case studies are presented to showcase the model's strengths and limitations in the KPG task. These case studies use real-world examples to demonstrate how the model leverages knowledge from other tasks and pretraining objectives to enhance keyphrase generation.

2 Related Work

2.1 Keyphrase Extraction and Generation

KPE involves selecting relevant phrases from a document, and there are two main approaches: supervised and unsupervised. Unsupervised methods typically use a two-step extraction process that involves heuristically identifying candidate phrases, then sorting and ranking them using graph-based approaches [17–21]. Supervised methods use labeled data and can be customized for specific linguistic and contextual characteristics. Earlier supervised approaches relied on manually-engineered features [6, 22, 23], but a sequence labeling approach using a conditional random field (CRF) was introduced in [7], and recent methods incorporate pre-trained word embeddings [8] like Word2Vec [24] or GloVe [25] to improve the accuracy. The improved sequence labeling approach uses a bidirectional long short-term memory (BiLSTM) + CRF layer to incorporate contextual information and model classification dependencies.

The transformer architecture, which has demonstrated improved performance in various natural language processing tasks, including as described in [26], has been

employed in several works for the KPE task, including TransKP [9] and TNT-KID [27]. Its ability to embed words in a sequence and provide representations that depend on the word and its context has led to the development of PLMs that specialize in providing contextualized embeddings. When combined with a trained BiLSTM-CRF layer, these embeddings have outperformed previous models, as demonstrated in [10]. Some works, such as KBIR [11], have designed self-supervised objectives specifically for keyphrase extraction tasks to further enhance the representation of these embeddings, resulting in state-of-the-art results.

However, the extractive model cannot handle the absent keyphrase. To overcome this limitation, KPG introduces a sequence-to-sequence generation method, first presented by [28] through CopyRNN, a Seq2Seq framework that employs attention and copy mechanism. Since then, researchers have proposed several enhancements to this methodology. [29] explored a semi-supervised method, [30] investigated a review mechanism to reduce duplicates, and [31] focused on leveraging title information. Meanwhile, [32] exploited deeper topics of the document, and [33] utilized linguistic constraints. Reinforcement learning was introduced by [34] and [35]. [36] proposed an exclusive hierarchical decoding framework, while [37] introduced a model that generates multiple keyphrases as delimiter-separated sequences. [38] proposed separate mechanisms to deal with present and absent keyphrase generation. [39] presented an AdaGM method to increase the discreteness of keyphrase generation, and [40] proposed an one2set method for generating diverse keyphrases as a set. Other works, including [41], [42], and [43], focused on jointly learning extraction and generation for keyphrase prediction. [44] introduced prompt-based learning with a non-autoregressive approach, which is constrained to generate absent keyphrases.

The SotA results were achieved by KeyBART [11], which presented a new pre-training technique for BART [12]. Unlike earlier pre-training methods that aimed to remove noise from the input text, KeyBART generates keyphrases related to the input text in concatenated sequential (CatSeq) format.

While previous studies, such as [45], have examined the effectiveness of ChatGPT in the KPG task, they did not assess its performance in full-length documents that exceed the model's maximum input limit without truncation, nor did they investigate its suitability for the news domain. Our study, on the other hand, provides a comprehensive analysis that extensively explores the capabilities of both KeyBART and ChatGPT across various use cases in the KPG task.

2.2 GPT Models

GPT models have been widely used in NLP tasks, such as language generation, language understanding, and question answering. GPT models, including GPT-2 [46], GPT-3 and GPT 3.5 [16], have achieved SotA performance in several NLP tasks and have become the de-facto standard in many applications. They exclusively use the transformer's decoder and are pre-trained in a massive corpora of data using self-supervised learning.

Reinforcement learning from human feedback (RLHF) allows language models to learn from explicit feedback provided by human annotators, leading to improved text quality. Originally developed for training robots [47, 48], recent studies have shown

the benefits of applying RLHF to fine-tune language models [49–59], including GPT models such as ChatGPT [15].

Multi-task learning has proven to be advantageous for GPT models since it entails instructing language models and is connected to cross-task generalization research in language models. Research has shown that fine-tuning language models on various NLP tasks with instructions can enhance their performance downstream, making it a powerful method for few-shot and zero-shot learning. These advantages have been corroborated in several studies [60],[61],[62], [63],[64].

By leveraging both RLHF and multi-task learning paradigms, ChatGPT has been fine-tuned from the GPT3.5 model to excel in chatbot development, surpassing its predecessors and showcasing its potential to revolutionize the field of conversational AI. While previous studies have benchmarked its performance in various NLP tasks, such as machine translation [65], there have been no previous studies exploring its potential in the KPG task.

3 Experimental Setup

3.1 Datasets

To evaluate the performance of KPG models, we employ the test set of six publicly available datasets covering both scientific and news domains, with varying document lengths. The datasets we use are as follows:

- Inspec[1] ([22]) is a scientific literature dataset that consists of 2,000 abstracts and their corresponding keyphrases, covering various topics The abstracts are from papers belonging to the scientific domains of Computers and Control and Information Technology published between 1998 to 2002. The dataset has a train,val and test split that contains 1000, 500 and 500 samples respectively.
- KP20K[2] ([28]), a large-scale dataset with over 528K articles for training, 20K articles for validation, and 20K articles for testing from the PubMed Central Open Access Subset, covering various domains including medicine, biology, and physics.
- The NUS[3] dataset ([6]) consists of 211 full scientific documents that have been manually annotated with their respective keyphrases. It is used exclusively for evaluation purposes, as the dataset comprises solely a test split.
- SemEval2010[4] ([66]), a dataset comprising 284 English full scientific papers from the ACM Digital Library, which are split into test and train sets containing 100 and 144 articles, respectively.
- The KPTimes[5] ([67]) dataset consists of 279,923 news articles from NY Times and 10K from JPTimes, curated by expert editors, and divided into train, validation, and test sets with 259,923, 10,000, and 20,000 samples, respectively.

[1] https://huggingface.co/datasets/midas/inspec
[2] https://huggingface.co/datasets/midas/kp20k
[3] https://huggingface.co/datasets/midas/nus
[4] https://huggingface.co/datasets/midas/semeval2010
[5] https://huggingface.co/datasets/midas/kptimes

Dataset	Test Size	Long Doc	Domain	Avg. Words	Avg. Extractive KPs	Avg. Abstractive KPs
Inspec	500	No	Scientific	135	6.56	3.26
KP20k	20,000	No	Scientific	160	2.34	2.94
NUS	211	Yes	Scientific	9287	8.02	3.08
SemEval2010	100	Yes	Scientific	8404	9.17	6.07
KPTimes	20,000	No	News	643	2.72	2.3
DUC2001	308	No	News	847	7.14	0.92
KPBioMed	2.4M	No	Biomedical	226	3.33	1.87
PubMed	1,319	Yes	Biomedical	4408.09	4.52	0.87

Table 1 Statistics of the test splits for the dataset used in our experiments

- The DUC2001[6] dataset ([68]) is a widely recognized corpus of news articles that includes 308 documents and 2,488 manually annotated keyphrases. It should be noted that this dataset only contains a test split and no training data.
- KPBioMed[7] ([69]) is a large-scale dataset for keyphrase generation, consisting of 2.4 million abstracts with author-assigned keyphrases. It is primarily focused on the biomedical domain. The dataset categorizes keyphrases under the PRMU (Present-Reordered-Mixed-Unseen) scheme.
- PubMed[8] ([70]) is a widely-used keyphrase extraction dataset in the biomedical field, comprising full-text documents and corresponding keyphrases from the PubMed Central Open Access Subset. The dataset provides a test split with 1,319 documents and 5,444 unique keyphrases.

Relevant statistics from the datasets can be found in Table 1.

3.2 Baselines

The evaluation of ChatGPT's performance involved comparing it with several other generative models, with respect to their ability to predict missing keyphrases across different generation frameworks. These models are:

- KeyBART[9] ([11]), which is a state-of-the-art model developed using BART and featuring a novel pre-training approach that generates relevant keyphrases in the CatSeq format, instead of just removing noise from input text.
- Prompt Based KPG ([44]), which utilizes a prompt-based learning method to generate missing keyphrases. The prompt is created based on the overlapping words between the absent keyphrase and the document, and a mask predict decoder is used to complete the keyphrase while adhering to the constraints of the prompt.

[6] https://huggingface.co/datasets/midas/duc2001
[7] https://huggingface.co/datasets/taln-ls2n/kpbiomed
[8] https://github.com/boudinfl/ake-datasets
[9] https://huggingface.co/bloomberg/KeyBART

Model	Training Domain	Max Input Tokens
ChatGPT	MultiDomain	4,096 (Summing Input and Output Tokens)
KeyBART	Scientific	1024
Prompt Base KPG	Scientific	384
UniKeyphrase	Scientific	384

Table 2 Insights from the main models used in our experiments

- UniKeyphrase ([43]), which is a unified framework for both present keyphrase extraction and absent keyphrase generation. This framework is based on a pre-trained prefix LM model.
- Pure generative models that use a Seq2Seq approach including CatSeq ([37]) and its enhanced versions such as CatSeqCorr ([30]), catSeqTG ([31]), and CatSeqD ([37]), along with ExHiRD-h model ([36]), for comparison.

Relevant information from the main models can be found in Table 2.

3.3 Evaluation Metrics

We used the $F1@K$ evaluation metric ([66]), where K represents the number of predicted keyphrases to be considered. Equations 1, 2, and 3 illustrate how to compute $F1@K$. Prior to evaluation, we preprocessed the ground truth and predicted keyphrases by converting them to lowercase, stemming them, and removing punctuation, and we used exact matching. Let Y denote the ground truth keyphrases, and $\bar{Y} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m)$ denote the predicted keyphrases. The metrics are defined as follows:

$$Precision@k = \frac{|Y \cap \bar{Y}_k|}{min\{|\bar{Y}_k|, k\}} \tag{1}$$

$$Recall@k = \frac{|Y \cap \bar{Y}_k|}{|Y|} \tag{2}$$

$$F1@k = \frac{2 * Precision@k * Recall@k}{Precision@k + Recall@k} \tag{3}$$

Here, $\bar{Y}_k$ denotes the top k elements of the predicted keyphrase set $\bar{Y}$. In our case, we set K to the total number of predicted keyphrases M, and 5 to represent the top 5 keyphrases generated by the model.

```
[{"role": "system", "content": "You are a helpful assistant that extracts keyphrases from the text"},
 {"role": "user", "content": text}]
```

Fig. 1 Example prompt used to generate keyphrases with ChatGPT

3.4 Setting

For KeyBART, the generated results are produced using a beam search with a beam width of 50, and the maximum length of the generated sequence is restricted to 40 tokens with a temperature of 0. When the number of input tokens exceeds the model's maximum input limit, the tokens are divided into non-overlapping chunks of the same size as the maximum input token limit, and the generated keyphrases are produced accordingly, concatenated and any duplicates are removed.

To generate keyphrases using ChatGPT, we utilize the gpt-3.5-turbo model in chat completion mode, with the prompt illustrated in Figure 1. In the prompt, the *text* variable represents the input text from which the keyphrases are to be generated. The generated sequence is restricted to 40 tokens with a temperature of 0, a frequency penalty of 0.8 and a presence penalty of 0. Although the model's response may vary in the chat format, it always includes a list of keyphrases presented in one of three ways: comma-separated, enumerated, or itemized. The response can be transformed into a list of keyphrases by post-processing it using regular expressions. Since the experiments involve processing long documents that exceed the maximum input token limit, the text is split into non-overlapping 2000-word segments and input using individual prompts. The final step involves concatenating the results and removing any duplicates to generate the complete list of keyphrases.

The primary aim of this study is to evaluate the performance of scientific domain models using datasets consisting of short documents. Unlike prior studies that relied on the title and abstract of papers in long document datasets, we will use the entire documents to assess the models' effectiveness on longer texts. Moreover, we will examine the models' adaptability across different domains by comparing their performance on datasets from both the news and biomedical fields. To conduct these two analyses, we will compare the results of the current SotA model, KeyBART, and ChatGPT.

We used the first five generated keyphrases for $F1@5$ to evaluate the models. If there were fewer than five keyphrases, we appended random incorrect keyphrases until there were five predictions. However, due to multiple separate generations for the same sample in long documents, this metric did not provide clear insights into the model's performance. Therefore, in this scenario we only benchmarked the results based on $F1@M$. In order to compare the ground truth keyphrases with the generated ones, we utilize the Porter Stemmer to normalize both as [28].

4 Results

This section presents the results of our experiments. ChatGPT outperforms the specialized models in the task in all scenarios, with the performance improvement becoming

Task	Model	Inspec		KP20k	
		F1@5	F1@M	F1@5	F1@M
Present Keyphrases	catSeq	0,225	0,262	0,291	0,367
	CatSeqD	0,219	0,263	0,285	0,363
	CatSeqCorr	0,227	0,269	0,289	0,365
	CatSeqTG	0,229	0,270	0,292	0,366
	ExHiRD-h	0,253	0,291	0,311	0,364
	SEG-Net	0,216	0,265	0,311	0,379
	UniKeyphrase	0,260	0,288	0,347	0,352
	Prompt Base KPG	0,260	0,294	**0,351**	0,355
	KeyBART	0,278	0,301	0,301	**0,398**
	ChatGPT	**0,352**	**0,403**	0,232	0,251
Absent Keyphrases	catSeq	0,004	0,008	0,015	0,032
	CatSeqD	0,006	0,011	0,015	0,031
	CatSeqCorr	0,005	0,009	0,015	0,032
	CatSeqTG	0,005	0,011	0,015	0,032
	ExHiRD-h	0,011	0,022	0,016	0,032
	SEG-Net	0,009	0,015	0,018	0,036
	UniKeyphrase	0,012	0,022	0,032	0,058
	Prompt Base KPG	0,017	0,022	0,032	0,042
	KeyBART	0,041	0,045	0,035	0,035
	ChatGPT	**0,049**	**0,059**	**0,044**	**0,056**

Table 3 Results for short scientific documents of keyphrase prediction on benchmarks. The bold-faced values indicate the best performances across the board.

more significant as the document length increases or as the document's domain moves farther away from the scientific domain.

4.1 Short Scientific Documents

As previously highlighted, Inspec and KP20k datasets are used to benchmark the performance of the models in the short scientific documents domains. Scenario in which all the models are specialized with the exception of Chat GPT. In Table 3 the results are shown.

Despite all models being specialized in the domain, ChatGPT surpasses them in all cases except for the KPE task in KP20K, where the fine-tuning of the models on this dataset explains their exceptional performance. However, ChatGPT outperforms them in the generation of abstractive keyphrases in both datasest. We speculate that this task benefits greatly from ChatGPT's multi-task learning, which enables the model to learn useful knowledge from other tasks and apply it to generate high-quality abstractive keyphrases; which could explain why the specialized models fine-tuned in the distribution were outperformed by ChatGPT.

Task	Model	SemEval2010 F1@M	NUS F1@M
Present Keyphrases	KeyBART	0,137	0,143
	ChatGPT	**0,186**	**0,1996**
Absent Keyphrases	KeyBART	0,019	0,010
	ChatGPT	**0,021**	**0,042**

Table 4 Results for long scientific documents of keyphrase prediction on benchmarks. The bold-faced values indicate the best performances across the board.

Task	Model	DUC2001		KPTimes	
		F1@5	F1@M	F1@5	F1@M
Present Keyphrases	KeyBART	0,023	0,079	0,023	0,063
	ChatGPT	**0,267**	**0,292**	**0,279**	**0,290**
Absent Keyphrases	KeyBART	0,001	0,001	0,006	0,010
	ChatGPT	**0,029**	**0,030**	**0,021**	**0,022**

Table 5 Results for news of keyphrase prediction on benchmarks. The bold-faced values indicate the best performances across the board.

4.2 Long Scientific Documents

To asses the performance of the models in long scientific documents the SemEval2010 and NUS datasets are utilized, where the number of tokens of each sample is several times larger than the maximum allowed by the models. The results of these experiments are presented in Table 4.

In this scenario, it is evident that ChatGPT outperforms KeyBART by a substantial margin. This can be credited to ChatGPT's higher input token limit, which enables it to capture more contextual information, such as distant word relationships. This is crucial for generating accurate and relevant keyphrases, giving ChatGPT a clear advantage over KeyBART in this task. Furthermore, we hypothesize that the incorporation of other tasks during training has facilitated the development of a larger language model without compromising its performance in this specific task. This approach could potentially be used to train larger language models specialized in KPG.

4.3 News Domain

To evaluate the domain adaptation abilities of the models, we will employ the news domain datasets DUC2001 and KPTimes. These datasets exhibit significant differences in both their domain and distributions when compared to the scientific domain. However, both models have been pre-trained on this domain. KeyBART, for example, is built on BART, which includes this distribution in its pre-training. The results are displayed in Table 5.

As demonstrated by the results, ChatGPT achieved significantly stronger performance in this domain, surpassing the results of KeyBART by a factor of three or more in every benchmark. The notable performance difference between ChatGPT and

Task	Model	KPBioMed		PubMed	
		F1@5	F1@M	F1@5	F1@M
Present Keyphrases	KeyBART	**0.116**	0.148	0.062	0.134
	ChatGPT	0.103	**0.315**	**0.174**	**0.433**
Absent Keyphrases	KeyBART	**0.075**	0.091	**0.076**	**0.182**
	ChatGPT	0.051	**0.131**	0.007	0.024

Table 6 Results for biomedical documents of keyphrase prediction on benchmarks. The bold-faced values indicate the best performances across the board.

KeyBART can be attributed to the fact that although KeyBART's weights include knowledge from the news domain due to BART's pre-training, this knowledge may have been relatively forgotten during KeyBART's pre-training and fine-tuning in the scientific domain. It's worth noting that ChatGPT may not have been explicitly trained on the news domain for the KPG task. However, during its training, other tasks in the news domain were included, and ChatGPT may have reused useful knowledge from these tasks to improve its performance on KPG.

4.4 Biomedical Domain

To assess the models' ability to adapt to the biomedical domain, we utilized the KPBioMed and PubMed datasets, which represent specialized subfields of scientific literature. These datasets pose significant challenges for the models due to their domain-specific content and varying document lengths. Table 6 provides a summary of the results.

In the comparison of KeyBART and ChatGPT's performance on the KPBioMed and PubMed datasets, KeyBART shows stronger results in abstractive keyphrase generation, while ChatGPT excels in extractive keyphrase generation.

For present keyphrases, KeyBART achieves higher precision on the KPBioMed dataset, as seen in its better F1@5 score (0.116 vs. ChatGPT's 0.103). This suggests that KeyBART, with its pretraining on scientific texts, is more effective at accurately predicting top keyphrases from shorter documents in the biomedical domain. However, in longer documents from the PubMed dataset, ChatGPT significantly outperforms KeyBART, particularly in the F1@M score (0.433 vs. 0.134). ChatGPT's ability to process larger contexts gives it an edge in handling more complex information across extensive text, leading to more accurate keyphrase generation.

For absent keyphrases, KeyBART consistently surpasses ChatGPT across both datasets. On KPBioMed, KeyBART achieves an F1@5 of 0.075, compared to ChatGPT's 0.051, and on PubMed, the difference is even more pronounced, with KeyBART scoring 0.076 vs. ChatGPT's 0.007. This suggests that while ChatGPT excels at generating extractive keyphrases, particularly from longer documents, KeyBART's specialized training enables it to generate more meaningful abstractive keyphrases. This advantage may be due to the fact that KeyBART predicts around 20% of its keyphrases for this domain in an abstractive manner, whereas ChatGPT's predictions are almost exclusively extractive, giving KeyBART a distinct edge in scenarios requiring deeper abstraction.

Highlighting Color	Associated Significance
Yellow	True keyphrase unpredicted by both models
Light Blue	True keyphrase correctly predicted by both models
Dark Blue	True keyphrase correctly predicted by KeyBART only
Green	True keyphrase correctly predicted by ChatGPT only

Table 7 Corresponding Significance of Each Highlight Color

5 Case Studies

This section presents several case studies, including both short and long scientific documents, as well as a news example. Our goal is to demonstrate how ChatGPT can enhance the Keyphrase Generation (KPG) task and highlight its benefits in real-world scenarios. For each sample, we used both the KeyBART and ChatGPT models to generate the keyphrases. To make it easier to view our results, we have color-coded the keyphrases in the accompanying images according to the scheme specified in Table 7. It is important to note that many of the keyphrases that were labeled as abstractive are actually extractive. Additionally, in short documents, most of the keyphrases labeled as abstractive appear later in the full document, underscoring the importance of processing the entire long document rather than just its summarized abstract.

5.1 Case Study 1: Short Scientific Documents

5.1.1 Case Study 1.1

This case study presents a short scientific document, specifically a sample of the Inspec test dataset with ID 2166, as shown in Figure 2.

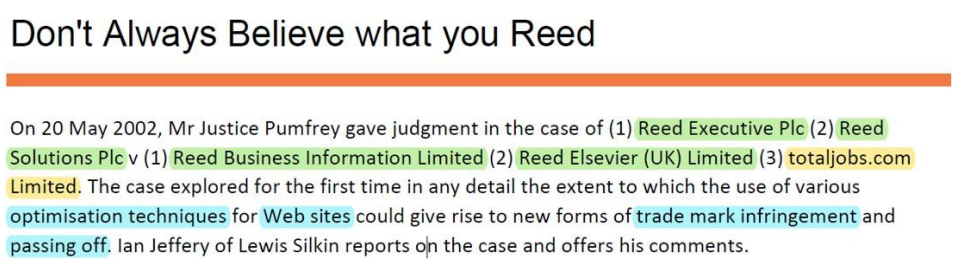


Fig. 2 Case Study 1.1: Short Scientific Document from Inspec Test Dataset

As observed, KeyBART was unable to predict half of the KPs in its specialized domain. We hypothesize that due to the document's brevity, the model struggles to identify the keywords' significance. However, ChatGPT was not hindered by this limitation and utilized its pretraining knowledge to assign importance to these phrases.

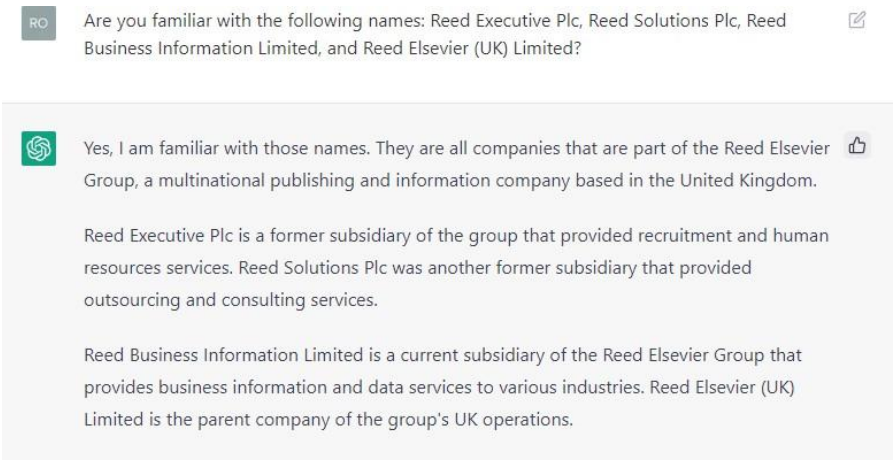
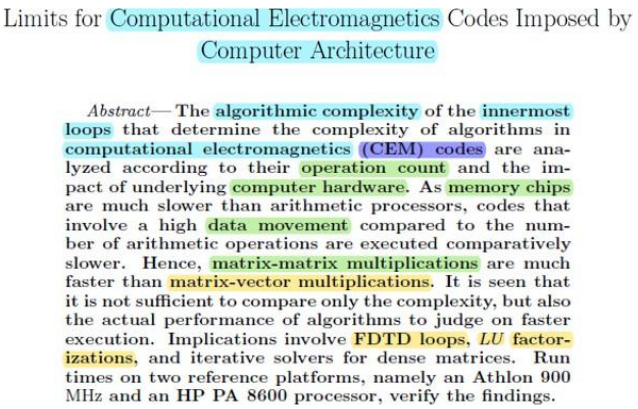


Fig. 3 Case Study 1.1: ChatGPT's knowledge in the field directly questioned


Limits for Computational Electromagnetics Codes Imposed by Computer Architecture

Abstract— The algorithmic complexity of the innermost loops that determine the complexity of algorithms in computational electromagnetics (CEM) codes are analyzed according to their operation count and the impact of underlying computer hardware. As memory chips are much slower than arithmetic processors, codes that involve a high data movement compared to the number of arithmetic operations are executed comparatively slower. Hence, matrix-matrix multiplications are much faster than matrix-vector multiplications. It is seen that it is not sufficient to compare only the complexity, but also the actual performance of algorithms to judge on faster execution. Implications involve FDTD loops, *LU* factorizations, and iterative solvers for dense matrices. Run times on two reference platforms, namely an Athlon 900 MHz and an HP PA 8600 processor, verify the findings.

Fig. 4 Case Study 1.2: Short Scientific Document from Inspec Test Dataset

We queried ChatGPT directly using its front-end to determine if it was familiar with the keyphrases exclusively predicted by it, and the response can be seen in Figure 3. The model not only recognized the terms but also grouped them together and provided a brief introduction, demonstrating how the model's knowledge from other tasks and pretraining objectives can be useful for KPG.

5.1.2 Case Study 1.2

The sample document chosen for this case study is a short scientific piece shown in Figure 4, sourced from the Inspec test dataset and bearing the identification number 2043.

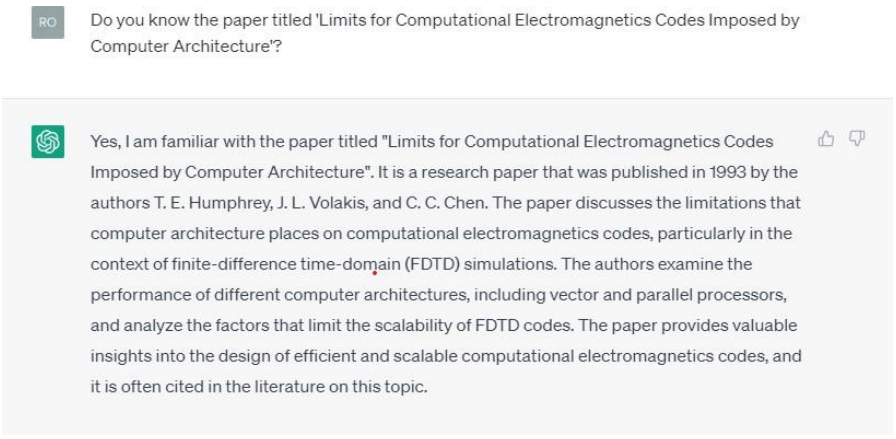


Fig. 5 Case Study 1.2: Direct Inquiry of ChatGPT's Familiarity with the Paper

Interestingly, ChatGPT exhibits superior performance in this abstract, which contains no repeated keyphrases. Such documents can be more challenging, as the importance of a word cannot be inferred from its redundant references. It is worth mentioning that these keyphrases are reiterated throughout the entire document, emphasizing the need for a comprehensive understanding of the full text to accurately identify keyphrases.

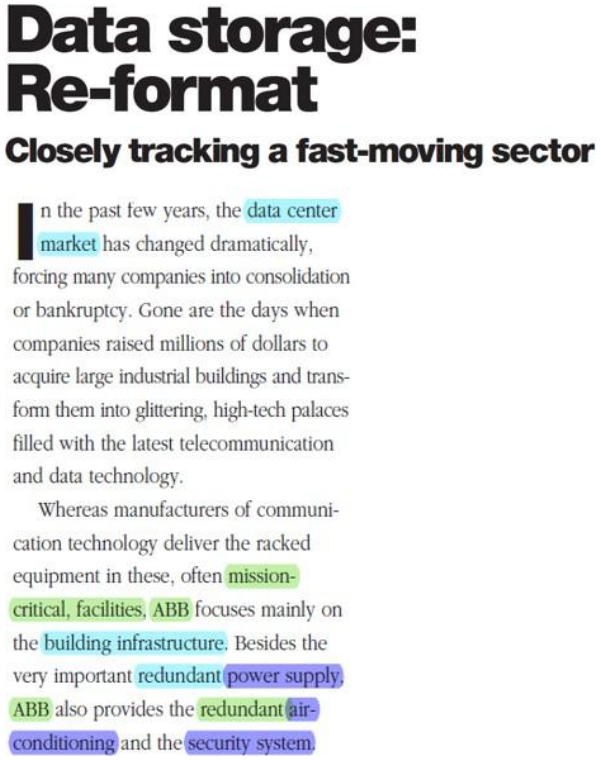
KeyBART's performance excels in the initial sections of the document but declines in the middle, and neither model successfully predicts any keyphrases in the latter parts. This observation suggests that the final sections are more complex, as the significance of a word cannot be deduced from its earlier mentions. This limitation impacts KeyBART more severely than ChatGPT.

In this instance, as demonstrated in Figure 5, we directly asked ChatGPT if it was familiar with the paper. Although the model could not correctly predict specific details from the document, such as the authors' names, it successfully summarized the paper's main points. This implies that ChatGPT possesses accurate latent knowledge of the article acquired during its pre-training, which in turn enhances its performance in the keyphrase generation task.

5.1.3 Case Study 1.3

This case study features another short scientific article, identified by number 2150 and sourced from the Inspec test dataset, as illustrated in Figure 6.

KeyBART and ChatGPT display comparable performance in this case, with a low intersection between their correct predictions. KeyBART succeeds in accurately predicting generic phrases from the scientific domain, such as "power supply," whereas ChatGPT is more adept at identifying specific terms, like the company name "ABB," likely due to its industry knowledge. Interestingly, both models have several inaccurately predicted keyphrases in common, such as "Data Storage," which could potentially be a keyphrase that the author overlooked. Keyphrase identification can sometimes be ambiguous.



Data storage: Re-format

Closely tracking a fast-moving sector

In the past few years, the data center market has changed dramatically, forcing many companies into consolidation or bankruptcy. Gone are the days when companies raised millions of dollars to acquire large industrial buildings and transform them into glittering, high-tech palaces filled with the latest telecommunication and data technology.

Whereas manufacturers of communication technology deliver the racked equipment in these, often mission-critical, facilities, ABB focuses mainly on the building infrastructure. Besides the very important redundant power supply, ABB also provides the redundant air-conditioning and the security system.

Fig. 6 Case Study 1.3: Short Scientific Document from Inspec Test Dataset

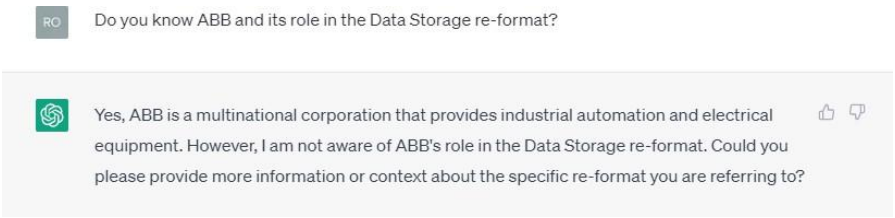


Fig. 7 Case Study 1.3: Investigating ChatGPT's Familiarity with the Company and Topic Discussed in the Paper

We asked the model directly about ABB, a company involved in data storage re-formatting. Figure 7 displays the results, which show that while the model accurately describes the company, it lacks knowledge about the specific topic described in the article. This instance represents ChatGPT's lowest performance on the Inspec dataset. It seems that the model relies heavily on its inner knowledge of the document, which may explain its lower accuracy in cases where it lacks specific information.

5.1.4 Human Evaluation of Case Study 1

Across all three case studies involving short scientific documents, ChatGPT generally outperformed KeyBART due to its broader pretraining knowledge and ability to handle abstractive keyphrases. KeyBART struggled in highly specialized domains and with shorter texts, where its performance relied heavily on the explicit repetition of terms or simpler extractive patterns. ChatGPT, by contrast, demonstrated more flexibility, using its latent knowledge to infer keyphrases even when they were not explicitly repeated or clearly extractive. However, both models showed limitations: KeyBART in handling domain-specific phrases, and ChatGPT in scenarios where its pretraining knowledge was less relevant or when precise, document-specific details were required. Overall, ChatGPT's adaptability and broader knowledge base gave it an edge in the keyphrase generation task, especially for complex or highly specialized documents, while KeyBART excelled in more generic scientific content.

5.2 Case Study 2: Long Scientific Documents

5.2.1 Case Study 2.1

This case evaluates the models' performance on long scientific documents, specifically a sample of the SemEval2010 test dataset with ID 'C-17'. Its title and abstract are shown in Figure 8.

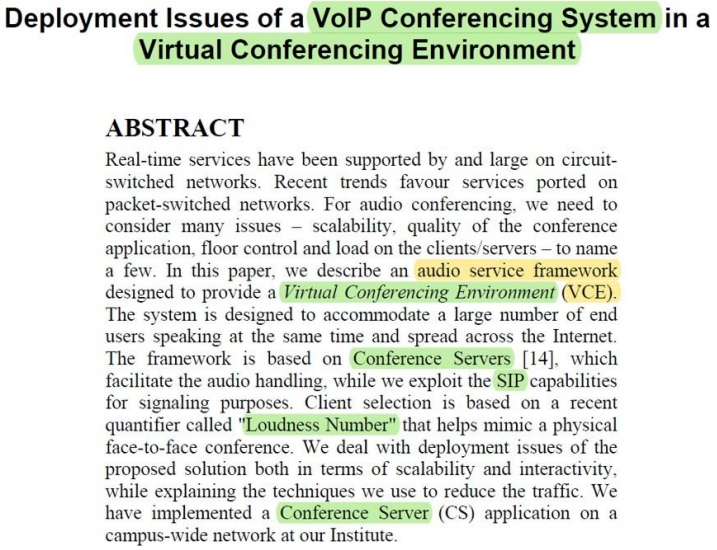
In this example, the lack of context in the title and abstract made it impossible for KeyBART to identify any keyword. However, due to its pre-training and field knowledge, it was still able to generate coherent abstractive keyphrases such as 'teleconferencing' and 'collaborative virtual environment.' In the case of long documents, the limitations posed by the maximum input tokens may prevent the model from comprehending long-term relationships between words necessary for generating keyphrases. The higher maximum input tokens in ChatGPT have contributed to its superior performance. As shown in Figure 9, keyphrases such as 'SIP' and 'Conference Server' gain greater importance in later parts of the document, such as the conclusion, which cannot be extrapolated without a holistic view of the entire document.

Domain knowledge gained from training on other tasks is another key component that explains ChatGPT's superior performance in such documents. Figure 10 illustrates how the model is capable of providing a coherent response when asked a question derived from the article's title, demonstrating the crucial role of domain knowledge in KPG from long articles.

5.2.2 Case Study 2.2

This example pertains to an extensive scientific paper from the SemEval2010 test dataset, identified as 'I-10'. The title and abstract can be viewed in Figure 11.

Observably, KeyBART struggles to identify 'incremental learning' from the title and fails to recognize the repeated term 'agent.' In contrast, ChatGPT accurately predicts both elements. Nevertheless, KeyBART does partially capture 'agent' with predictions such as 'multi-agent system' and 'autonomous agent.' Furthermore, it



Deployment Issues of a VoIP Conferencing System in a Virtual Conferencing Environment

ABSTRACT

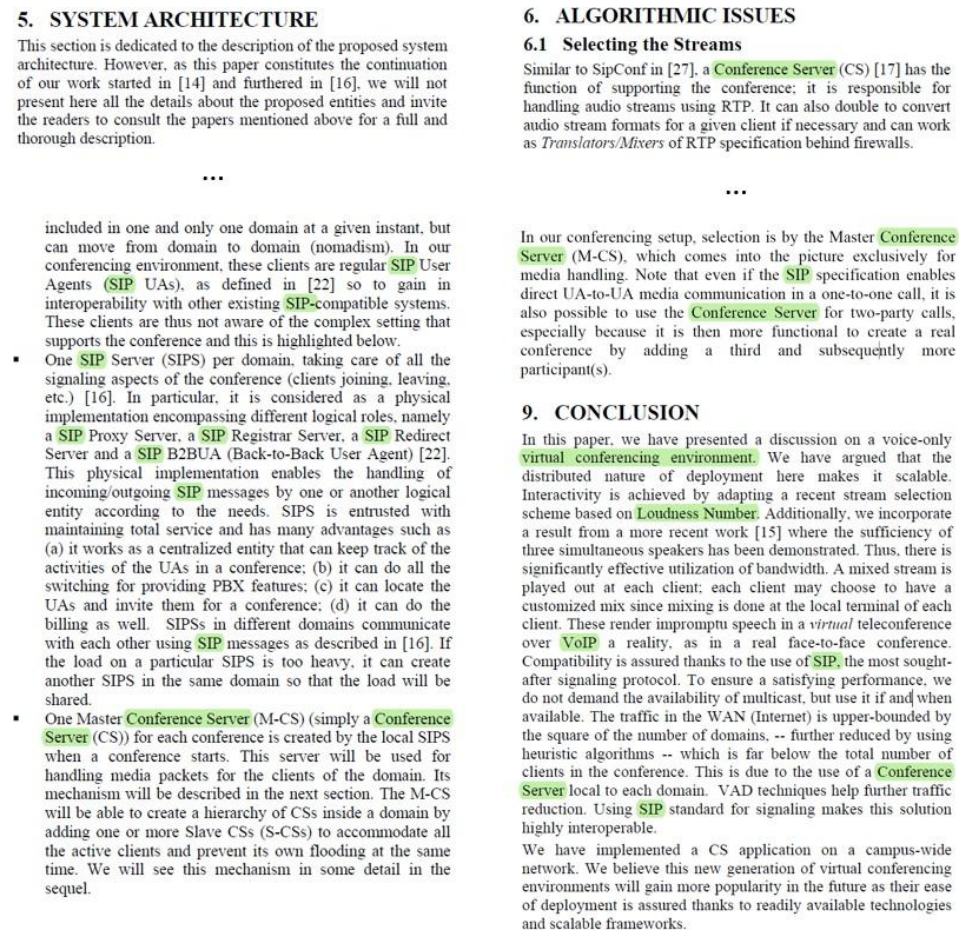
Real-time services have been supported by and large on circuit-switched networks. Recent trends favour services ported on packet-switched networks. For audio conferencing, we need to consider many issues – scalability, quality of the conference application, floor control and load on the clients/servers – to name a few. In this paper, we describe an audio service framework designed to provide a *Virtual Conferencing Environment* (VCE). The system is designed to accommodate a large number of end users speaking at the same time and spread across the Internet. The framework is based on Conference Servers [14], which facilitate the audio handling, while we exploit the SIP capabilities for signaling purposes. Client selection is based on a recent quantifier called "Loudness Number" that helps mimic a physical face-to-face conference. We deal with deployment issues of the proposed solution both in terms of scalability and interactivity, while explaining the techniques we use to reduce the traffic. We have implemented a Conference Server (CS) application on a campus-wide network at our Institute.

Fig. 8 Case Study 2.1: Long Scientific Document from SemEval2010 Test Dataset - Title & Abstract

accurately anticipates the abstractive keyphrase 'mas-consistency' through alternative expressions like 'mas-consistent learning,' which ChatGPT did not manage to recognize.

Key portions of the document, as well as the conclusion, can be found in Figure 12 Notably, both models failed to identify 'knowledge' as a keyphrase, despite its repetition more than 10 times throughout the document. This oversight could result from the models lacking full document context, causing them to perceive 'knowledge' as a common word. Another instance is KeyBART's inability to recognize 'Incremental Learning,' which appears in significant sections later in the document, such as a subsection title. This limitation might be related to its narrow token context or a lack of familiarity with the paper's specialized area.

In order to comprehend the implicit knowledge contained in the document produced by ChatGPT, we requested that it provide us with a brief summary of the key concepts it covers. The resulting summary is displayed in Figure 13. However, the model incorrectly characterized the collaborative learning method outlined in the document as reinforcement learning. Despite this error, the model did correctly identify that the primary focus of the paper was artificial intelligence, indicating that its prediction was not entirely incorrect. Additionally, we directly queried the model regarding its understanding of the difference between collaborative learning and reinforcement learning, and the results are shown in 14. The model demonstrated a clear understanding of the distinction between the two terms, suggesting that its misidentification of the paper's content may have been an informed guess.



5. SYSTEM ARCHITECTURE

This section is dedicated to the description of the proposed system architecture. However, as this paper constitutes the continuation of our work started in [14] and furthered in [16], we will not present here all the details about the proposed entities and invite the readers to consult the papers mentioned above for a full and thorough description.

...

included in one and only one domain at a given instant, but can move from domain to domain (nomadism). In our conferencing environment, these clients are regular SIP User Agents (SIP UAs), as defined in [22] so to gain in interoperability with other existing SIP-compatible systems. These clients are thus not aware of the complex setting that supports the conference and this is highlighted below.

- One SIP Server (SIPS) per domain, taking care of all the signaling aspects of the conference (clients joining, leaving, etc.) [16]. In particular, it is considered as a physical implementation encompassing different logical roles, namely a SIP Proxy Server, a SIP Registrar Server, a SIP Redirect Server and a SIP B2BUA (Back-to-Back User Agent) [22]. This physical implementation enables the handling of incoming/outgoing SIP messages by one or another logical entity according to the needs. SIPS is entrusted with maintaining total service and has many advantages such as (a) it works as a centralized entity that can keep track of the activities of the UAs in a conference; (b) it can do all the switching for providing PBX features; (c) it can locate the UAs and invite them for a conference; (d) it can do the billing as well. SIPSs in different domains communicate with each other using SIP messages as described in [16]. If the load on a particular SIPS is too heavy, it can create another SIPS in the same domain so that the load will be shared.
- One Master Conference Server (M-CS) (simply a Conference Server (CS)) for each conference is created by the local SIPS when a conference starts. This server will be used for handling media packets for the clients of the domain. Its mechanism will be described in the next section. The M-CS will be able to create a hierarchy of CSs inside a domain by adding one or more Slave CSs (S-CSs) to accommodate all the active clients and prevent its own flooding at the same time. We will see this mechanism in some detail in the sequel.

6. ALGORITHMIC ISSUES

6.1 Selecting the Streams

Similar to SipConf in [27], a Conference Server (CS) [17] has the function of supporting the conference; it is responsible for handling audio streams using RTP. It can also double to convert audio stream formats for a given client if necessary and can work as *Translators/Mixers* of RTP specification behind firewalls.

...

In our conferencing setup, selection is by the Master Conference Server (M-CS), which comes into the picture exclusively for media handling. Note that even if the SIP specification enables direct UA-to-UA media communication in a one-to-one call, it is also possible to use the Conference Server for two-party calls, especially because it is then more functional to create a real conference by adding a third and subsequently more participant(s).

9. CONCLUSION

In this paper, we have presented a discussion on a voice-only virtual conferencing environment. We have argued that the distributed nature of deployment here makes it scalable. Interactivity is achieved by adapting a recent stream selection scheme based on Loudness Number. Additionally, we incorporate a result from a more recent work [15] where the sufficiency of three simultaneous speakers has been demonstrated. Thus, there is significantly effective utilization of bandwidth. A mixed stream is played out at each client; each client may choose to have a customized mix since mixing is done at the local terminal of each client. These render impromptu speech in a *virtual* teleconference over VoIP a reality, as in a real face-to-face conference. Compatibility is assured thanks to the use of SIP, the most sought-after signaling protocol. To ensure a satisfying performance, we do not demand the availability of multicast, but use it if and when available. The traffic in the WAN (Internet) is upper-bounded by the square of the number of domains, -- further reduced by using heuristic algorithms -- which is far below the total number of clients in the conference. This is due to the use of a Conference Server local to each domain. VAD techniques help further traffic reduction. Using SIP standard for signaling makes this solution highly interoperable.

We have implemented a CS application on a campus-wide network. We believe this new generation of virtual conferencing environments will gain more popularity in the future as their ease of deployment is assured thanks to readily available technologies and scalable frameworks.

Fig. 9 Case Study 2.1: Long Scientific Document from SemEval2010 Test Dataset - Parts of the Body & Conclusion

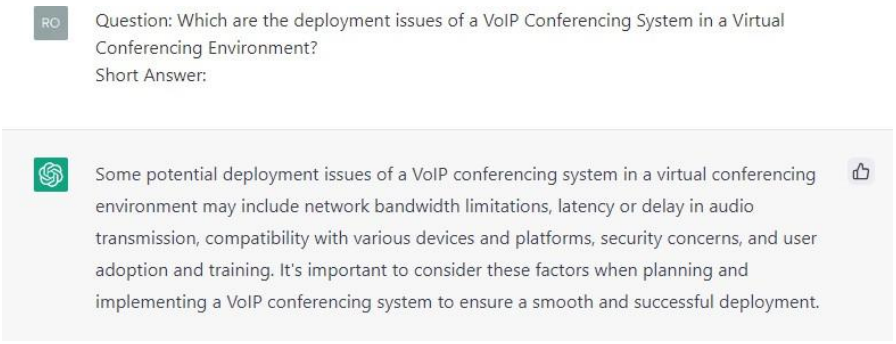
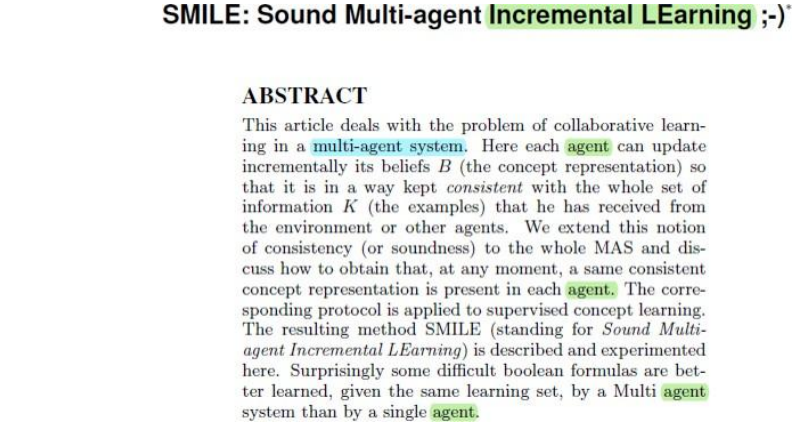


Fig. 10 Case Study 2.1: ChatGPT's knowledge in the domain of the article

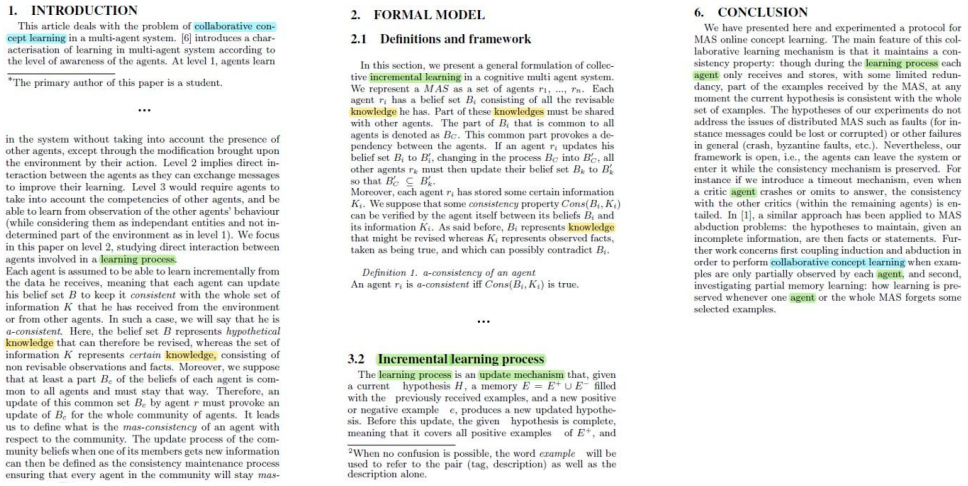


SMILE: Sound Multi-agent Incremental LEarning ;-)*

ABSTRACT

This article deals with the problem of collaborative learning in a multi-agent system. Here each agent can update incrementally its beliefs B (the concept representation) so that it is in a way kept *consistent* with the whole set of information K (the examples) that he has received from the environment or other agents. We extend this notion of consistency (or soundness) to the whole MAS and discuss how to obtain that, at any moment, a same consistent concept representation is present in each agent. The corresponding protocol is applied to supervised concept learning. The resulting method SMILE (standing for *Sound Multi-agent Incremental LEarning*) is described and experimented here. Surprisingly some difficult boolean formulas are better learned, given the same learning set, by a Multi agent system than by a single agent.

Fig. 11 Case Study 2.2: Long Scientific Document from SemEval2010 Test Dataset - Title & Abstract



1. INTRODUCTION

This article deals with the problem of collaborative concept learning in a multi-agent system. [6] introduces a characterisation of learning in multi-agent system according to the level of awareness of the agents. At level 1, agents learn

*The primary author of this paper is a student.

...

in the system without taking into account the presence of other agents, except through the modification brought upon the environment by their action. Level 2 implies direct interaction between the agents as they can exchange messages to improve their learning. Level 3 would require agents to take into account the competencies of other agents, and be able to learn from observation of the other agents' behaviour (while considering them as independant entities and not indetermined part of the environment as in level 1). We focus in this paper on level 2, studying direct interaction between agents involved in a learning process.
Each agent is assumed to be able to learn incrementally from the data he receives, meaning that each agent can update his belief set B to keep it *consistent* with the whole set of information K that he has received from the environment or from other agents. In such a case, we will say that he is *a-consistent*. Here, the belief set B represents *hypothetical* knowledge that can therefore be revised, whereas the set of information K represents *certain* knowledge, consisting of non revisable observations and facts. Moreover, we suppose that at least a part B_c of the beliefs of each agent is common to all agents and must stay that way. Therefore, an update of this common set B_c by agent r must provoke an update of B_c for the whole community of agents. It leads us to define what is the *mas-consistency* of an agent with respect to the community. The update process of the community beliefs when one of its members gets new information can then be defined as the consistency maintenance process ensuring that every agent in the community will stay *mas-*

2. FORMAL MODEL

2.1 Definitions and framework

In this section, we present a general formulation of collective incremental learning in a cognitive multi agent system. We represent a MAS as a set of agents $r_1, ..., r_n$. Each agent r_i has a belief set B_i consisting of all the revisable knowledge he has. Part of these knowledges must be shared with other agents. The part of B_i that is common to all agents is denoted as B_C. This common part provokes a dependency between the agents. If an agent r_i updates his belief set B_i to B'_i, changing in the process B_C into B'_C, all other agents r_k must then update their belief set B_k to B'_k so that $B'_C \subseteq B'_k$.
Moreover, each agent r_i has stored some certain information K_i. We suppose that some *consistency* property $Cons(B_i, K_i)$ can be verified by the agent itself between its beliefs B_i and its information K_i. As said before, B_i represents knowledge that might be revised whereas K_i represents observed facts, taken as being true, and which can possibly contradict B_i.

Definition 1. a-consistency of an agent
An agent r_i is *a-consistent* iff $Cons(B_i, K_i)$ is true.

...

3.2 Incremental learning process

The learning process is an update mechanism that, given a current hypothesis H, a memory $E = E^+ \cup E^-$ filled with the previously received examples, and a new positive or negative example e, produces a new updated hypothesis. Before this update, the given hypothesis is complete, meaning that it covers all positive examples of E^+, and

[2]When no confusion is possible, the word *example* will be used to refer to the pair (tag, description) as well as the description alone.

6. CONCLUSION

We have presented here and experimented a protocol for MAS online concept learning. The main feature of this collaborative learning mechanism is that it maintains a consistency property: though during the learning process each agent only receives and stores, with some limited redundancy, part of the examples received by the MAS, at any moment the current hypothesis is consistent with the whole set of examples. The hypotheses of our experiments do not address the issues of distributed MAS such as faults (for instance messages could be lost or corrupted) or other failures in general (crash, byzantine faults, etc.). Nevertheless, our framework is open, i.e., the agents can leave the system or enter it while the consistency mechanism is preserved. For instance if we introduce a timeout mechanism, even when a critic agent crashes or omits to answer, the consistency with the other critics (within the remaining agents) is entailed. In [1], a similar approach has been applied to MAS abduction problems: the hypotheses to maintain, given an incomplete information, are then facts or statements. Further work concerns first coupling induction and abduction in order to perform collaborative concept learning when examples are only partially observed by each agent, and second, investigating partial memory learning: how learning is preserved whenever one agent or the whole MAS forgets some selected examples.

Fig. 12 Case Study 2.2: Long Scientific Document from SemEval2010 Test Dataset - Parts of the Body & Conclusion

5.2.3 Case Study 2.3

This instance relates to a long scientific document in the SemEval2010 evaluation dataset, designated as 'I-10'. The title and abstract are observable in Figure 15.

In this example, both models accurately identify the keyphrases covered in the document's title and abstract. However, as shown in Figure 16, neither can identify phrases located at crucial points in the document, such as 'Reputation-based adaption' or 'commitment-based semantics', the latter even being included in the conclusion.

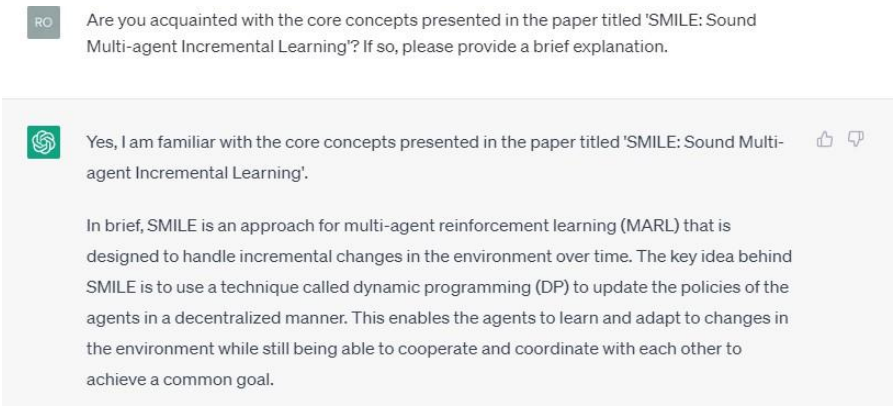


Fig. 13 Case Study 2.2: ChatGPT's Familiarity with the Paper

We speculate that this may be due to the lack of long-term relationship embedding mechanisms in both models.

It's worth noting that KeyBART outperforms ChatGPT in this sample. KeyBART can correctly identify all the keyphrases that ChatGPT predicts correctly and, in addition, correctly predicts the extractive keyphrase 'state transition systems' and comes close to correctly predicting the abstractive keyphrase 'social notion', with predictions such as 'social science' or 'social attitude'.

In Figure 17, we prompt ChatGPT to provide definitions of the keyphrases it did not predict by contextualizing the field they belong to. Surprisingly, ChatGPT can correctly recognize the field and provide accurate definitions for both phrases, despite not being able to identify them as keyphrases initially.

5.2.4 Human Evaluation of Case Study 2

For long scientific documents, ChatGPT consistently outperformed KeyBART due to its larger token limit and pretraining, allowing it to capture keyphrases that appear throughout the entire document, such as in conclusions. KeyBART struggled with token limitations, often missing keyphrases that required a more holistic view. However, KeyBART showed strength in extractive keyphrase prediction, occasionally outperforming ChatGPT on specific technical terms. Both models struggled with repetitive terms, like "knowledge," suggesting limitations in recognizing their significance. Overall, ChatGPT's broader context understanding made it better suited for long documents, while KeyBART excelled in extractive cases.

5.3 Case Study 3: News Domain

5.3.1 Case Study 3.1

In this study, we assess the model's ability to generalize across domains by analyzing a sample from the DUC2001 dataset with ID 'AP891006-0029'. As depicted in Figure 18, the article belongs to the sports news domain, which is markedly different from the

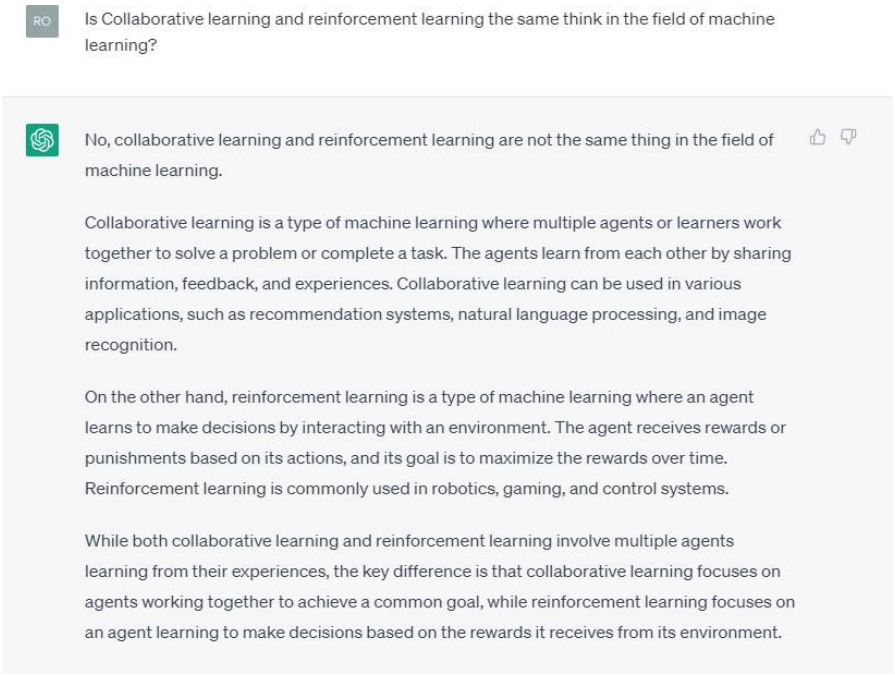


Fig. 14 Case Study 2.2: Examining ChatGPT's Comprehension of Reinforcement and Collaborative Learning

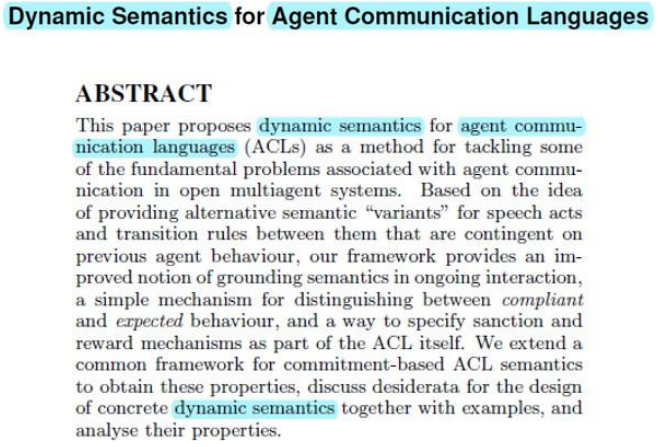
scientific domain. This allows us to evaluate the model's capacity to handle a broad range of domains beyond scientific literature.

The drop in KeyBART's performance was expected given that the domain of the sample article is substantially different from its training data. However, the model was still able to accurately generate the scientific keyphrase 'anabolic steroids'. The model's incorrectly generated keyphrases, such as 'biomedical and behavioral research' and 'human factors', demonstrate that it still approached the article as if it were in the scientific domain, which is coherent given that this is where the model was trained. Regarding ChatGPT, its multidomain training enables it to be more robust, as it can leverage the knowledge it acquired from various tasks to generate high-quality keyphrases in any domain that was included in its training.

As the article covers an important event in sports history, we can directly test the model's knowledge of it. Figure 19 shows that ChatGPT is able to describe the events described in the article by posing a simple question, indicating that the model has latent knowledge that goes beyond the textual content. This capability enables the model to identify relevant keyphrases more accurately.

5.3.2 Case Study 3.2

This case study evaluates the model's ability to generalize across various domains by analyzing a specific instance from the DUC2001 dataset identified as 'WSJ910529-0003'.

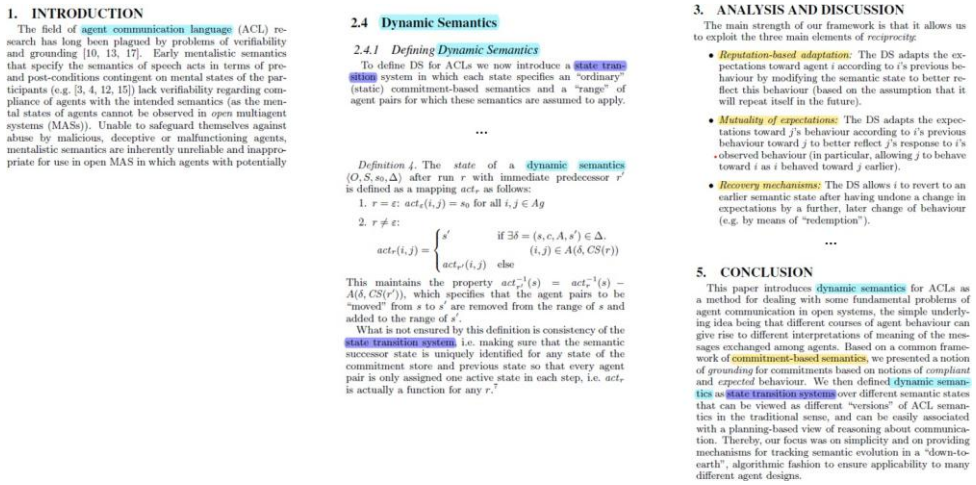


Dynamic Semantics for Agent Communication Languages

ABSTRACT

This paper proposes dynamic semantics for agent communication languages (ACLs) as a method for tackling some of the fundamental problems associated with agent communication in open multiagent systems. Based on the idea of providing alternative semantic "variants" for speech acts and transition rules between them that are contingent on previous agent behaviour, our framework provides an improved notion of grounding semantics in ongoing interaction, a simple mechanism for distinguishing between *compliant* and *expected* behaviour, and a way to specify sanction and reward mechanisms as part of the ACL itself. We extend a common framework for commitment-based ACL semantics to obtain these properties, discuss desiderata for the design of concrete dynamic semantics together with examples, and analyse their properties.

Fig. 15 Case Study 2.3: Long Scientific Document from SemEval2010 Test Dataset - Title & Abstract



1. INTRODUCTION

The field of agent communication language (ACL) research has long been plagued by problems of verifiability and grounding [10, 13, 17]. Early mentalistic semantics that specify the semantics of speech acts in terms of pre- and post-conditions contingent on mental states of the participants (e.g. [3, 4, 12, 15]) lack verifiability regarding compliance of agents with the intended semantics (as the mental states of agents cannot be observed in *open* multiagent systems (MASs)). Unable to safeguard themselves against abuse by malicious, deceptive or malfunctioning agents, mentalistic semantics are inherently unreliable and inappropriate for use in open MAS in which agents with potentially

2.4 Dynamic Semantics

2.4.1 *Defining Dynamic Semantics*

To define DS for ACLs we now introduce a state transition system in which each state specifies an "ordinary" (static) commitment-based semantics and a "range" of agent pairs for which these semantics are assumed to apply.

...

Definition 4. The *state* of a dynamic semantics $\langle O, S, s_0, \Delta \rangle$ after run r with immediate predecessor r' is defined as a mapping act_r as follows:

1. $r = \varepsilon$: $act_\varepsilon(i,j) = s_0$ for all $i, j \in Ag$
2. $r \neq \varepsilon$:

$$act_r(i,j) = \begin{cases} s' & \text{if } \exists \delta = (s, c, A, s') \in \Delta. \\ & \quad (i,j) \in A(\delta, CS(r)) \\ act_{r'}(i,j) & \text{else} \end{cases}$$

This maintains the property $act_r^{-1}(s) = act_{r'}^{-1}(s) - A(\delta, CS(r'))$, which specifies that the agent pairs to be "moved" from s to s' are removed from the range of s and added to the range of s'.

What is not ensured by this definition is consistency of the state transition system, i.e. making sure that the semantic successor state is uniquely identified for any state of the commitment store and previous state so that every agent pair is only assigned one active state in each step, i.e. act_r is actually a function for any r.[7]

3. ANALYSIS AND DISCUSSION

The main strength of our framework is that it allows us to exploit the three main elements of *reciprocity*.

- *Reputation-based adaptation:* The DS adapts the expectations toward agent i according to i's previous behaviour by modifying the semantic state to better reflect this behaviour (based on the assumption that it will repeat itself in the future).
- *Mutuality of expectations:* The DS adapts the expectations toward j's behaviour according to i's previous behaviour toward j to better reflect j's response to i's observed behaviour (in particular, allowing j to behave toward i as i behaved toward j earlier).
- *Recovery mechanisms:* The DS allows i to revert to an earlier semantic state after having undone a change in expectations by a further, later change of behaviour (e.g. by means of "redemption").

...

5. CONCLUSION

This paper introduces dynamic semantics for ACLs as a method for dealing with some fundamental problems of agent communication in open systems, the simple underlying idea being that different courses of agent behaviour can give rise to different interpretations of meaning of the messages exchanged among agents. Based on a common framework of commitment-based semantics, we presented a notion of *grounding* for commitments based on notions of *compliant* and *expected* behaviour. We then defined dynamic semantics as state transition systems over different semantic states that can be viewed as different "versions" of ACL semantics in the traditional sense, and can be easily associated with a planning-based view of reasoning about communication. Thereby, our focus was on simplicity and on providing mechanisms for tracking semantic evolution in a "down-to-earth", algorithmic fashion to ensure applicability to many different agent designs.

Fig. 16 Case Study 2.3: Long Scientific Document from SemEval2010 Test Dataset - Parts of the Body & Conclusion

The article, which belongs to the gossip press domain and discusses a famous actress's medical condition, is shown in Figure 20.

As demonstrated, ChatGPT can accurately predict almost all the keyphrases, whereas KeyBART fails to predict any. KeyBART generates mostly abstractive keyphrases related to the medical domain, such as 'medical records' and 'occupational safety', which is expected since it is constrained to generate keyphrases only from the

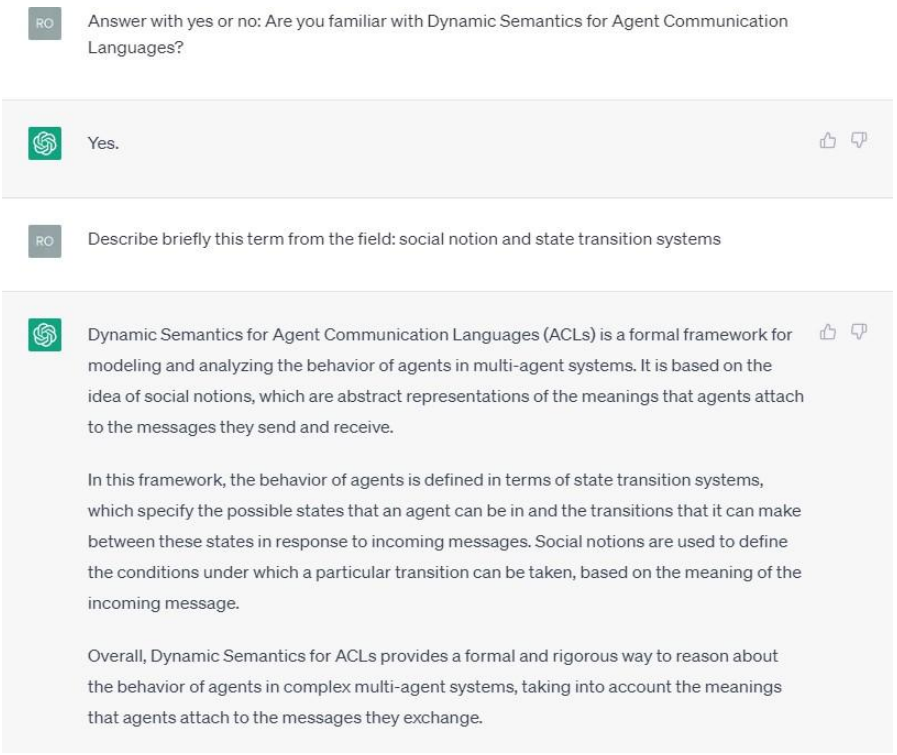


Fig. 17 Case Study 2.3: Analyzing ChatGPT's Understanding of Unpredicted Keyphrases in the Document

Ben Johnson, the world-class sprinter who was knocked off track and field's pedestal after testing positive for steroids, says it's wrong for athletes to use the muscle-building substance. "I got caught in Seoul. I lost my gold medal," the Canadian told reporters as legislation to classify anabolic steroids as a controlled substance was introduced on Thursday. "I'm here to tell the people of this country it's wrong to cheat, not to take it. It's bad for your health." Watching Johnson was his chief nemesis: Carl Lewis, the man who was awarded the Olympic gold medals Johnson lost. "I think it's great," Lewis said of the legislation. "They're making a move and it's very positive. I'm happy to see it." However, the flamboyant, pony-tailed Lewis told reporters: "I don't understand why Ben Johnson's here." Lewis said he attended the news conference because he was working on his autobiography, and one of the chapters deals with steroids. He said he didn't intend to upstage Johnson. Reps. Mel Levine, D-Calif., Henry Waxman, D-Calif., and Benjamin Gilman, R-N.Y., invited Johnson to attend as they presented the legislation that would make anabolic steroids a controlled substance, similar to the designation given to cocaine and heroin. The lawmakers emphasized the increasing abuse of steroids by college, high school, and even junior high school athletes who believe the substance will enhance their performance. "America is about to have an adolescent time-bomb explode in its hands," Levine said. "But if we act quickly enough to restrict steroid distribution, and to increase the penalties for illicit distribution, we can prevent this plague from spreading." Levine referred to the abuse of steroids as "the silent side of the drug disease in this country." He applauded Johnson's courage for attending the news conference. Johnson later stepped up to the microphones and in a quiet voice with a slight stutter told other athletes not to make the same mistake he did, urging them "to come forward, to come clean." Lewis dismissed reporters' questions that he had used steroids, indicating he would be willing to run against Johnson "if he comes back and he's clean." While Lewis held the spotlight, Johnson slipped away to adjoining congressional offices. "He's not here to compete with Carl Lewis. I hope he will someday," said Ed Futerman, Johnson's lawyer.

Fig. 18 Case Study 3.1: News from DUC2001 Dataset



Fig. 19 Case Study 3.1: ChatGPT's knowledge from the events described in the article

The lawsuit, which has been pending for nine months, arose out of two articles published by the Enquirer in June 1990 reporting on Miss Taylor's condition and activities at St. John's Hospital, Santa Monica, Calif., where she was treated last spring for pneumonia. The Enquirer said that after gaining access to all of Miss Taylor's medical records, it is satisfied that the articles reporting on the actress's medical condition and the report that she was drinking were in error. The paper said it published the articles in good faith reliance on information provided to it, but the information was inaccurate. Iain Calder, Enquirer editor, said in a statement, "we regret the inaccuracies in the articles but are pleased that this dispute has come to an amicable end." Miss Taylor said she feels completely vindicated," and that after the newspaper's management determined the articles were in error, the Enquirer "acted promptly and in good faith." Miss Taylor initially sought damages of $20 million in Los Angeles Superior Court, according to Neil Papiano, her attorney. Although Mr. Papiano wouldn't specify the size of the settlement, he said we were persuaded that it was certainly large enough that we shouldn't go to trial." As previously reported, G.P. Group Inc., the Lantana, Fla., publisher of the Enquirer and Star tabloids, plans to raise $350 million by offering 43% of the firm in an initial public offering.

Fig. 20 Case Study 3.2: News from DUC2001 Dataset

domain included in its training. Therefore, it cannot adapt well to a new domain that is far from its training data.

We evaluated ChatGPT's knowledge by asking it directly about the actress and the events mentioned in the article. The results are shown in Figure 21, which reveals that the model has knowledge about the events, but it struggles with details such

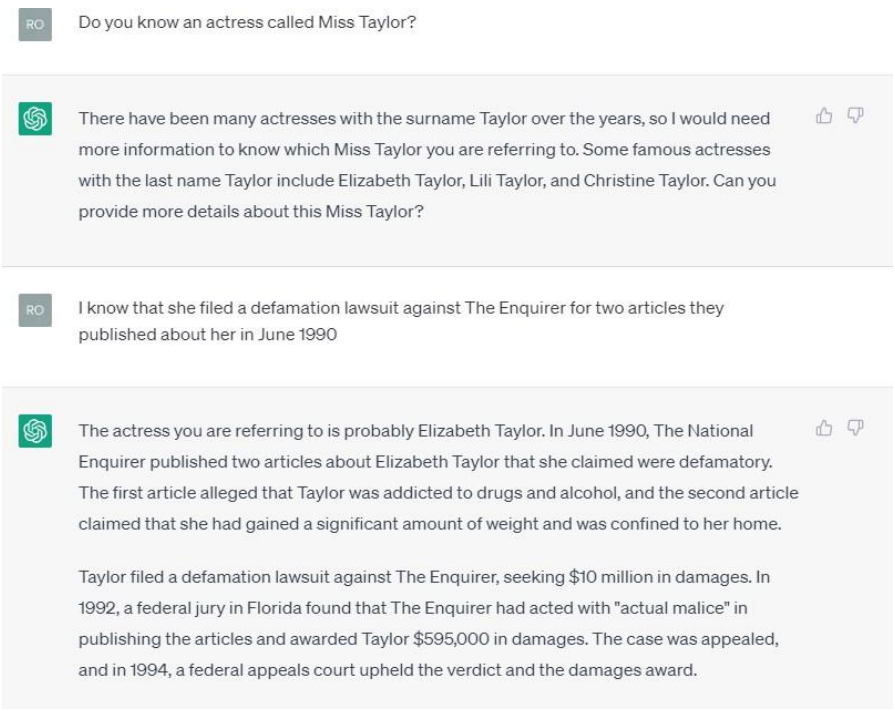


Fig. 21 Case Study 3.2: ChatGPT's Knowledge of the Actress and the Events Described in the Article

as monetary values and locations. This knowledge is obtained from ChatGPT's pre-training on other tasks, and it significantly enhances the model's performance in the KPG task.

5.3.3 Case Study 3.3

This case study examines the model's capacity to generalize across different domains by analyzing a specific instance identified as 'LA103089-0070' from the DUC2001 dataset. The article, which belongs to the aviation or military news domain, is displayed in Figure 22 and is one of the closest articles to the scientific domain that can be found in the dataset.

In this example, both models demonstrate a low level of performance, with ChatGPT's performance slightly superior to that of KeyBART's. KeyBART's predictions consist of both abstractive and extractive keyphrases from the scientific domain, indicating that the model is not resilient enough to function in an untrained domain. Some of its predictions, such as 'air traffic control' or 'suicide prevention', are not relevant to the text. The keyphrases 'crash' and 'victim' are frequently repeated, but neither of the models can identify their importance.

In this sample, we directly asked ChatGPT to explain the reasons behind its poor performance in this specific article. The corresponding prompt and answer are depicted in Figures 23 and 24, respectively. The model attributes its low performance to the

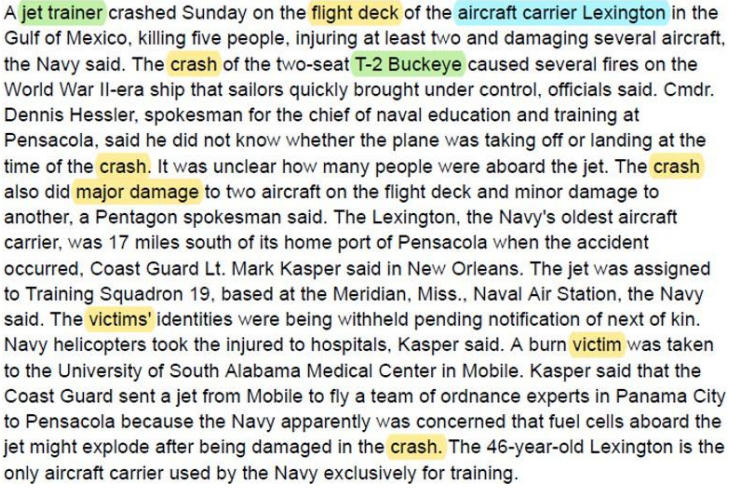


Fig. 22 Case Study 3.3: News from DUC2001 Dataset

lack of a clear focus, limited context, and limited vocabulary, which are plausible explanations that have been repeatedly identified as challenges in KPG.

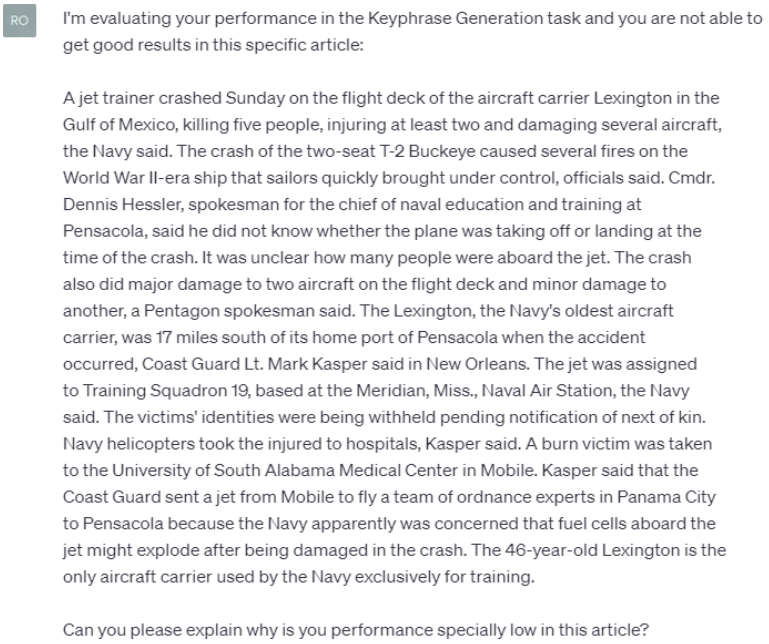


Fig. 23 Case Study 3.3: Directly questioning ChatGPT about its low KPG performance in this article

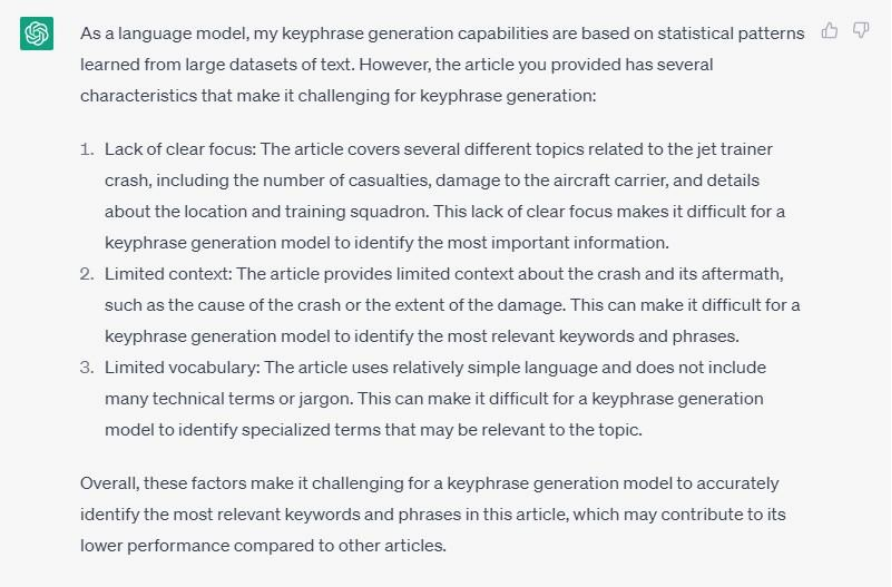


Fig. 24 Case Study 3.3: ChatGPT's response to why it performed poorly in KPG in this article

5.3.4 Human Evaluation of Case Study 3

Across the news domain case studies, ChatGPT consistently outperformed KeyBART due to its multidomain pretraining, enabling it to adapt better to various topics outside the scientific field. KeyBART, trained primarily on scientific data, struggled to generalize and produced irrelevant keyphrases, particularly in the sports and gossip press articles. In contrast, ChatGPT's broader knowledge allowed it to predict keyphrases more accurately, even in unfamiliar domains like aviation and military news. However, both models faced challenges when articles had unclear focus or limited context, with neither fully capturing critical repetitive terms like "crash" and "victim." Overall, ChatGPT's flexibility in handling diverse content made it more effective for keyphrase generation in the news domain.

5.4 Case Study 4: Biomedical Documents

5.4.1 Case Study 4.1

This case study analyzes a biomedical paper from the PubMed test dataset, specifically a sample originally published in the PubMed journal [71]. The evaluation shows that KeyBART performed poorly, correctly predicting only the generic keyword "cancer." Despite its limitations, KeyBART generated other relevant but broad keyphrases, such as "cell cycle," "oncology," and "molecular medicine," which, although not included in the original labels, could serve as suitable keyphrases for a broader, more general context, as illustrated in figure 25.

Primary lung cancer is the leading cause of cancer deaths in most industrialized countries, and non-small-cell lung cancer (NSCLC) accounts for 75–80% of primary lung cancer. Surgery is the most effective therapeutic modality for the cure, but the postoperative prognosis remains poor.[1–3] To improve the prognosis, establishment of biological markers other than tumor-node-metastasis (TNM) factors that determine prognosis and response to a particular treatment is essential. Although many possible biological markers such as mutations of *p53* and *k-ras* genes have been reported, none of them has been established as a marker in decision-making of the treatment of NSCLC. In addition, most NSCLC patients present with advanced stages and are appropriate candidates for surgery. For inoperable patients, chemotherapy with or without radiotherapy may be given. Recently, tyrosine kinase inhibitors targeting epidermal growth factor receptor (EGFR), gefitinib and erlotinib, have been introduced into the therapy of NSCLC, and experimental and clinical studies have revealed that these agents induce dramatic antitumor effects for tumors with activating mutations within the EGFR kinase domain.[4,5] Thus, other molecular abnormalities in NSCLC should be revealed in the development of useful prognostic markers and in the development of effective molecular targeting agents.

Fig. 26 Case Study 4.1: Biomedical paper introduction

Perimembrane Aurora-A Expression is a Significant Prognostic Factor in Correlation with Proliferative Activity in Non-Small-Cell Lung Cancer (NSCLC)

Eiji Ogawa, MD,[1,2] Kazumasa Takenaka, MD, PhD,[1] Hiromichi Katakura, MD, PhD,[1] Masashi Adachi, MD,[1] Yosuke Otake, MD, PhD,[1] Yoshinobu Toda, PhD,[3] Hirokazu Kotani, MD, PhD,[2,3] Toshiaki Manabe, MD, PhD,[2] Hiromi Wada, MD, PhD,[1] and Fumihiro Tanaka, MD, PhD[✉1]

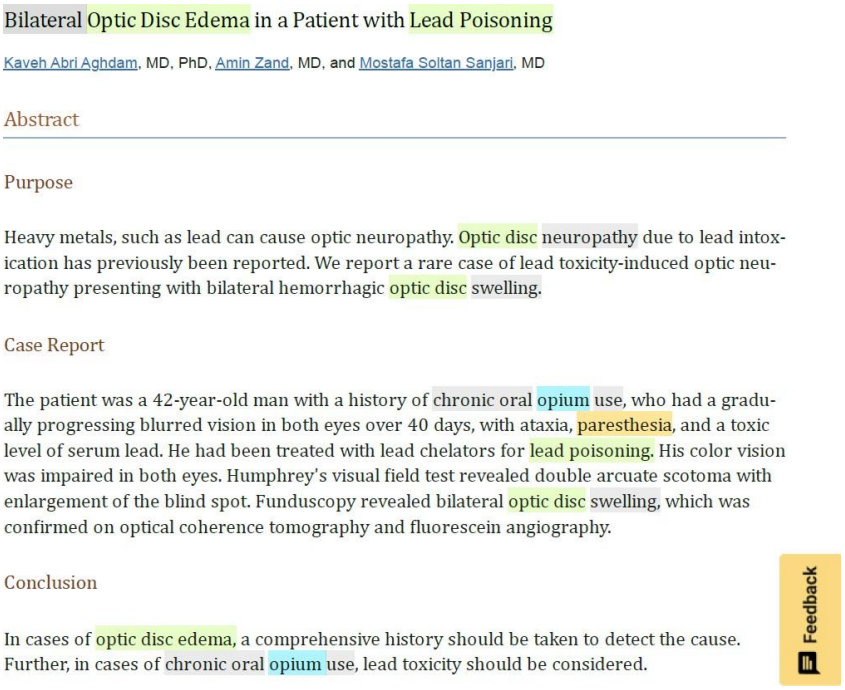
Abstract

Purpose

Aurora-A, also known as STK15/BTAK, is a member of the protein serine/threonine kinase family, and experimental studies have revealed that Aurora-A plays critical roles in cell mitosis and in carcinogenesis. However, no clinical studies on Aurora-A expression in non-small-cell lung cancer (NSCLC) have been reported. Thus, the present study was conducted to assess the clinical significance of Aurora-A status.

Fig. 25 Case Study 4.1: Biomedical paper title and abstract

The term "Aurora-A" is a central concept in the document, appearing frequently, including in the title, and being mentioned 127 times throughout, as seen in figure 26. Despite its prominence, only ChatGPT was able to accurately predict this keyphrase. As a specialized term within the biomedical field, "Aurora-A" likely required domain-specific knowledge to identify its significance, something KeyBART seemed to lack despite the repeated occurrences. Furthermore, ChatGPT not only recognized "Aurora-A" but also generated related keyphrases, such as "Aurora-A expression," which provided a more comprehensive understanding of the document's overall context. This demonstrates ChatGPT's stronger grasp of the document's core themes in long documents and technical content.



Bilateral Optic Disc Edema in a Patient with Lead Poisoning

Kaveh Abri Aghdam, MD, PhD, Amin Zand, MD, and Mostafa Soltan Sanjari, MD

Abstract

Purpose

Heavy metals, such as lead can cause optic neuropathy. Optic disc neuropathy due to lead intoxication has previously been reported. We report a rare case of lead toxicity-induced optic neuropathy presenting with bilateral hemorrhagic optic disc swelling.

Case Report

The patient was a 42-year-old man with a history of chronic oral opium use, who had a gradually progressing blurred vision in both eyes over 40 days, with ataxia, paresthesia, and a toxic level of serum lead. He had been treated with lead chelators for lead poisoning. His color vision was impaired in both eyes. Humphrey's visual field test revealed double arcuate scotoma with enlargement of the blind spot. Funduscopy revealed bilateral optic disc swelling, which was confirmed on optical coherence tomography and fluorescein angiography.

Conclusion

In cases of optic disc edema, a comprehensive history should be taken to detect the cause. Further, in cases of chronic oral opium use, lead toxicity should be considered.

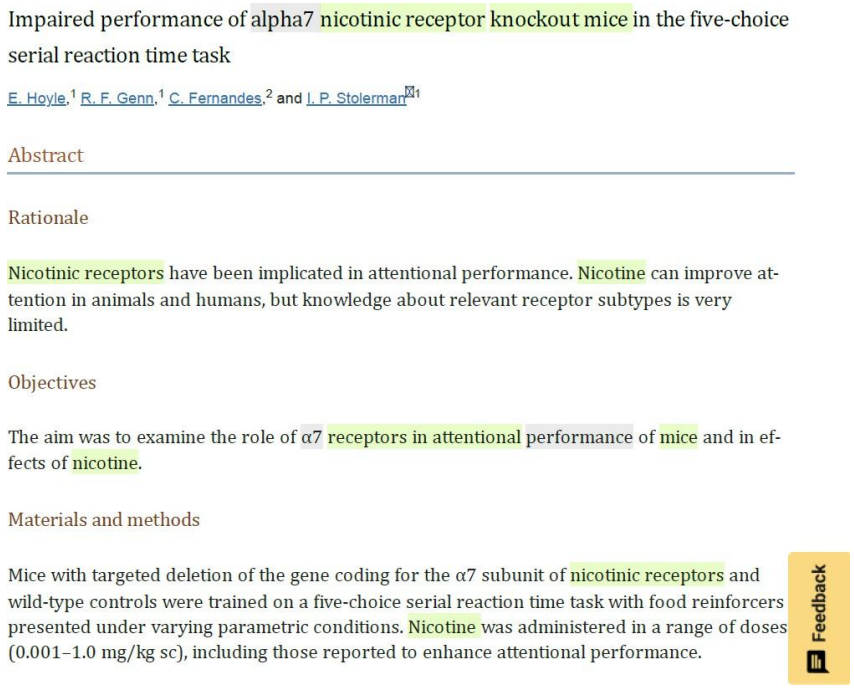
Feedback

Fig. 27 Case Study 4.2: Biomedical paper title and abstract

Neither model predicted "cancer" as a keyphrase. Its infrequent appearance in the document suggests that both models failed to recognize its significance as a key term. Identifying "cancer" as a crucial concept would require a deeper understanding of the document's context within the biomedical domain, something that both KeyBART and ChatGPT lacked in this case. This highlights the necessity for strong latent knowledge to grasp the importance of terms like "cancer," even when they are not explicitly repeated often.

5.4.2 Case Study 4.2

This case study examines a biomedical abstract from the KPBioMed test dataset, specifically a sample originally published in the PubMed journal [72], which can be found in figure 27. The evaluation shows that both models performed well, successfully identifying most of the keyphrases. However, they often did so in combination with other terms, rather than isolating the keyphrases on their own. For instance, both models predicted "Optic Disc" alongside words like "bilateral," "neuropathy," or "swelling." Since "Optic Disc" is the common denominator and keyphrases are meant to provide concise summaries of the article, it stands out as the more appropriate candidate. However, annotators may still disagree on which terms should or should not be included in these keyphrases, highlighting the inherent ambiguity in keyphrase extraction and generation tasks.



Impaired performance of alpha7 nicotinic receptor knockout mice in the five-choice serial reaction time task

E. Hoyle,[1] R. F. Genn,[1] C. Fernandes,[2] and I. P. Stolerman[1]

Abstract

Rationale

Nicotinic receptors have been implicated in attentional performance. Nicotine can improve attention in animals and humans, but knowledge about relevant receptor subtypes is very limited.

Objectives

The aim was to examine the role of α7 receptors in attentional performance of mice and in effects of nicotine.

Materials and methods

Mice with targeted deletion of the gene coding for the α7 subunit of nicotinic receptors and wild-type controls were trained on a five-choice serial reaction time task with food reinforcers presented under varying parametric conditions. Nicotine was administered in a range of doses (0.001–1.0 mg/kg sc), including those reported to enhance attentional performance.

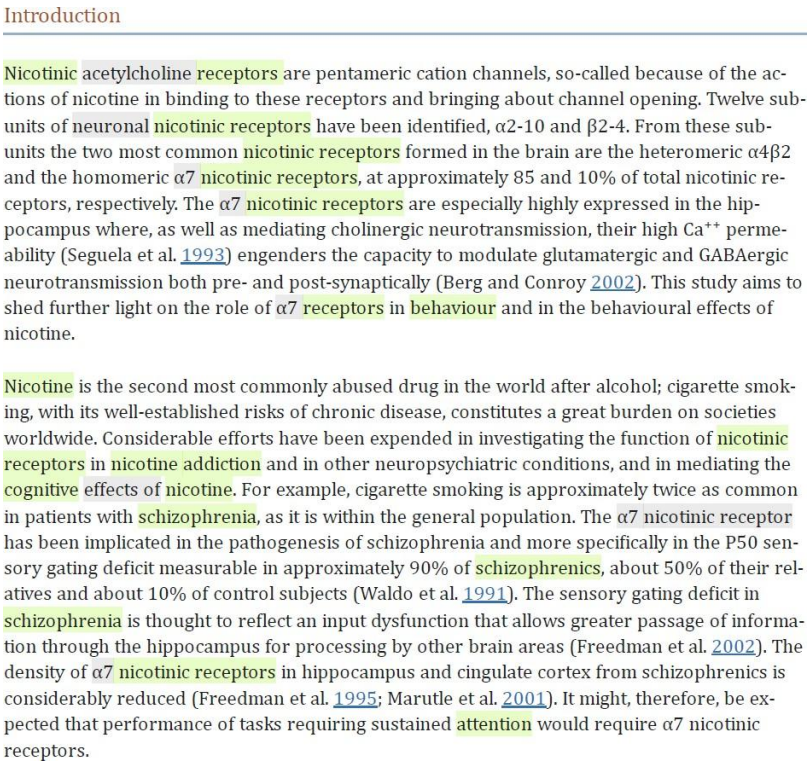
Feedback

Fig. 28 Case Study 4.3: Biomedical paper title and abstract

Both models failed to predict "paresthesia," a term that, based solely on the abstract, might not immediately stand out as a keyphrase. However, when considering the entire document, its repetition and central role in the article's content make it a clear candidate for keyphrase inclusion. This underscores the importance of taking a holistic view of the document for effective keyphrase extraction and generation, as certain terms may only reveal their significance when the full context is considered.

5.4.3 Case Study 4.3

This example focuses on a biomedical paper from the PubMed test dataset, specifically a sample originally published in Psychopharmacology (Berl) [73]. Both models performed exceptionally well, as illustrated in figure 28 and figure 29, accurately predicting all keyphrases labeled by the authors. Notably, both models identified "alpha7" as a keyphrase, which is a highly relevant term. Upon further investigation, it becomes clear that many databases also use "alpha7" as a keyphrase, confirming its significance as one of the most important terms in the article. This makes it particularly useful for downstream KPE tasks, such as document retrieval. Additionally, it is worth noting that the models reordered the keyphrases differently, even grouping four of them together. This demonstrates not only the inherent ambiguity in defining what constitutes a keyphrase but also how their order and overlap can vary across predictions.



Introduction

Nicotinic acetylcholine receptors are pentameric cation channels, so-called because of the actions of nicotine in binding to these receptors and bringing about channel opening. Twelve subunits of neuronal nicotinic receptors have been identified, α2-10 and β2-4. From these subunits the two most common nicotinic receptors formed in the brain are the heteromeric α4β2 and the homomeric α7 nicotinic receptors, at approximately 85 and 10% of total nicotinic receptors, respectively. The α7 nicotinic receptors are especially highly expressed in the hippocampus where, as well as mediating cholinergic neurotransmission, their high Ca^{++} permeability (Seguela et al. 1993) engenders the capacity to modulate glutamatergic and GABAergic neurotransmission both pre- and post-synaptically (Berg and Conroy 2002). This study aims to shed further light on the role of α7 receptors in behaviour and in the behavioural effects of nicotine.

Nicotine is the second most commonly abused drug in the world after alcohol; cigarette smoking, with its well-established risks of chronic disease, constitutes a great burden on societies worldwide. Considerable efforts have been expended in investigating the function of nicotinic receptors in nicotine addiction and in other neuropsychiatric conditions, and in mediating the cognitive effects of nicotine. For example, cigarette smoking is approximately twice as common in patients with schizophrenia, as it is within the general population. The α7 nicotinic receptor has been implicated in the pathogenesis of schizophrenia and more specifically in the P50 sensory gating deficit measurable in approximately 90% of schizophrenics, about 50% of their relatives and about 10% of control subjects (Waldo et al. 1991). The sensory gating deficit in schizophrenia is thought to reflect an input dysfunction that allows greater passage of information through the hippocampus for processing by other brain areas (Freedman et al. 2002). The density of α7 nicotinic receptors in hippocampus and cingulate cortex from schizophrenics is considerably reduced (Freedman et al. 1995; Marutle et al. 2001). It might, therefore, be expected that performance of tasks requiring sustained attention would require α7 nicotinic receptors.

Fig. 29 Case Study 4.3: Biomedical paper added context

5.4.4 Human Evaluation of Case Study 4

In biomedical documents, ChatGPT consistently outperformed KeyBART, particularly when handling specialized terms like "Aurora-A," which KeyBART failed to recognize despite its frequency in the text. ChatGPT's broader pretraining and domain-specific knowledge enabled it to accurately generate key phrases, providing more comprehensive and contextually relevant keyphrase predictions. KeyBART, while able to produce generic keyphrases such as "oncology" and "molecular medicine," struggled with more specific technical terms, highlighting its limitations in specialized domains.

In other cases, both models performed similarly, but ChatGPT's ability to capture related phrases in addition to the core key terms gave it a slight edge. Both models struggled when the importance of certain terms, like "paresthesia," became clear only with a holistic view of the document. Overall, ChatGPT's superior domain knowledge made it more effective in generating accurate and context-aware keyphrases for biomedical content, while KeyBART performed well with more general terms but lacked precision in specialized contexts.

6 Conclusions and Future Work

As demonstrated in this study, ChatGPT consistently outperforms its counterparts across various benchmarks, particularly in handling long documents and non-scientific domains. Its superior performance can be attributed to the model's increased input token limit and its multidomain training approach, which allows it to leverage in-domain knowledge from other tasks and apply this for keyphrase generation. The use of a vast dataset further supports the model's ability to manage larger inputs without compromising the performance in keyphrase generation tasks.

These findings underscore the potential of multitask and multidomain learning in improving keyphrase generation quality, addressing challenges related to domain adaptation, and handling long document keyphrase generation. While all models exhibit performance degradation when dealing with long documents, ChatGPT's performance remains notable, and this study highlights the importance of developing more advanced methods for capturing long-term dependencies between words to provide a holistic understanding of the document.

Future work can explore several promising directions. For instance, ablation studies could be conducted to further elucidate the individual contributions of the various model components and training strategies. Additionally, while ChatGPT shows strong performance across many domains, a more detailed analysis in the biomedical domain suggests potential for improvement in certain abstractive keyphrase generation tasks, where domain-specific models like KeyBART perform competitively. This opens opportunities for further fine-tuning and domain-specific adaptation.

Another interesting area of future research is the role of prompt engineering. While this study focused on a fixed prompt design, exploring the effects of prompt variations could offer valuable insights into enhancing prompt-based learning, further refining the model's performance across different tasks and domains.

In conclusion, by investigating these directions, future work can continue to refine keyphrase generation models, further expanding their utility in both general and specialized applications.

Acknowledgments

We would like to express our gratitude to Debanjan Mahata, who served as our master and introduced us to the field of NLP and KPE. His guidance and patience have been invaluable throughout this research project, and we are grateful for his mentorship and support.

We would also like to thank Alejandro Pérez and Sergio Gago for providing the computational resources that were essential for developing and testing the ideas presented in this paper. Their generosity and support have been instrumental in the success of this research project.

Additionally, we acknowledge that an earlier version of this manuscript was made available as a preprint on arXiv[10]. We appreciate the broader research community for engaging with this preliminary version.

[10] https://arxiv.org/abs/2304.14177

Finally, we would like to acknowledge the countless individuals and organizations who have contributed to the field of NLP and Keyphrase Extraction, as their work has provided the foundation for this research. We are grateful for their ongoing efforts and dedication to advancing this field, and we hope that this paper will contribute to their ongoing work.

Thank you all for your contributions and support.

Statements and Declarations

Competing interests

The authors declare no competing interests related to this study. All methodologies, analyses, and interpretations of data were conducted independently and without influence from external entities. This research was purely academic and aimed at contributing to the existing body of knowledge on keyphrase generation.

Authors Contribution Statement

Roberto Martínez-Cruz played a pivotal role in shaping the study's conception and design. His expertise was instrumental in the development of the study's methodology and the analytical processing of the data. Alvaro J. López-López and José Portela were integral to the data gathering and compilation efforts. Additionally, their insightful critiques significantly enriched the study's intellectual depth. Each author contributed diligently to the manuscript's creation, and collectively, they have given their consent for the final manuscript to be published.

Ethical and Informed Consent

This research did not involve any human participants, data, or tissues, and therefore did not require ethical approval or informed consent. All datasets utilized in this study are publicly available and were openly published for research purposes, adhering to all applicable ethical standards.

Data Availability and Access

The datasets supporting the conclusions of this article are all publicly available and openly published for research purposes. Links to these datasets can be found in Section 3.1 of this paper. The availability of these datasets ensures transparency and allows for the reproducibility of the research findings.

References

[1] Hulth, A., Megyesi, B.B.: A study on automatically extracted keywords in text categorization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for

Computational Linguistics, pp. 537–544. Association for Computational Linguistics, Sydney, Australia (2006). https://doi.org/10.3115/1220175.1220243 . https://aclanthology.org/P06-1068

[2] Hammouda, K.M., Matute, D.N., Kamel, M.S.: Corephrase: Keyphrase extraction for document clustering. In: Perner, P., Imiya, A. (eds.) Machine Learning and Data Mining in Pattern Recognition, pp. 265–274. Springer, Berlin, Heidelberg (2005)

[3] Qazvinian, V., Radev, D.R., Özgür, A.: Citation summarization through keyphrase extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 895–903. Coling 2010 Organizing Committee, Beijing, China (2010). https://aclanthology.org/C10-1101

[4] Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications (2017)

[5] Sanyal, D.K., Bhowmick, P.K., Das, P.P., Chattopadhyay, S., Santosh, T.Y.S.S.: Enhancing access to scholarly publications with surrogate resources. Scientometrics **121**(2), 1129–1164 (2019) https://doi.org/10.1007/s11192-019-03227-

[6] Nguyen, T.D., Kan, M.-Y.: Keyphrase extraction in scientific publications. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, pp. 317–326. Springer, Berlin, Heidelberg (2007)

[7] Gollapalli, S.D., Li, X.-l., Yang, P.: Incorporating expert knowledge into keyphrase extraction. Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (2017) https://doi.org/10.1609/aaai.v31i1.10986

[8] Alzaidy, R., Caragea, C., Giles, C.L.: Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents. In: The World Wide Web Conference. WWW '19, pp. 2551–2557. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313642 . https://doi.org/10.1145/3308558.3313642

[9] Rungta, M., Kumar, R., Dhaliwal, M.P., Tiwari, H., Vala, V.: Transkp: Transformer based key-phrase extraction. 2020 International Joint Conference on Neural Networks (IJCNN), 1–7 (2020)

[10] Sahrawat, D., Mahata, D., Zhang, H., Kulkarni, M., Sharma, A., Gosangi, R., Stent, A., Kumar, Y., Shah, R.R., Zimmermann, R.: Keyphrase extraction as sequence labeling using contextualized embeddings. In: European Conference on Information Retrieval, pp. 328–335 (2020). Springer

[11] Kulkarni, M., Mahata, D., Arora, R., Bhowmik, R.: Learning rich representation of

keyphrases from text. In: Findings of the Association for Computational Linguistics: NAACL 2022, pp. 891–906. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.findings-naacl.67 . https://aclanthology.org/2022.findings-naacl.67

[12] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (2019)

[13] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (2020)

[14] Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training. (2018)

[15] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)

[16] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners (2020)

[17] Mihalcea, R., Tarau, P.: TextRank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-3252

[18] Bougouin, A., Boudin, F., Daille, B.: TopicRank: Graph-based topic ranking for keyphrase extraction. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 543–551. Asian Federation of Natural Language Processing, Nagoya, Japan (2013). https://aclanthology.org/I13-1062

[19] Wang, R., Liu, W., McDonald, C.: Corpus-independent generic keyphrase extraction using word embedding vectors. In: Software Engineering Research Conference, vol. 39, pp. 1–8 (2014)

[20] Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., Jaggi, M.: Simple unsupervised keyphrase extraction using sentence embeddings. arXiv preprint arXiv:1801.04470 (2018)

[21] Mahata, D., Kuriakose, J., Shah, R., Zimmermann, R.: Key2vec: Automatic

ranked keyphrase extraction from scientific articles using phrase embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 634–639 (2018)

[22] Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. EMNLP '03, pp. 216–223. Association for Computational Linguistics, USA (2003). https://doi.org/10.3115/1119355.1119383 . https://doi.org/10.3115/1119355.1119383

[23] Kim, S.N., Kan, M.-Y.: Re-examining automatic keyphrase extraction approaches in scientific articles. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009), pp. 9–16. Association for Computational Linguistics, Singapore (2009). https://aclanthology.org/W09-2902

[24] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv (2013). https://doi.org/10.48550/ARXIV.1301.3781 . https://arxiv.org/abs/1301.3781

[25] Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1162 . https://aclanthology.org/D14-1162

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv (2017). https://doi.org/10.48550/ARXIV.1706.03762 . https://arxiv.org/abs/1706.03762

[27] Martinc, M., Škrlj, B., Pollak, S.: TNT-KID: Transformer-based neural tagger for keyword identification. Natural Language Engineering **28**(4), 409–448 (2021) https://doi.org/10.1017/s1351324921000127

[28] Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., Chi, Y.: Deep keyphrase generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 582–592. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1054 . https://aclanthology.org/P17-1054

[29] Ye, H., Wang, L.: Semi-supervised learning for neural keyphrase generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4142–4153. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1447 . https://aclanthology.org/D18-1447

[30] Chen, J., Zhang, X., Wu, Y., Yan, Z., Li, Z.: Keyphrase Generation with Correlation Constraints (2018)

[31] Chen, W., Gao, Y., Zhang, J., King, I., Lyu, M.R.: Title-guided encoding for keyphrase generation. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19. AAAI Press, ??? (2019). https://doi.org/10.1609/aaai.v33i01.33016268 . https://doi.org/10.1609/aaai.v33i01.33016268

[32] Wang, Y., Li, J., Chan, H.P., King, I., Lyu, M.R., Shi, S.: Topic-aware neural keyphrase generation for social media language. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2516–2526. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1240 . https://aclanthology.org/P19-1240

[33] Zhao, J., Zhang, Y.: Incorporating linguistic constraints into keyphrase generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5224–5233. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1515 . https://aclanthology.org/P19-1515

[34] Chan, H.P., Chen, W., Wang, L., King, I.: Neural Keyphrase Generation via Reinforcement Learning with Adaptive Rewards (2019)

[35] Swaminathan, A., Zhang, H., Mahata, D., Gosangi, R., Shah, R.R., Stent, A.: A preliminary exploration of GANs for keyphrase generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8021–8030. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.645 . https://aclanthology.org/2020.emnlp-main.645

[36] Chen, W., Chan, H.P., Li, P., King, I.: Exclusive hierarchical decoding for deep keyphrase generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1095–1105. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.103 . https://aclanthology.org/2020.acl-main.103

[37] Yuan, X., Wang, T., Meng, R., Thaker, K., Brusilovsky, P., He, D., Trischler, A.: One size does not fit all: Generating and evaluating variable number of keyphrases. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7961–7975. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.710 . https://aclanthology.org/2020.acl-main.710

[38] Zhao, J., Bao, J., Wang, Y., Wu, Y., He, X., Zhou, B.: SGG: Learning to Select,

Guide, and Generate for Keyphrase Generation (2021)

[39] Huang, X., Xu, T., Jiao, L., Zu, Y., Zhang, Y.: Adaptive beam search decoding for discrete keyphrase generation. Proceedings of the AAAI Conference on Artificial Intelligence **35**(14), 13082–13089 (2021) https://doi.org/10.1609/aaai.v35i14.17546

[40] Ye, J., Gui, T., Luo, Y., Xu, Y., Zhang, Q.: One2Set: Generating diverse keyphrases as a set. 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), pp. 4598–4608. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.354 . https://aclanthology.org/2021.acl-long.354

[41] Chen, W., Chan, H.P., Li, P., Bing, L., King, I.: An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. ArXiv **abs/1904.03454** (2019)

[42] Ahmad, W., Bai, X., Lee, S., Chang, K.-W.: Select, extract and generate: Neural keyphrase generation with layer-wise coverage attention. 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), pp. 1389–1404. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.111 . https://aclanthology.org/2021.acl-long.111

[43] Wu, H., Liu, W., Li, L., Nie, D., Chen, T., Zhang, F., Wang, D.: UniKeyphrase: A Unified Extraction and Generation Framework for Keyphrase Prediction (2021)

[44] Wu, H., Ma, B., Liu, W., Chen, T., Nie, D.: Fast and constrained absent keyphrase generation by prompt-based learning. Proceedings of the AAAI Conference on Artificial Intelligence **36**(10), 11495–11503 (2022) https://doi.org/10.1609/aaai.v36i10.21402

[45] Song, M., Jiang, H., Shi, S., Yao, S., Lu, S., Feng, Y., Liu, H., Jing, L.: Is ChatGPT A Good Keyphrase Generator? A Preliminary Study (2023)

[46] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. (2019)

[47] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., ??? (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

[48] Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., Amodei, D.: Reward learning from human preferences and demonstrations in Atari (2018)

[49] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-Tuning Language Models from Human Preferences (2020)

[50] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.: Learning to summarize from human feedback (2022)

[51] Böhm, F., Gao, Y., Meyer, C.M., Shapira, O., Dagan, I., Gurevych, I.: Better Rewards Yield Better Summaries: Learning to Summarise Without References (2019)

[52] Wu, J., Ouyang, L., Ziegler, D.M., Stiennon, N., Lowe, R., Leike, J., Christiano, P.: Recursively Summarizing Books with Human Feedback (2021)

[53] Jaques, N., Ghandeharioun, A., Shen, J.H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., Picard, R.: Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog (2019)

[54] Kreutzer, J., Khadivi, S., Matusov, E., Riezler, S.: Can neural machine translation be improved with user feedback? In: North American Chapter of the Association for Computational Linguistics (2018)

[55] Lawrence, C., Riezler, S.: Improving a neural semantic parser by counterfactual learning from human bandit feedback. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1820–1830. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1169 . https://aclanthology.org/P18-1169

[56] Zhou, W., Xu, K.: Learning to Compare for Better Training and Evaluation of Open Domain Natural Language Generation Models (2020)

[57] Cho, W.S., Zhang, P., Zhang, Y., Li, X., Galley, M., Brockett, C., Wang, M., Gao, J.: Towards Coherent and Cohesive Long-form Text Generation (2019)

[58] Perez, E., Karamcheti, S., Fergus, R., Weston, J., Kiela, D., Cho, K.: Finding generalizable evidence by learning to convince q&a models. ArXiv **abs/1909.05863** (2019)

[59] Madaan, A., Tandon, N., Clark, P., Yang, Y.: Memory-assisted prompt editing to improve GPT-3 after deployment (2023)

[60] Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers), pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1031 . https://aclanthology.org/P18-1031

[61] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)

[62] Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.-W.: Unified Language Model Pre-training for Natural Language Understanding and Generation (2019)

[63] McCann, B., Keskar, N.S., Xiong, C., Socher, R.: The Natural Language Decathlon: Multitask Learning as Question Answering (2018)

[64] Keskar, N.S., McCann, B., Xiong, C., Socher, R.: Unifying Question Answering, Text Classification, and Regression via Span Extraction (2019)

[65] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation (2023)

[66] Kim, S.N., Medelyan, O., Kan, M.-Y., Baldwin, T.: Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10, pp. 21–26. Association for Computational Linguistics, USA (2010)

[67] Gallina, Y., Boudin, F., Daille, B.: KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents (2019)

[68] Wan, X., Xiao, J.: Single document keyphrase extraction using neighborhood knowledge. In: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2. AAAI'08, pp. 855–860. AAAI Press, ??? (2008)

[69] Houbre, M., Boudin, F., Daille, B.: A large-scale dataset for biomedical keyphrase generation. arXiv preprint arXiv:2211.12124 (2022)

[70] Schutz, A.T., et al.: Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods. M. App. Sc Thesis (2008)

[71] Ogawa, E., Takenaka, K., Katakura, H., Adachi, M., Otake, Y., Toda, Y., Kotani, H., Manabe, T., Wada, H., Tanaka, F.: Perimembrane aurora-a expression is a significant prognostic factor in correlation with proliferative activity in non-small-cell lung cancer (nsclc). Annals of Surgical Oncology **15**(2), 547–554 (2008) https://doi.org/10.1245/s10434-007-9653-8

[72] Abri Aghdam, K., Zand, A., Soltan Sanjari, M.: Bilateral optic disc edema in a patient with lead poisoning. Journal of Ophthalmic and Vision Research **14**(4),

513–517 (2019) https://doi.org/10.18502/jovr.v14i4.5465

[73] Hoyle, E., Genn, R.F., Fernandes, C., Stolerman, I.P.: Impaired performance of alpha7 nicotinic receptor knockout mice in the five-choice serial reaction time task. Psychopharmacology (Berl) **189**(2), 211–223 (2006) https://doi.org/10.1007/s00213-006-0549-2