Towards Trustworthy Summarization of Cardiovascular Articles: A Factuality-and-Uncertainty-Aware Biomedical LLM Approach

Eleni Partalidou[1,2], Tatiana Passali[1], Chrysoula Zerva[3,4,5], Grigorios Tsoumakas[1,2], Sophia Ananiadou[2,6]

[1]Aristotle University of Thessaloniki, [2]Archimedes, Athena Research Center, Greece,
[3]Instituto Superior Técnico, Universidade de Lisboa, [4]Instituto de Telecomunicações, Portugal,
[5]ELLIS Unit Lisbon, [6]The University of Manchester

Correspondence: epartala@csd.auth.gr

Abstract

While large, biomedical documents with complex terminology are in need of being understood more easily and efficiently, summarizing this kind of content can be problematic, as Large Language Models (LLMs) aren't always trustworthy. Considering the importance of comprehending Cardiovascular Diseases, we study in depth the ability of different state-of-the-art biomedical LLMs to generate factual and certain summaries in this topic, and examine which generation choices can influence their trustworthiness. To that end, besides using factuality metrics, we employ techniques for token-level uncertainty estimation, an area that has received little attention from the scientific community. Our results reveal dissimilarities between LLMs and generation methods, and highlight connections between factuality and uncertainty metrics, thereby laying the groundwork for further investigation in the area.

1 Introduction

Biomedical researchers worldwide try to solve vital medical problems and publish scientific discoveries. Due to the exponentially increasing amount of scientific publications, summarizing them is vital, as they enable medical practitioners to keep up with the literature in an efficient manner. For that reason, it is crucial that the summary is accurate, as a minor mistake in explaining a medical concept or an unclear medical advice to treat a disease can have severe consequences for the health of patients. Large Language Models (LLMs) have recently been used to process and understand this kind of information in depth.

In recent years, LLMs have gained much attention from the scientific community, as they have been especially transformative for generative tasks, such as text summarization, machine translation, and question answering (Jurafsky and Martin, 2025). Text summarization, the task of our interest, is the process of creating a brief, accurate, and coherent summary of a longer text document. LLMs have greatly facilitated this task by providing the option to generate new text with the most salient information (i.e., abstractive summarization; Shakil et al. (2024)). In biomedicine, scientific findings tend to be reported in large documents with complex terminology, so summarizing scientific content can make important, clinical information accessible to researchers and clinicians more easily and efficiently (Xie et al., 2023a). However, LLMs may misrepresent their confidence and have specifically been shown to overestimate their knowledge and certainty level (they don't know what they don't know). As a result, they may confidently generate summaries with hallucinations or ambiguities (Baan et al., 2023; Hu et al., 2023) that can lead to misinformation with potentially severe consequences in medical contexts.

Although previous studies have explored uncertainty in biomedicine (Zerva et al., 2017; Kim et al., 2025), most either attempt to detect confidence expressions in text (Vasilakes et al., 2022), or focus on sequence-level uncertainty (Farquhar et al., 2024; Wagner et al., 2024; Qiu and Miikkulainen, 2024; Nikitin et al., 2024), frequently requiring the use of external models, repetitive sampling, or dedicated loss functions. Instead, we focus on simple, flexible, token-level uncertainty metrics, which can detect fine-grained local uncertainties, while also avoiding sequence-level limitations, such as length bias and over-correction that arise when collapsing token distributions into a single score. This area has received little attention in biomedical summarization, despite its importance in early detection and avoidance of misleading outputs. We thus attempt to address this gap and present early findings, assessing several biomedical LLMs on summarization of literature related to Cardiovascular Diseases.

We argue that a trustworthy model should not

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 200–207
November 9, 2025 ©2025 Association for Computational Linguistics

only achieve high factuality but also high confidence, especially for factually correct generations. We thus examine different factors that could influence factuality and uncertainty, such as decoding strategies and fine-tuning. We find that the proposed factuality and confidence metrics significantly correlate for some model variants, motivating further token-level analysis and more dedicated uncertainty metrics.

2 Related work

Biomedical summarization has become an important task and recent studies show an increased interest at it with the proposal of novel approaches based on LLMs (Xie et al., 2023a). Firstly, (Luo et al., 2022) introduced a new task of readability controllable summarization for biomedical documents, which aims to recognize users' readability demands and generate summaries that better suit their needs. Moreover, (Luo et al., 2023) proposed a novel citation-aware scientific paper summarization framework based on a citation graph, able to accurately locate and incorporate the salient contents from references, as well as capture varying relevance between source papers and their references. Lastly, (Xie et al., 2023b) addressed the issues of low-coherence summaries and the lack of explainability in black-box models by proposing a domain knowledge-enhanced graph topic transformer for explainable biomedical text summarization.

Evaluation of factuality in biomedical text generation is an ongoing challenge. (Zha et al., 2023) introduced AlignScore, a holistic metric, based on a general function of information alignment of text and its unified framework, which achieved substantial improvements over previous metrics. (Min et al., 2023) advocated a new evaluation metric that computes factual accuracy from pieces of generated text and was used to compare the performance of different LLMs. Additionally, (Bishop et al., 2023) proposed a new evaluation framework, LongDocFACTScore, for detecting human factuality targeting specifically summarized, long documents. Finally, (Luo et al., 2024) introduced a human-annotated dataset of LLM-generated summaries of clinical texts (TreatFact) and revealed significant performance gaps in terms of factuality for open-source LLMs.

Previous work has comprehensively examined **uncertainty in Natural Language Generation** (NLG) systems (Baan et al., 2023; Hu et al., 2023) and has explored strategies to address uncertainty with the goal of making LLMs more trustworthy, especially in biomedicine (Zerva et al., 2017; Kim et al., 2025). (Xu et al., 2020) studied summarization decoders in both blackbox and whitebox ways by focusing on the entropy of the models' predictions and revealed that features, such as the sentence position and the syntactic distance between adjacent pairs of tokens, influence uncertainty. (Ulmer et al., 2024) focused on token-level uncertainty and proposed a method for non-exchangeable conformal prediction, which was shown to improve text generation quality. Finally, (Fadeeva et al., 2024) introduced a token-level uncertainty method named Claim Conditioned Probability (CCP), disentangling claim-specific uncertainty from model decisions on surface forms, etc.

3 Methodology

We propose two different metrics of uncertainty, we test them to commonly used decoding methods, and we measure their correlation to factual accuracy.

3.1 Decoding strategies

We evaluate several decoding strategies for LLMs to identify the one that produces the least uncertain abstractive summaries. Specifically, we compare:

- Greedy search: At each timestep it selects the word with the highest probability.
- Top-k sampling: The k most likely words are filtered and the probability mass is redistributed among them (Fan et al., 2018).
- Top-p sampling: It chooses from the smallest possible set of words, whose cumulative probability exceeds a threshold p. The probability mass is then redistributed among them.

We note that the token-level uncertainty metrics (Section 3.3) can be applied across decoding methods, and, as they do not require sampling several times, they are also applicable to greedy decoding.

3.2 Factuality metrics

The factuality metrics process the summary (claim) at the sequence-level and require ground truth (evidence) for computation, which in our case is the abstract of the article.

HHEM. A series of models for detecting hallucinations in LLMs. These models collect a list of claims and associated evidence and compute a score between 0 and 1, where 0 means that the hypothesis is not evidenced at all and 1 means that the hypothesis is fully supported (Bao et al., 2024).

AlignScore. An automatic factual consistency metric, built on RoBERTa-large, applying a unified information alignment function between a claim and evidence. It splits each claim into sequences of specific length and each evidence into sentences, generates pairs, and computes an average score from the maximum alignment scores of the pairs. The score is between 0 (no factual accuracy) and 1 (full factual accuracy) (Zha et al., 2023).

3.3 Uncertainty metrics

Below we present the token-level uncertainty metrics we use. Even though they compute a value at each step, we average the values at sequence level.

Token Certainty. As a simple metric of model certainty at the token level, we use the maximum probability assigned to any token in the vocabulary at each decoding step. Thus, token certainty is defined as:

$$C = \max_i P(w_i), \qquad (1)$$

where $P(w_i)$ is the probability assigned to token w_i in the vocabulary.

Token Entropy. Beyond computing token certainty based on probabilities, we define a complementary metric based on the entropy of the token probabilities at each step, hence accounting for the full probability distribution over the vocabulary. It is computed as:

$$E = -\sum_{i=1}^{V} (P(w_i) \log(P(w_i)), \qquad (2)$$

where $P(w_i)$ is the probability assigned to token w_i in the vocabulary and V is the vocabulary size.

4 Experimental Setup

Below we describe the different features that are set up for the conduction of the experiments.

4.1 Biomedical LLMs

We use decoder-only LLMs that have been fine-tuned on biomedical content and give full access to the parameters for a more focused experimentation [1]. Specifically, we select the following variants:

BioMistral-7B (Labrak et al., 2024) is a suite of Mistral-based open source models pre-trained using textual data from PubMed Central Open Access. BioMistral is the first biomedical, multilingual LLM, demonstrating superior performance compared to existing open-source medical models. For the scope of our research, we use the default, 7B parameters version.

Meditron3-8B [2] is a LLaMA3.1, 8B model from a suite of open-source LLMs adapted to the medical domain named Meditron3. The models of this collection are co-designed by a global group of clinicians, humanitarian practitioners, and data scientists.

Phi4-14B is a decoder-only transformer of Microsoft built upon a blend of synthetic datasets, data from filtered public domain websites and acquired academic books, and Q&A datasets (Abdin et al., 2024). For compatibility with our work, we make use of the 14B parameters model from Meditron3, a model based on the Microsoft one.

Qwen2.5 models are another category of the Meditron3 collection fine-tuned from the organization of Qwen (Yang et al., 2024). Evaluation of used 7B and 14B parameters models showed that they are a better option for capturing real-world utility, especially in terms of contextual adaptation in under-represented settings.

4.2 Cardiology dataset

Cardiovascular diseases (CVDs) are the leading cause of death worldwide and a major contributor to reduced quality of life, with their prevalence driven by lifestyle and healthcare factors (Mensah et al., 2023; Mendis et al., 2011). Early detection and effective management are therefore essential to improving patient outcomes and reducing healthcare burdens. To support research in this area, we use biomedical literature from PubMed [3]. The dataset that we base our work on originates from (Cohan et al., 2018), which contains an amount of PubMed, long, and structured documents and we keep the same training, validation, and test splits. Additionally, the majority of the records contain one or more indexes named Medical Subject Headings (MeSH) [4]. The condition applied to filter the appropriate records is checking whether at least one of the MeSH terms falls into the category "Cardiovascular Diseases". Moreover, we ignore the

[1] For all models we use the version available on HuggingFace (Wolf et al., 2019).

[2] https://github.com/OpenMeditron

[3] https://pubmed.ncbi.nlm.nih.gov/

[4] https://www.ncbi.nlm.nih.gov/mesh/

records that have more than 8,192 tokens when processing, due to memory constraints. After these filterings, a total of 3,924 records for training, 230 records for validation, and 205 records for inference remain.

4.3 Input representation

The model input prompt is structured as follows:

$PROMPT\ article\ RESPONSE\ abstract$

for fine-tuning and:

$PROMPT\ article\ RESPONSE$

for inference, where $PROMPT$ is "Summarize the following biomedical article in a clear and concise manner, in no more than 300 words:" and $RESPONSE$ is "Summary:".

4.4 Hyperparameter settings

Our experiments are conducted on an Amazon, p5.48xlarge instance equipped with 192 vCPUs, 2,048 GiB RAM, and 8 NVIDIA H100 GPUs, each with 80 GiB of memory. Additionally, LoRA is applied to the models, and each biomedical model is fine-tuned with the cardiology dataset on 3 epochs with a batch size of 1, learning rate of 5^e-5, and the AdamW optimizer. Lastly, for the text generation strategies, we set K to 50 in the top-k sampling method and p to 0.70 in the top-p sampling method.

5 Results

In this section we present the comparisons across the metrics and models described above, accounting for different aspects, like the overall performance of the LLMs, the effect of fine-tuning on factuality and uncertainty, as well as differences between the decoding strategies. Finally, we assess the correlation between the factuality and uncertainty metrics.

5.1 Model Performance and Contribution of instruction fine-tuning

At first, we want to observe the level of contribution of instruction fine-tuning on the models. In Table 1 we present the experiments using greedy decoding. For the majority of the models, we do not observe significant improvements in terms of factuality and only small improvements in terms of certainty, because instruction fine-tuning pushes the LLMs to generate long outputs with knowledge they haven't seen before (Wu et al., 2025). However, as we want to keep the added information into all the models, we continue the experiments with the instruction fine-tuned ones.

We then compare the overall performance and trustworthiness of LLMs, focusing on the fine-tuned versions. Using the average rank shown in Table 1, it can be observed that the Qwen models are the best option across metrics, while Meditron-8B lags behind in both cases.

5.2 Investigation of decoding strategies

It is also important to understand whether different decoding strategies can impact the trustworthiness of a summary. For this comparison, we use the Qwen-7B and Qwen-14B models, since they outperform the rest with greedy decoding. From Table 2, it is evident that the sampling methods generate the most trustworthy summaries, i.e., outperform greedy decoding across metrics, with the token-entropy values decreasing greatly, producing both more accurate summaries, but also demonstrating higher model confidence during generation.

5.3 Correlations between the factuality and uncertainty metrics

As an initiative of finding relationships between the factuality and uncertainty metrics, we compute their correlation using Pearson's r. The sequence-level and token-level measures are paired with each other. The results in Table 3 show that the two types of metrics are correlated and that the most meaningful insights come from individual models, not decoding strategies alone.

6 Conclusions & Future Work

We evaluated the trustworthiness of state-of-the-art biomedical LLMs on summarization using both factuality and token-level uncertainty metrics. Results showed that model choice and decoding strategy influenced trustworthiness, even though we applied standard values on the sampling strategies for K and p, with Qwen variants performing best and sampling-based methods, especially top-p, producing more factual and confident summaries.

Several promising directions for future work include expanding the evaluation to larger and more diverse biomedical datasets to improve statistical reliability and test the generalizability of token-level uncertainty metrics across domains. Another direction is investigating different decoding hyperparameters to gain insights into how generation settings affect factuality and uncertainty. Finally, evaluating larger biomedical LLMs, including closed-source models, and incorporating human evalua-

Model	HHEM		AlignScore		Token Certainty		Token Entropy		Average Rank	
	ZS	FT	ZS	FT	ZS	FT	ZS	FT	Fact.	Uncert.
BioMistral-7B	0.1807	0.1915	**0.4493**	0.3364	**0.7736**	0.6746	**0.4007**	0.5239	3	2.5
Meditron-8B	0.1991	0.1994	0.3104	0.3112	0.6153	0.6239	2.1242	2.0916	4	5
Phi-14B	**0.2420**	0.2415	0.2319	0.2313	0.7487	0.7450	1.5654	1.5900	3	3.5
Qwen-7B	0.2263	0.2251	0.3263	0.3266	0.7445	0.7471	0.9976	0.9868	**2.5**	**2**
Qwen-14B	0.2367	0.2324	0.3158	0.3131	0.7678	0.7709	1.1612	1.1448	**2.5**	**2**

Table 1: Comparison of factuality and uncertainty across LLMs and effect of instruction fine-tuning. **Bold** values represent the best score for each metric and underlined ones the best score for each column. Results of zero-shot models are shown in the ZS columns and these of fine-tuned ones in the FT. Average ranks are shown separately for factuality (Fact.) and uncertainty (Uncert.).

	Method	HHEM	AlignScore	Certainty	Entropy
Qwen-7B	Greedy	0.2251	0.3266	0.7471	0.9868
	Top-k	0.2206	0.3161	0.7698	0.2361
	Top-p	**0.2354**	**0.3369**	**0.7988**	**0.1577**
Qwen-14B	Greedy	0.2324	0.3131	0.7709	1.1448
	Top-k	0.2375	**0.3496**	0.7343	0.3265
	Top-p	**0.2414**	0.3416	**0.7868**	**0.2033**

Table 2: Decoding strategy comparison for Qwen-7B/14B on factuality and uncertainty. **Bold** marks the best per metric within each model.

Correlation	Model	r	p
Certainty-HHEM	Meditron-8B	1.0	0.00
Entropy-AlignScore	Meditron-8B	-0.99	0.01
Certainty- Entropy	BioMistral-7B	-0.96	0.04
Entropy-AlignScore	Phi-14B	-0.96	0.04

Table 3: Observation of Pearson r correlation between the factuality and uncertainty metrics.

tion, along with automatic metrics, would further strengthen the reliability of the results.

Limitations

Although our work gives a great initiative for factuality and token-level uncertainty quantification in biomedical applications, there are still some areas that could be explored. At first, token-level uncertainty metrics can be used to other specializations of medicine in the future, other than cardiology, in order to proof the generalization of our work. Additionally, more descriptive, token-level metrics can be incorporated into the experiments for further uncertainty detection and quantification. Moreover, as we use Pearson coefficients, which do not account for confounding factors, like model architecture, decoding strategy, or dataset characteristics, some correlations may reflect model-level biases, rather than true causal relationships. The lack of comparison of automatic metrics to human evaluation is another limitation, which could strength the paper claims, if conducted. Lastly, due to the fact that factuality is a very important topic, future enhancements could investigate deeply factual accuracy and relativity to uncertainty.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Also, AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility. In addition, this research was partially funded by an unrestricted gift from Google, by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), by Fundação para a Ciência e Tecnologia (FCT) through 'OptiGov' (DOI 10.54499/2024.07385.IACDC), funded by the PRR under the measure RE-C05-i08.m04, by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), and also by FCT/MECI through national funds and, when applicable, cofunded EU initiatives under contract UID/50008 for Instituto de Telecomunicações. Lastly, Sophia Ananiadou acknowledges support from the British Heart Foundation Centre of Research Excellence, University of Manchester (Award Code: RC000797).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero

Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Jennifer A Bishop, Qianqian Xie, and Sophia Ananiadou. 2023. Longdocfactscore: Evaluating the factuality of long document abstractive summarisation. *arXiv preprint arXiv:2309.12455*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, and 1 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. *arXiv preprint arXiv:2210.04705*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Citationsum: Citation-aware graph contrastive learning for scientific paper summarization. In *Proceedings of the ACM web conference 2023*, pages 1843–1852.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. Factual consistency evaluation of summarization in the era of large language models. *Expert Systems with Applications*, 254:124456.

Shanthi Mendis, Pekka Puska, B editors Norrving, World Health Organization, and 1 others. 2011. *Global atlas on cardiovascular disease prevention and control*. World Health Organization.

George A Mensah, Valentin Fuster, Christopher JL Murray, Gregory A Roth, Global Burden of Cardiovascular Diseases, and Risks Collaborators. 2023. Global burden of cardiovascular diseases and risks, 1990-2022. *Journal of the American College of Cardiology*, 82(25):2350–2473.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533.

Hassan Shakil, Ahmad Farooq, and Jugal Kalita. 2024. Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, 603:128255.

Dennis Ulmer, Chrysoula Zerva, and André FT Martins. 2024. Non-exchangeable conformal language generation with nearest neighbors. *arXiv preprint arXiv:2402.00707*.

Jake Vasilakes, Chrysoula Zerva, Makoto Miwa, and Sophia Ananiadou. 2022. Learning disentangled representations of negation and uncertainty. *arXiv preprint arXiv:2204.00511*.

Nico Wagner, Michael Desmond, Rahul Nair, Zahra Ashktorab, Elizabeth M Daly, Qian Pan, Martín Santillán Cooper, James M Johnson, and Werner Geyer. 2024. Black-box uncertainty quantification method for llm-as-a-judge. *arXiv preprint arXiv:2410.11594*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tianyi Wu, Jingwei Ni, Bryan Hooi, Jiaheng Zhang, Elliott Ash, See-Kiong Ng, Mrinmaya Sachan, and Markus Leippold. 2025. Balancing truthfulness and informativeness with uncertainty-aware instruction fine-tuning. *arXiv preprint arXiv:2502.11962*.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023a. A survey for biomedical text summarization: From pre-trained to large language models. *arXiv preprint arXiv:2304.08763*.

Qianqian Xie, Prayag Tiwari, and Sophia Ananiadou. 2023b. Knowledge-enhanced graph topic transformer for explainable biomedical text summarization. *IEEE journal of biomedical and health informatics*, 28(4):1836–1847.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. *arXiv preprint arXiv:2010.07882*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Chrysoula Zerva, Riza Batista-Navarro, Philip Day, and Sophia Ananiadou. 2017. Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics*, 33(23):3784–3792.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

A Examples of summaries with low and high uncertainty

Below we present the least and most uncertain summaries of different cases using the fine-tuned, Qwen-7B model.

Greedy search (Least uncertain)

PMCID: PMC2913399

Summary: A network is generated by determining the first- and second - order interactions of cd proteins associated with a given morphological subgroup in a human protein interaction network consisting of refined experimental proteomics data . This network is described in high detail in lage et al (2007 , 2008) , and online (http://www.cbs.dtu.dk/suppl/dgf/) . Interactions of the cd proteins are integrated into a network by always including direct interactions between cd proteins , and only including indirect interactions mediated through proteins with q percent of its interactions to the cd set . Various thresholds for q are iteratively tested and value of q for the final network is chosen based on which value gives the optimal network significance , this procedure is described in detail in bergholdt et al (2007) and d'hertog et al (2007) . The method for determining network significances can be seen below . Detailed views of the networks can be seen in supplementary figures s1 , s2 , s3 and s4 . The significance of each of the generated 19 networks was determined by randomization testing as described in detail earlier (bergholdt et al , 2007 ; d'hertog et al , 2007) . Specifically , for an input set of ninput proteins yielding an interaction network (connected component) with g input proteins and t total proteins , a network score (ns

Greedy search (Most uncertain)

PMCID: PMC4635763

Summary: A 67-year-old woman presented to the emergency department with a sudden collapse and was found to have a cardiac myxoma, a rare tumor of the heart. The patient's initial presentation was characterized by ST-elevation myocardial infarction, but her coronary arteries appeared normal on immediate angiography. She also experienced a stroke, which is relatively common in cardiac myxoma patients. The patient underwent successful surgical resection of the myxoma, and her condition improved. This case highlights the importance of basic echocardiography in diagnosing cardiac myxomas and emphasizes the need for timely surgical intervention to prevent complications.

Top-k sampling (Least uncertain)

PMCID: PMC3297038

Summary: There are similarities between children and adults suffering from heart failure, such as the preferred pharmacological treatment, the use of pacemakers and heart transplants, the inability of the patient to reach the predicted heart rate for the patient's age during cardiopulmonary exercise testing, and the ergoespirometric response under similar clinical conditions. In adults, endothelial dysfunction is related to the development of diastolic dysfunction, Chagas disease, left ventricular hypertrophy, ischemic cardiomyopathy, heart failure, obesity, type 1 diabetes, hyperlipidemia, arterial hypertension, peripheral arterial disease, chronic kidney disease, and atherosclerosis because the dysfunction predisposes the vasculature to vasoconstriction, leukocyte adherence, platelet activation, and vascular inflammation. Nevertheless, there is a lack of data regarding endothelial function in children with cardiomyopathy. The severity of endothelial dysfunction is related to the cardiovascular risk, the severity of cardiovascular symptoms, and the inability to exercise and represents a predictor for cardiac transplant and death. It is known that diseases, such as Kawasaki's disease, hyperlipidemia, obesity, and type 1 diabetes, play important roles in systemic inflammation and endothelial dysfunction. These diseases may increase the likelihood of cardiovascular events and may predispose children to the development of cardiomyopathy. Based on these considerations, we reviewed the published literature on endothelial function in pre-pubertal children to evaluate the endothelial function in prepub

Top-k sampling (Most uncertain)

PMCID: PMC5290108

Summary: Intra-amniotic inflammation may increase the risk of adverse perinatal outcomes in preterm infants, but this risk is mainly due to low gestational age at birth rather than direct effects of inflammation. Culture-proven intra-amniotic infection does not appear to be a risk factor for adverse perinatal outcomes.

Top-p sampling (Least uncertain)

PMCID: PMC4666010

Summary: this is a retrospective cross - sectional study . there were 4,377 participants , all of whom underwent cardiac ct in a health promotion center at gangnam severance hospital (seoul , korea) between january 2008 and february 2013 . from 4,377 participants , we excluded men (n = 2,584) and premenopausal women (n = 748) . postmenopausal women 50 years or older with no menstrual periods for more than 12 consecutive months and women with elevated follicle - stimulating hormone levels (> 30 iu / l) were included in this study (n = 1,045) . women with diabetes mellitus (n = 59) , excessive alcohol consumption (n = 7) , viral hepatitis (positive results for hepatitis b surface antigen or anti hepatitis c virus ; n = 19) , liver cirrhosis or malignancy on ultrasonography (n = 6) , and self - reported or medically verified history of cvd (n = 35) were excluded from this study . height and weight were measured , and body mass index (bmi) was calculated by dividing weight (kg) by the square of height (m) . lifestyle , personal medical history of acute and

Top-p sampling (Most uncertain)

PMCID: PMC5192323

Summary: The present study demonstrates that plasma CTRP9 levels are independently and positively associated with carotid intima-media thickness in patients with type 2 diabetes without chronic kidney disease. This study further proposes that plasma CTRP9 level is a potential biomarker of atherosclerosis in type 2 diabetes patients without renal complications.