Evaluating Unsupervised Argument Aligners via Generation of Conclusions of Structured Scientific Abstracts

Yingqiang Gao[†], **Nianlong Gu**[‡], **Jessica Lam**[†], **James Henderson**[§]
Richard H.R. Hahnloser[†]
[†]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland
{yingqiang.gao, lamjessica, rich}@ini.ethz.ch
[‡]Linguistic Research Infrastructure, University of Zurich, Switzerland
nianlong.gu@uzh.ch
[§]Idiap Research Institute, Switzerland
james.henderson@idiap.ch

Abstract

Scientific abstracts provide a concise summary of research findings, making them a valuable resource for extracting scientific arguments. In this study, we assess various unsupervised approaches for extracting arguments as aligned premise-conclusion pairs: semantic similarity, text perplexity, and mutual information. We aggregate structured abstracts from PubMed Central Open Access (PMCOA) papers published in 2022 and evaluate the argument aligners in terms of the performance of language models that we fine-tune to generate the conclusions from the extracted premise given as input prompts. We find that mutual information outperforms other measures on this task, suggesting that the reasoning process in scientific abstracts hinges mostly on linguistic constructs beyond simple textual similarity.[1]

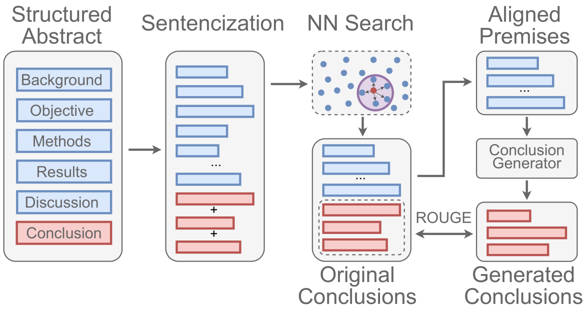

Figure 1: The evaluation pipeline for argument aligners. First, the structured abstract is split into premise and conclusion sentences. Then, the argument aligner uses nearest neighbor search to find relevant premises for conclusions. Finally, a trained language model generates conclusions from the selected premises. The best aligner is the one that selects the most sufficient premises for generated conclusions with the highest ROUGE score, compared to the original conclusions.

1 Introduction

Scientific reasoning involves pairing conclusions with premises, which encompasses information such as pre-existing knowledge, observations, and experimental results (Hesse, 1974; Al Khatib et al., 2021). This reasoning process is inherently directional: While inductive reasoning establishes logical links from the causal premises to the resulting conclusions (Gao et al., 2022), abductive reasoning aligns the most plausible premises for given conclusions (Ovchinnikova et al., 2014; Young et al., 2022; Li et al., 2023; Zhao et al., 2023).

The goal of mining scientific arguments is to discover the argumentative structure within academic papers (Binder et al., 2022). Despite the great success in recent studies (Fergadis et al., 2021; Wadden et al., 2022a,b), a crucial aspect of evaluating the alignment quality involves assessing the logical strength and quality of arguments (Kees et al., 2021; Wachsmuth et al., 2017), which entails determining the sufficiency of an argument's premises for deriving its conclusions. Normally, sufficient premises furnish comprehensive details for deducing conclusions, whereas insufficient premises lack essential prerequisites, making them compatible with flawed conclusions. Being able to assess argument sufficiency would not only allow the identification of well-argumented premise-conclusion pairs, but also help with evaluating the argument aligners that were used in the first place to pair premises and conclusions (Gurcke et al., 2021).

In this work, inspired by previous studies on text alignment (Nikolov and Hahnloser, 2019; Jiang et al., 2020), we investigate the sufficiency of premises aligned by various unsupervised argument aligners, i.e. normalized point-wise mutual information (npmi, Bouma (2009); Padmakumar and He (2021)), normalized perplexity (nppl, Miaschi et al. (2021)), and semantic (cosine) similar-

[1] Code and data available at https://github.com/CharizardAcademy/ARG-ALIGN.git

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 151–160
March 17-22, 2024 ©2024 Association for Computational Linguistics

ity (csim, Reimers and Gurevych (2019)). Drawing inspiration from Johnson and Blair (2006) and Wright et al. (2022), we assess the sufficiency of premises by evaluating the extent (measured with ROUGE score) to which a language model can generate the paired conclusion from them.

Our main **contributions** are: 1) We constructed a dataset named ARG-ALIGN, which comprises more than 17k pairs of premises and conclusions aggregated from structured scientific abstracts from the PubMed Central Open Access (PMCOA) corpus; 2) We assessed the sufficiency of the aligned premises by reconstructing the corresponding conclusions using language models; 3) We highlighted that premises in scientific abstracts may contain redundant information in terms of the drawn conclusions.

2 Unsupervised Argument Aligners

Given an abstract that contains a premise segment of $n \geq 5$ sentences $\mathcal{P} = (p_i)_{i=1}^n$ and a conclusion segment $\mathcal{C}$, unsupervised argument aligners compute alignment scores $d(p, \mathcal{C})$ between each premise sentence p and the entire conclusion segment $\mathcal{C}$. We set ourselves the goal of finding the $k = 5$ premise sentences $\mathcal{P}_k^* = (p_{i_j})_{j=1}^k$ that are most relevant to $\mathcal{C}$ in terms of their relatedness, as judged by a language model.

We consider the conclusion segment $\mathcal{C}$ as a single text rather than as a list of individual sentences because a paper typically has one primary research finding that is stated over possibly multiple conclusion sentences. The argument aligners therefore should identify premise sentences that are relevant to inferring $\mathcal{C}$ as a whole.

In contrast to previous studies that focused on inductive argument alignment, where $\mathcal{C}$ is identified based on $\mathcal{P}$ (Wadden et al., 2020), we focus on abductive argument alignment, where $\mathcal{P}$ is identified based on $\mathcal{C}$. This choice is motivated by the fact that the conclusion sentences in structured abstracts can be easily located by searching for the CONCLUSIONS discourse section using regular expressions, whereas premise sentences are distributed across all discourse sections and therefore more difficult to identify.

To abductively align a premise sentence p with the conclusion segment $\mathcal{C}$, we explore four unsupervised argument aligners with different alignment scores:

csim Semantic relevance using embedding-based cosine similarity.

$$\mathrm{csim}(p, \mathcal{C}) = 1 - \frac{e_p \cdot e_{\mathcal{C}}}{\|e_p\| \cdot \|e_{\mathcal{C}}\|},$$

where

$$e_p = \frac{1}{|p|} \sum_{w_p \in p} e(w_p), \quad e_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{w_c \in \mathcal{C}} e(w_c)$$

denote the SENTENCE-BERT (SBERT, Reimers and Gurevych (2019)) embeddings of p and $\mathcal{C}$, respectively, and $|\cdot|$ denotes the number of words. We hypothesize that the larger csim, the better p aligns with $\mathcal{C}$.

nppl Normalized perplexity.

$$\mathrm{nppl}(p|\mathcal{C}) = \frac{\mathrm{ppl}(p|\mathcal{C})}{\mathcal{U}(p|\mathcal{C})},$$

where the perplexity score is calculated as

$$\begin{aligned}\mathrm{ppl}(p|\mathcal{C}) &= \exp\left(-\frac{\log P(p|\mathcal{C})}{|p| + |\mathcal{C}|}\right) \\ &= \exp\left(-\frac{\sum_{i=1}^{|p|} \log P(w_{p,i}|\mathcal{C}, w_{p,1:i-1})}{|p| + |\mathcal{C}|}\right),\end{aligned}$$

here $P(w_{p,i}|\mathcal{C}, w_{p,i:i-1})$ indicates the probability of the i-th premise word $w_{p,i}$ taken from the concatenation of $\mathcal{C}$ and p. The normalizing factor $\mathcal{U}(p|\mathcal{C})$ is based on the likelihood of an arbitrary text of length $|p| + |\mathcal{C}|$, in which each word is uniformly sampled from the vocabulary V of the argument aligner:

$$\mathcal{U}(p|\mathcal{C}) = \exp\left(-\frac{\sum_{i=1}^{|p|+|\mathcal{C}|} \log |V|^{-1}}{|p| + |\mathcal{C}|}\right) = |V|,$$

where $|V|$ is the size of V. We hypothesize that the smaller nppl, the better p aligns with $\mathcal{C}$.

npmi Normalized point-wise mutual information.

$$\begin{aligned}\mathrm{npmi}(p|\mathcal{C}) &= \frac{\mathrm{pmi}(p|\mathcal{C})}{h(p, \mathcal{C})} = -\frac{\log P(p) + \log P(p|\mathcal{C})}{\log P(\mathcal{C}) + \log P(p|\mathcal{C})} \\ &= -\frac{\log P(p) + \sum_{i=1}^{|p|} \log P(w_{p,i}|\mathcal{C}, w_{p,1:i-1})}{\log P(\mathcal{C}) + \sum_{i=1}^{|p|} \log P(w_{p,i}|\mathcal{C}, w_{p,1:i-1})},\end{aligned}$$

where $h(p, \mathcal{C})$ denotes the joint self-information (Futrell and Hahn, 2022). We hypothesize that the larger npmi, the better p aligns with $\mathcal{C}$.

rand An argument aligner that selects five random premise sentences from $\mathcal{P}$.

To calculate nppl and npmi scores with low computational cost, we use a simple pre-trained GPT-2 model ($|V| = 50,257$, Radford et al. (2019)) and compute the log likelihoods by taking the logits of the last decoder layer.

3 Methodology

In line with the concept presented by Gurcke et al. (2021), our objective is to investigate the extent to which the premises, when aligned with the conclusions using our argument aligners, can effectively contribute to the reconstruction of those conclusions.

3.1 Dataset

Although previous works have resulted in datasets for scientific argument mining (Lauscher et al., 2018; Mayer et al., 2020; Achakulvisut et al., 2019) and natural language inference (Sadat and Caragea, 2022; Khot et al., 2018), none deals with pairing premises and conclusions in scientific abstracts. Therefore, we created a dataset called ARG-ALIGN (detailed statistics in Table 1) by aggregating structured abstracts from papers in PubMed Central Open Access (PMCOA, National Library of Medicine (2003)) that are segmented into multiple discourse sections such as BACKGROUND, OBJECTIVES, METHODS, RESULTS, and CONCLUSIONS.

Count	Training	Validation	Test
# structured abstracts	13,939	1,745	1,752
# premise sentences	69,695	8,725	8,760
# conclusion sentences	28,668	3,627	3,605

Table 1: Overall statistics of our ARG-ALIGN dataset.

To ensure that our GPT-2-based argument aligners are naive with regards to our aggregated dataset, we intentionally selected structured abstracts from papers that were published in the year 2022, which was after the release of GPT-2. Following the instructions in Gao et al. (2023), we take the text under the CONCLUSIONS section as the conclusion segment $\mathcal{C}$ and all other sentences of the abstract as candidate premise sentences $\mathcal{P}$. We only use abstracts containing a maximum of three conclusion sentences to ensure they fit within the input constraints when reconstructing them from the premises.

3.2 Conclusion Generators

For conclusion generation, we fine-tuned two Seq2seq models: 1) T5-large with 770M parameters (Raffel et al., 2020); and 2) BART-large with 400M parameters (Lewis et al., 2020), as well as three large language models (LLMs): a) LLaMA-v1 with 7B parameters (Touvron et al., 2023); b) Galactica with 6.7B parameters (Taylor et al., 2022); and c) GPT-3.5-turbo with 170B parameters (OpenAI, 2023). All conclusion generators were fine-tuned on a single NVIDIA GeForce RTX 3090 GPU card, except GPT-3.5-turbo[2] which we fine-tuned via the OpenAI API. Specifically, we fine-tuned LLaMA and Galactica with a parameter-efficient (Liu et al., 2022) quantized low-rank adapter technique (Dettmers et al., 2023).

3.3 Evaluation

Following Gurcke et al. (2021) and Syed et al. (2021), we evaluate the individual argument aligners by measuring the sufficiency of the aligned premise sentences $\mathcal{P}_k^*$ for the corresponding conclusion segment $\mathcal{C}$, where the sufficiency is measured in terms of the average ROUGE F1 score (Lin, 2004) between the generated conclusion and the original conclusion $\mathcal{C}$.

4 Results and Discussion

We present conclusion generation results for different argument aligners in Table 2. In addition to the four argument aligners, we also report the sufficiency of taking all sentences as premises for generating the conclusion (denoted as **full**). Note that we did not use T5-large on this task due to its input length limitation of 512 tokens.

We found that all argument aligners selected premise sentences of encouraging sufficiency, evident from their average ROUGE-2 scores consistently exceeding 10. Interestingly, premises aligned using npmi consistently generated the best conclusion, suggesting that npmi captures well the dichotomy of premises and conclusions in scientific arguments.

Somewhat surprisingly, we found that full (unrestricted) premises tended to degrade the generated conclusions, as evidenced by lower ROUGE scores. Perhaps, full premises may contain irrelevant content in relation to the conclusions that overshadows

[2] Fine-tuning GPT-3.5-turbo with the OpenAI API `https://platform.openai.com/docs/api-reference` has costed 32.93 US dollars.

conclusion generators	csim			nppl			npmi			rand			full		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
T5-770M‡	32.43	12.90	24.43	32.04	12.66	24.32	**32.47**	**13.12**	**24.47**	30.10	11.17	22.76	-	-	-
BART-400M‡	33.91	13.47	24.74	33.68	13.52	24.73	**34.18**	**13.91**	**24.96**	31.32	11.73	22.89	34.16	13.40	24.44
LLaMA-v1-7B‡	33.75	13.99	25.35	33.90	13.99	**25.84**	**33.94**	**14.13**	25.75	31.71	12.39	23.97	33.73	13.66	25.34
Galactica-6.7B‡	34.62	14.54	26.39	34.37	14.41	26.18	34.87	**14.89**	**26.57**	32.93	13.16	25.00	**35.50**	14.62	26.42
GPT-3.5-turbo†	31.57	10.62	20.90	31.16	10.59	20.63	**31.99**	**11.17**	**21.39**	29.29	8.87	19.38	30.84	10.25	20.24
GPT-3.5-turbo‡	35.38	14.36	26.56	35.03	14.27	26.32	**35.60**	**14.89**	**26.85**	33.45	12.80	25.17	35.49	14.58	26.68

Table 2: Results on generating the conclusion from premises extracted by different argument aligners, measured as ROUGE F1 scores. † indicates zero-shot models without fine-tuning and ‡ indicates the fine-tuned models.

the relevant information for conclusion generation.

Finally, the fine-tuned BART-large conclusion generator outperformed the 425 times larger zero-shot GPT-3.5 generator. We suggest that because LLMs such as GPT-3.5 excel at generating text of low perplexity (Mitrović et al., 2023), it is likely that GPT-3.5 has a preference to use less common vocabulary and expressions when generating the conclusion, resulting in lower ROUGE scores. However, we noticed that after fine-tuning, GPT-3.5 has acquired the ability to incorporate words more typical of scientific language, leading to improved ROUGE scores.

5 Related Works

Computational argument sufficiency was first studied by Stab and Gurevych (2017). They viewed argument sufficiency as a binary classification task and trained a CNN classifier to predict whether an argument is sufficient or not. Later, the concept of argument sufficiency was extended to include argument strength, with strong arguments steering conversations towards more crucial topics compared to weak arguments. Hunter (2022) proposed assessing the strength of deductive arguments by probabilistically modeling the necessity and sufficiency of premises for claims with a defeasible logic. Their four-dimensional probabilistic measures of argument strength provided a theoretical foundation of computational argument evaluation.

Computational argument evaluation often involves utilizing language models for assessing premise-conclusion pairs. For example, conclusion generation focuses on the challenge of inferring conclusions from a provided collection of premises, approaching it as a text generation task (Alshomary et al., 2021; Tang et al., 2022; Syed et al., 2021). Shieh et al. (2019) investigated the effectiveness of Seq2seq models in generating conclusions from Random Clinical Trials (RCTs), indicating the capability of these models to perform scientific reasoning. Other works focused on generating sentence- and paragraph-level counter-arguments, with carefully designed control mechanisms (Hua et al. (2019); Schiller et al. (2021); Saha and Srihari (2023); Alshomary and Wachsmuth (2023)) such that the generated conclusions contain more detailed information.

6 Conclusions

In this study, we explored semantic similarity, text perplexity, and mutual information as unsupervised argument aligners. We quantified these metrics on the task of pairing premises with conclusions in PMCOA paper abstracts. Our primary objective was to probe the sufficiency of aligned premises by using them to reconstruct the conclusions.

Our findings indicate that semantic similarity, often considered a straightforward measure of text relevance, did not emerge as the best criterion for premise-conclusion alignment. This surprising result suggests that the process of scientific reasoning within abstracts is not solely driven by text-based similarity, but rather encompasses nuanced perspectives involving the cohesiveness of premise sentences amongst each other, as captured by $P(p)$.

This study highlights the need for a deeper understanding of the intricacies involved in the construction of well-aligned argument pairs in scientific papers. Our research sheds light on the multifaceted nature of scientific reasoning and the importance of exploring alternative approaches that better capture the underlying connections between premises and conclusions. As we move forward, it becomes evident that refining the techniques for aligning arguments will contribute to more accurate and insightful representations of scientific discourse, with the potential of improving the information dissemination and knowledge synthesis within the scientific community.

7 Limitations

The main limitations of our work are:

- When normalizing perplexity scores for the nppl aligner, we make the assumption that words are sampled uniformly from the vocabulary. However, this approach may not be the most effective way. We propose that employing a more refined sampling strategy that takes into account the lexical preferences for premises and conclusions in scientific abstracts could potentially result in improved performance.

- The calculation of npmi is point-wise, which does not consider the relation between individual premise sentences such as sentence order.

- Our method relies on structured scientific abstracts. When applying our approach to non-structured scientific abstracts, conclusions would have to be annotated in the first place.

- Figure 2 in Appexdix A shows that all the argument aligners tend to prefer premise sentences at the start of abstracts. We leave the investigation into this preference for future work.

In the future, we will investigate multi-step scientific reasoning by extending our findings to more complex argumentation schemes.

8 Acknowledgements

We acknowledge the support from Swiss National Science Foundation NCCR Evolving Language, Agreement No.51NF40_180888. We also thank the anonymous reviewers for the constructive comments and feedback.

References

Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*.

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967.

Arne Binder, Leonhard Hennig, and Bhuvanesh Verma. 2022. Full-text argumentation mining on scientific publications. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 54–66.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022a. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022b. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111.

Richard Futrell and Michael Hahn. 2022. Information theory as a bridge between language function and language form. *Frontiers in Communication*, 7:657725.

Yingqiang Gao, Nianlong Gu, Jessica Lam, and Richard HR Hahnloser. 2022. Do discourse indicators reflect the main arguments in scientific papers? In *Proceedings of the 9th Workshop on Argument Mining*, pages 34–50.

Yingqiang Gao, Jessica Lam, Nianlong Gu, and Richard Hahnloser. 2023. Greedycas: Unsupervised scientific abstract segmentation with normalized mutual information. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6093–6108.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77.

Mary B Hesse. 1974. *The structure of scientific inference*. Univ of California Press.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672.

Anthony Hunter. 2022. Argument strength in probabilistic argumentation based on defeasible rules. *International Journal of Approximate Reasoning*, 146:79–105.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. Idea.

Nataliia Kees, Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2021. Active learning for argument strength estimation. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 144–150.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46.

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

Mengze Li, Tianbao Wang, Jiahe Xu, Kairong Han, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Shiliang Pu, and Fei Wu. 2023. Multi-modal action chain abductive reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4617–4628.

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

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020-24th European Conference on Artificial Intelligence*.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2021. What makes my model perplexed? a linguistic investigation on neural language models perplexity. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 40–47.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

National Library of Medicine. 2003. PMC Open Access Subset. Internet.

Nikola I Nikolov and Richard Hahnloser. 2019. Large-scale hierarchical alignment for data-driven text rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 844–853.

OpenAI. 2023. ChatGPT.

Ekaterina Ovchinnikova, Niloofar Montazeri, Theodore Alexandrov, Jerry R Hobbs, Michael C McCord, and Rutu Mulkar-Mehta. 2014. Abductive reasoning with a large knowledge base for discourse processing. *Computing Meaning: Volume 4*, pages 107–127.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Sougata Saha and Rohini Srihari. 2023. Argu: A controllable factual argument generator. *arXiv preprint arXiv:2305.05334*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396.

Cassandra EL Seah, Zheyuan Zhang, Sijin Sun, Esther Wiskerke, Sarah Daniels, Talya Porat, and Rafael A Calvo. 2022. Designing mindfulness conversational agents for people with early-stage dementia and their caregivers: Thematic analysis of expert and user perspectives. *JMIR aging*, 5(4):e40360.

Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2019. Towards understanding of medical randomized controlled trials by conclusion generation. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 108–117.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493.

Liyan Tang, Shravan Kooragayalu, Yanshan Wang, Ying Ding, Greg Durrett, Justin F Rousseau, and Yifan Peng. 2022. Echogen: Generating conclusions from echocardiogram notes. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 359–368.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

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

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460.

Nathan Young, Qiming Bao, Joshua Bensemann, and Michael J Witbrock. 2022. Abductionrules: Training transformers to explain unexpected inputs. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 218–227.

Wenting Zhao, Justin T Chiu, Claire Cardie, and Alexander M Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. *arXiv preprint arXiv:2305.14618*.

A Comparison of Argument Aligners

Figure 2 illustrates the relative positioning of premise sentences aligned by various argument aligners. It is evident that csim, nppl, and npmi metrics display an inclination toward selecting premise sentences located at the start of structured abstracts. The content located in the beginning of structured abstracts typically is the motivation for the study and holds an importance for setting an expectation of the downstream conclusion.

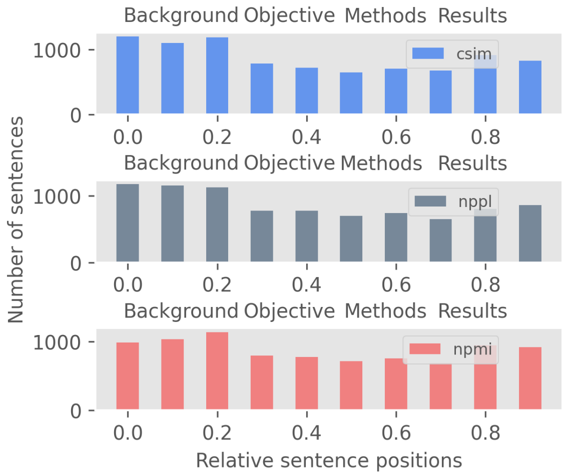

Figure 2: Relative positions within the abstracts (with 0 indicating the start and 1 the end of abstract) for premise sentences picked by different argument aligners.

In order to explore to which extent do the initial premises in the abstract play a role in generating conclusions, we also use the first five premise sentences to generate the conclusions. Since the first five sentences form a consecutive sequence, we did not introduce any additional separation tokens during the model's training process.

conclusion generators	first five premises		
	R-1	R-2	R-L
T5-770M[‡]	30.76	11.74	23.25
BART-400M[‡]	31.79	10.75	22.92
LlaMA-v1-7B[‡]	32.45	12.85	24.66
Galactica-6.7B[‡]	33.26	13.40	25.23
GPT-3.5-turbo[†]	30.13	9.33	19.93
GPT-3.5-turbo[‡]	34.66	13.54	25.95

Table 3: Results on generating the conclusion from the first five premise sentences in structured abstracts, measured as ROUGE F1 scores.

The findings presented in Table 3 demonstrate that in general, the first five premise sentences perform better than the random baseline. This suggests that, to some extent, pertinent information for drawing conclusions can be found in the initial portion of abstracts.

To assess whether ROUGE scores can accurately represent the degree of alignment between premises and conclusions, we randomly selected 100 structured abstracts from the test set. We then computed the correlation coefficients between the average ROUGE F1 scores between premises and conclusions (specifically R-1, R-2, and R-L) and the alignment scores (csim, nppl, and npmi) independently.

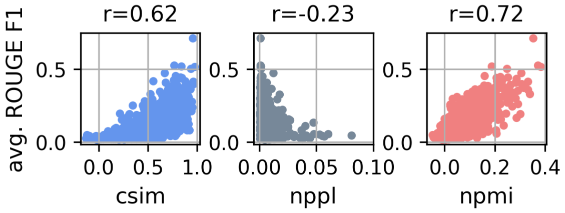

Figure 3: Average ROUGE F1 score between aligned premises and original conclusions, calculated for 100 randomly selected structured abstracts in the Test set. $p < 10^{-10}$ for csim, nppl, and npmi observed (Pearson correlation test).

Figure 3 indicates that csim and npmi align positively with the average ROUGE F1 scores, suggesting a tendency of these metrics to align lexically similar premises with conclusions. By contrast, nppl aligns negatively with ROUGE scores. Our findings highlight a degree of consistency within scientific reasoning, where a logically sound connection between a premise and conclusion is more likely (indicated by high npmi), exhibits greater semantic similarity (indicated by high csim), and is characterized by enhanced coherence (indicated by low nppl).

B Dataset Example

Table 4 shows an example from our ARG-ALIGN dataset. The information of discourse sections is removed for clarity.

Title: Designing Mindfulness Conversational Agents for People With Early-Stage Dementia and Their Caregivers: Thematic Analysis of Expert and User Perspectives (Seah et al., 2022)	
Premises	The number of people with dementia is expected to grow worldwide. **Among the ways to support both persons with early-stage dementia and their caregivers (dyads), researchers are studying mindfulness interventions**. *However, few studies have explored technology-enhanced mindfulness interventions for dyads and the needs of persons with dementia and their caregivers*. ***The main aim of this study was to elicit essential needs from people with dementia, their caregivers, dementia experts, and mindfulness experts to identify themes that can be used in the design of mindfulness conversational agents for dyads***. **Semistructured interviews were conducted with 5 dementia experts, 5 mindfulness experts, 5 people with early-stage dementia, and 5 dementia caregivers**. Interviews were transcribed and coded on NVivo (QSR International) before themes were identified through a bottom-up inductive approach. ***The results revealed that dyadic mindfulness is preferred and that implementation formats such as conversational agents have potential***. ***A total of 5 common themes were also identified from expert and user feedback, which should be used to design mindfulness conversational agents for persons with dementia and their caregivers***. *The 5 themes included enhancing accessibility, cultivating positivity, providing simplified tangible and thought-based activities, encouraging a mindful mindset shift, and enhancing relationships.*
Conclusion	In essence, this research concluded with 5 themes that mindfulness conversational agents could be designed based on to meet the needs of persons with dementia and their caregivers.

Table 4: An example in our proposed ARG-ALIGN dataset. We use **bold font**, underline, *italic font* to indicate premise sentences select by the csim, nppl, and npmi argument aligners respectively.

C Fine-tuning Details

Given that argument aligners may select premise sentences that are not contiguously located within the abstracts, we employed a special token <SENTENCEMISSING> to indicate missing premise sentences that were not selected by the argument aligners. This approach encourages the models to learn to generate conclusions from non-contiguous premises.

The training settings for different models are as follows:

Seq2seq Following the original training prompts used in Raffel et al. (2020), we first concatenated the aligned premises with <SENTENCEMISSING> and then augmented the concatenation with the suffix "summarize: " when fine-tuning T5-large. For BART-large, the aligned premises were simply concatenated with <SENTENCEMISSING> and used as input. Both T5-large and BART-large models were optimized with AdamW (Loshchilov and Hutter, 2018) with batch size of 2, learning rate initialized at $1e^{-5}$, and adapted with 10% warm-up steps by the linear scheduler, and fine-tuned for five epochs. We report the performance from the checkpoints with the best results on the validation set. The maximal output length during the inference is set to 128.

LLM We fine-tuned LLaMA-v1-7B and Galactica-6.7B using QLoRA (Dettmers et al., 2023) with batch size of 4 and int8 quantization (Dettmers et al., 2022b). For inferences, a temperature of 1.0 was utilized to ensure that the models do not exhibit a strong confidence for specific words during generation. We concatenated the aligned premises with the conclusions to form the following prompt:

> Premise: [aligned premises concatenated with <SENTENCEMISSING>] Conclusions: [concatenated conclusions]

Notice that for the LLMs, only the logits of the conclusion tokens are used to optimize the adapter's parameters. To accelerate the inference, we first converted the fine-tuned PEFT models to huggingface models, then we compiled them with CTranslate2[3] toolkit (Klein et al., 2020). Both LLaMA-v1-7B and Galactica-6.7B were trained for three epochs. We use bitsandbytes[4] toolkit (Dettmers et al., 2022a) for int8 matrix multiplication.

For the zero-shot GPT-3.5-turbo model, we used the following prompt:

> Your task: Please generate a conclusion text that can be drawn from the following sentences used as premises: [aligned premises concatenated with <SENTENCEMISSING>].
>
> Requirements:
>
> 1. Infer the conclusion text only from the given premises.
>
> 2. Please return only the generated conclusion text. The conclusion text should be minimally verbose and should not contain any irrelevant decorative text. For example, if the conclusion you inferred is "Pluto is not a planet.", do not respond with "The conclusion that can be drawn from the given premises is that Pluto is not a planet.". Text like "This conclusion can be drawn from the given premises" should not be part of the generated conclusion text.

For the fine-tuned GPT-3.5-turbo model, we used the same prompt as for the LLaMA-v1-7B and Galactica-6.7B model.

D Results of Oracle Aligner

To investigate the maximum potential performance in generating conclusions from aligned premises, we developed an oracle argument aligner that picks the five premise sentences associated with generated conclusions of highest ROUGE scores. We opt for T5-large and BART-large as the conclusion generators due to their fast inference speed. The oracle ROUGE scores and the percentage thereof achieved by the top non-oracle argument aligner (npmi) are presented in Table 5.

models	R-1 / npmi%	R-2 / npmi%	R-L / npmi%
T5-770M‡	45.77 / 70.94	24.07 / 54.51	36.76 / 66.57
BART-400M‡	46.80 / 73.03	24.14 / 57.62	36.54 / 68.31

Table 5: Oracle results using fine-tuned BART-large and T5-large as conclusion generators.

The npmi aligner achieves more than 70% of the theoretical maximum ROUGE-1, over 54% for ROUGE-2, and over 66% for ROUGE-L. This observation highlights npmi's capacity to effectively select sufficient premises.

[3] MIT license, available at https://github.com/OpenNMT/CTranslate2.

[4] MIT license, available at https://github.com/TimDettmers/bitsandbytes.