The Impact of Large Language Models in Academia: from Writing to Speaking

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly impacting human society, particularly in textual information. Based on more than 30,000 papers and 1,000 presentations from machine learning conferences, we examined and compared the words used in writing and speaking, representing the first large-scale study of how LLMs influence the two main modes of verbal communication and expression within the same group of people. Our empirical results show that LLM-style words such as *"significant"* have been used more frequently in abstracts and oral presentations. The implicit impact on human expression like writing and speaking is beginning to emerge and is likely to grow in the future. We take the first step in building an automated monitoring platform to record its longitudinal changes to call attention to the implicit influence and ripple effect of LLMs on human society.

1 Introduction

The development and popularity of large language models (LLMs) (OpenAI, 2024; Anthropic, 2024; OpenAI, 2023) have alerted more researchers to the impact of LLMs on human society. In this paper, we focus on the impact of LLMs in academia, especially on writing and speaking.

While the rapid increase in usage and impact of LLMs have been demonstrated in academic papers (Liang et al., 2024b; Geng and Trotta, 2024), few studies have addressed the influence of LLMs beyond writing. Only recently, a preprint pointed out the impact of LLMs on the words used in speaking, as collected in YouTube videos (Yakura et al., 2024). The similarities and differences in how writing and speaking are influenced, particularly for the same population, have not been explored.

People can use LLMs to write emails or accomplish tasks other than paper writing, which possibly changes their English expression and is reflected in their academic output later. Just like the use of Google Translate can affect the English expression of non-native English speakers (Resende and Way, 2021), a similar influence might be at play with LLM users and eventually influence even the way people speak.

Besides, detecting a mixture of machine-generated and human-written text is another difficulty being actively researched (Lee et al., 2022; Gao et al., 2024). Researchers have paid more attention to whether a piece of text is generated by LLMs, while the implicit impact of LLMs is often underestimated. Here we refer to people who do not directly use LLMs to create content but are influenced through exposure to such content.

In the face of these challenges and gaps, our contributions are three-fold:

1. We are the first to analyze and compare the impact of LLM on the writing and speaking in the same group of people.
2. We propose a simplified simulation-based method for estimating LLM impact.
3. We are calling attention to the implicit impact of LLMs, as the words used in the machine learning conference presentations show signs of being influenced by LLMs.

2 Data and Methodology

2.1 Datasets

To better explore and compare how the *"same"* group of people are affected in writing and speaking by LLMs, we crawled presentations and meta-information of papers from three machine learning conferences. The abstracts of papers rather than the full papers were used in the analysis, as the former are more representative. More than 30,000 papers and 1,000 talks were collected, detailed in Appendix B.

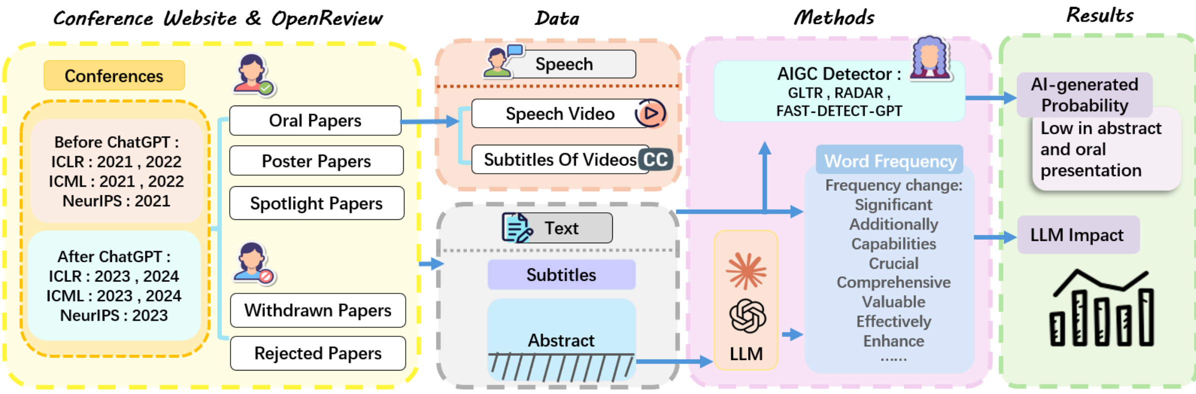


Figure 1: Overview of processing and analysis.

Then we use some Machine-Generated Text (MGT) detectors or analyze the changes in word frequency. The whole process is shown in Figure 1.

2.2 Word Frequency Analysis

Given that word frequencies are always changing, the issue of noise cannot be ignored. To reduce the error caused by the randomness of word usage, the target words are considered as a group, denoted as $W_I = \{w_i \mid i \in I\}$, where i is the frequency ranking in the corresponding dataset.

For a group of words W_{I_0}, the control group with shift n is defined as $W_{I_n} = \{w_{i+n} \mid i \in I_0\}$. Given a corpus S, the corresponding frequency $F_n(S)$ is $F_n(S) = \sum_{w \in I_n} f_w(S)$, where f_w means the frequency of word w in set S. The frequency ratio between two different corpus is $R_n(S, S') = \frac{F_n(S)}{F_n(S')}$.

Constructing control groups to analyze changes in word frequency has been used before (Matsui, 2024; Yakura et al., 2024). In this paper, the words in each group have roughly the same frequency based on the ranking in the dataset, which shows whether the change in frequency of the target words is unusual.

2.3 LLM Simulations and Impact Estimation

Some researchers have estimated the impact of LLMs by excess vocabulary only (Kobak et al., 2024), but the words in the abstract are also related to the topic of papers, and the hot topics of machine learning conferences change frequently. Therefore, it is also helpful to perform LLM simulations, and compare texts before and after processing for a reliable estimation of LLM impact.

If the frequency of word i is $f(S_1)$ and $f(S_2)$ in a corpus before and after LLM processing, the frequency change rate r_i is estimated as $\hat{r}_i = \frac{f(S_2) - f(S_1)}{f(S_1)}$. Then for the *"proportion"* (impact) of LLMs texts $\eta(S)$, the following equation is a simplified and direct version of the method proposed by Geng and Trotta (2024),

$$f_i^d(S) - f_i^*(S) = \eta(S) f_i^*(S) \hat{r}_i + \delta_i(S) \quad (1)$$

where $f_i^d(S)$ represents the frequency of word i in the set of texts S, $f_i^*(S)$ represents the one if LLMs do not affect writing abstracts, and $\delta_i(S)$ is a noise term.

The estimate of LLM impact given by Ordinary Least Squares (OLS) is expressed as

$$\hat{\eta}(S) = \frac{\sum_{i \in I} (f_i^d(S) - f_i^*(S)) f_i^*(S) \hat{r}_i}{\sum_{i \in I} (f_i^*(S) \hat{r}_i)^2} \quad (2)$$

where I is the set of words used for estimation, and different I give us different estimates.

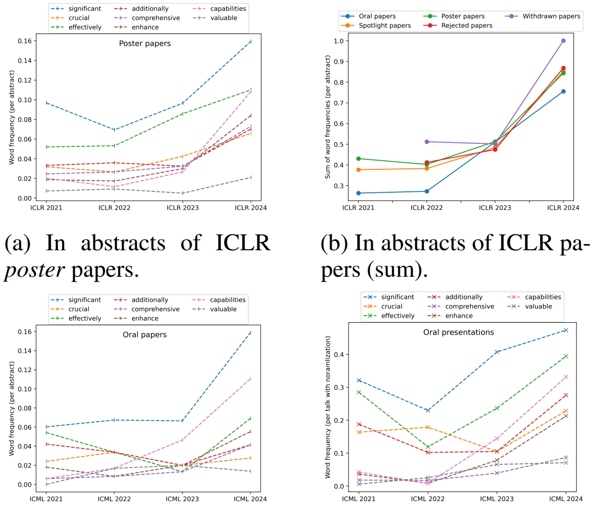
3 Results

3.1 AIGC Detectors

Several AIGC detectors[1] like Fast-DetectGPT (Bao et al., 2023), GLTR (Gehrmann et al., 2019) and RADAR (Hu et al., 2023) are used to detect the probability that the abstracts and speeches are generated by AI. The results show no significant difference between the pre-ChatGPT and post-ChatGPT era, more details and results are in Appendix C.1.

Therefore, we dive deeper into fine-grained word frequency analysis, as researchers have discovered that the frequency of some words increased rapidly across academic papers in different disciplines after the end of 2022 (Liang et al., 2024a).

[1]We do not use GPTZero because the size of its context window is smaller than the length of oral presentation.



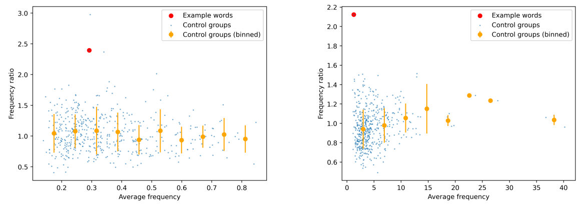
(a) In abstracts of ICLR *poster* papers.

(b) In abstracts of ICLR papers (sum).

(c) In abstracts of ICML *oral* papers.

(d) In presentations of ICML *oral* papers.

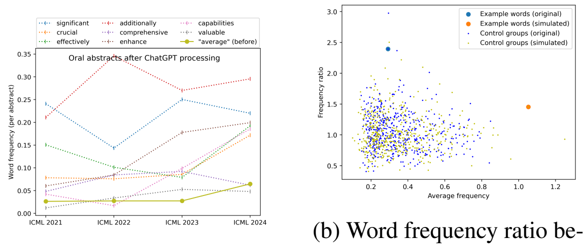
Figure 2: Word frequency in abstracts and presentations.



(a) Frequency ratio in abstracts of *oral* papers.

(b) Frequency ratio in presentations of *oral* papers.

Figure 3: Word frequency ratio in abstracts and talks of ICML *oral* papers. The error bars represent one standard deviation in each bin.



(a) Word frequency after ChatGPT processing.

(b) Word frequency ratio between simulated abstracts and original abstracts.

Figure 4: Word frequency in abstracts of ICML *oral*.

3.2 Changes in Word Frequencies

We found a similar trend for some words in the abstracts of conference papers. For example, the 8 words (**significant**, **crucial**, **effectively**, **additionally**, **comprehensive**, **enhance**, **capabilities**, **valuable**), listed as examples by Geng and Trotta (2024), are much more often observed in the abstracts of ICLR 2024 accepted papers than before, as shown in Figure 2a. Besides, Figure 2b indicates that the frequency sums of the 8 words are *"significantly"* higher in the ICLR 2024 abstracts than in abstracts of 2021 and 2022, in all three categories of accepted papers, as well as in the *reject* and *withdrawn* papers. The results also suggest the correlation between the use of these words and their destination in ICLR, with a lower frequency in *oral* and a higher frequency in *withdrawn* papers.

Figure 8 in the appendix provides more examples of this trend, including cases with the full text of the articles rather than just the abstracts. **In fact, our estimates of LLM impact are based on 40 different groups of words, each group consisting of hundreds of words. The 8 words here are only used as illustrative examples.**

Although these conference papers are mostly submitted to the arXiv as well, their average quality is higher than the same type of arXiv papers. These eight words shown above were taken together for better comparison and were not selected based on the data in this paper, reconfirming the LLM impact in academic research.

We further compared word frequency used in the abstracts and speeches of the *oral* papers. Since the time for presentations varied from year to year, the number of words was normalized based on the total number of words used in 2021. The results for ICML *oral* papers and presentations are presented in Figure 2c and Figure 2d, which shows that these example words are more frequently used in the abstracts of papers after 2022. Words in the speeches share the same trend as well, though not as strongly as in the abstracts.

3.3 Distribution of Frequency Ratios

The word frequencies in the abstracts of poster papers in NeurIPS from 2021 to 2023 were used for ranking i the words to form the control group W_I with the shift as defined in section 2.2. The frequency ratio $R_n(S, S')$ for the abstracts and talks of the ICML oral papers in 2024 compared to those from 2021 to 2022 are shown in Figure 3, with shift n from -250 to 250 forming the control groups.

The frequency ratios of the example word group are 3.4 standard deviations and 5.8 standard deviations away from the mean within the bin in abstracts and speeches, respectively. The vast majority of word groups do not have as much frequency change as the example word group.

3.4 LLM Simulations

The simulations of LLM were performed on GPT-3.5 with a simple prompt: *"Revise the following sentences"*.

The word frequency analysis on abstracts revised

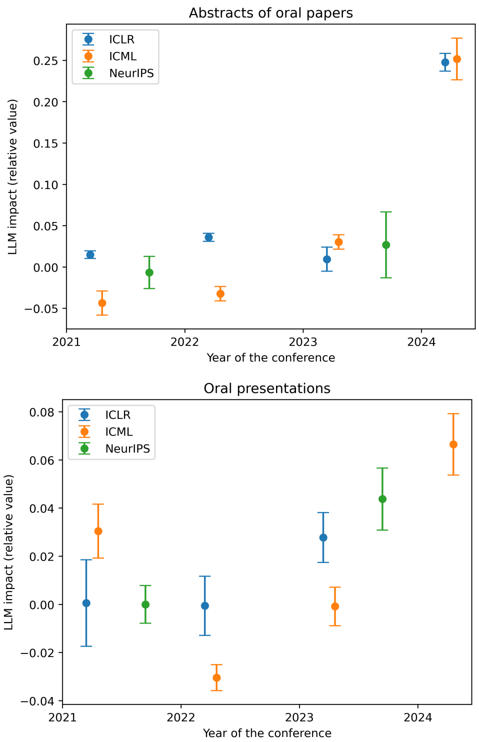


Figure 5: Estimation of LLM's impact on oral papers. Error bars indicate the standard deviation of the estimates using different groups of words.

by ChatGPT is in Figure 4a, obtained after the same calculations as for Figures 2a and 2c. The frequency of these words has all increased after ChatGPT processing, which reconfirms that ChatGPT favors these words. The comparison of simulated and original data in Figure 4b tells us that words sensitive to ChatGPT are in the minority.

Based on our results, it is hard to believe that the change in word frequency revealed in Figure 2 is a coincidence. Further, the results above have illustrated the correlation between word frequency changes and word preferences of LLMs.

3.5 LLM Impact Estimations

The ChatGPT-modified abstracts of spotlight (poster) papers from ICML 2022 were utilized to calculate $\hat{r}_i$ and LLM impact estimations. And $f_i^*(S)$ is approximated by the word frequency of the abstracts of poster or spotlight papers in 2021 for each conference, after normalization to have the same total number of words. To reduce the effect of topic-specific words, only the words that ranked in the top 10,000 of the Google Ngram dataset[2] were considered in the calculations, which represent the most frequent words in Google Books. Finally, hundreds of words are used for estimating, not just the eight previously exemplified.

Because of the lack of presentation data before 2021, the $f_i^*(S)$ of speeches were also approximated with the contents of the abstracts. Estimates of LLM impact based on different word selection criteria determined by $f_i^*(S)$ and $\hat{r}_i$ are shown in in Figure 5. Since the words written and spoken are different, the means of the estimates for 2021 and 2022 were also calibrated to "0", with corresponding adjustments for 2023 and 2024. Not surprisingly, LLM impact increased in the abstracts for the 2024 conference. In the presentations, the estimate of LLM impact has also increased, but not as "significantly" as in the abstracts.

Note that the LLM impact term here is a relative value, which is an estimate based on the output of a specific prompt via ChatGPT. Different prompts and different LLMs will have different outputs, and the implicit impact also plays a role.

4 Lifelong Platform for LLM Influence Monitoring

To investigate LLMs' potential indirect influences on human writing and speaking patterns, we developed an automated monitoring framework as shown in Figure 7. Our system collects research paper and oral presentations, then employs LLM simulations and algorithmic analysis to measure these influences in real-time, examining word frequency, sentence structure, and readability metrics. Through longitudinal analysis, we aim to uncover subtle influences that emerge over time. Data and software will be open-sourced under `CC 4.0`.

5 Discussions and Conclusions

Speech is one of the scenarios in which LLMs have an implicit impact (broadly), for it is usually safe to assume that the speaker is not using an LLM while presenting. In addition, because of the difference between written and spoken language, it is actually difficult to directly compare LLM impact on writing and speaking.

The rapid increase in AI-generated content requires us to pay more attention, as synthetic data can lead to model collapse (Shumailov et al., 2024; Briesch et al., 2023) and even knowledge collapse (Peterson, 2024). In the near future, a paper considered likely to be the product of LLMs may be so only because the authors have read too many papers containing LLM-style sentences. The implicit impact of LLMs seems unstoppable.

[2] https://www.kaggle.com/datasets/wheelercode/english-word-frequency-list

Limitations

LLMs have started a paradigm revolution in AI and transformed the game completely. While the discussion of the social impact of LLMs began long before the storm hit (Solaiman et al., 2019), it took some time to really *"delve into"* it.

There are many sources of implicit impacts, and it is also difficult to have a standard definition. For instance, the titles of papers (Matsui, 2024; Kobak et al., 2024; Astarita et al., 2024) analyzing the use of LLMs in academic papers begin with *"delving into"* (one of ChatGPT's signature words). This fits our definition of implicit impact in the broad sense, but some argue that it's just a means for authors to get readers' attention.

It is also true that researchers may have prepared presentations or slides using LLMs, in which case their choice of wording may have been influenced. The sample sizes in speaking are not as large as in writing, but they are homogenized and representative.

We concentrated on the frequency of words and did not address other forms. The use of words reflects the most basic information, and some changes should have occurred in the way they are expressed.

Ethics Statement

Our paper primarily focuses on LLM's influence on writing and speaking, and we must first declare that our research adheres to all applicable ethical standards. This study is intended to promote academic discussion and technological progress and all experiments are conducted in a strictly controlled environment. Our research encourages relevant developers to enhance the supervision of LLMs, thereby making them more trustworthy. We ensured that all datasets and benchmarks used in the study comply with their intended purposes and standards.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Anthropic. 2024. Claude 3.5: A sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2024-09-04.

Simone Astarita, Sandor Kruk, Jan Reerink, and Pablo Gómez. 2024. Delving into the utilisation of chatgpt in scientific publications in astronomy. *arXiv preprint arXiv:2406.17324*.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Honglin Bao, Mengyi Sun, and Misha Teplitskiy. 2024. Where there's a will there's a way: Chatgpt is used more for science in countries where it is prohibited. *arXiv preprint arXiv:2406.11583*.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*.

Dongping Chen, Jiawen Shi, Yao Wan, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2024. Self-cognition in large language models: An exploratory study. *arXiv preprint arXiv:2407.01505*.

Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Chujie Gao, Dongping Chen, Qihui Zhang, Yue Huang, Yao Wan, and Lichao Sun. 2024. Llm-as-a-coauthor: The challenges of detecting llm-human mixcase. *arXiv preprint arXiv:2401.05952*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Mingmeng Geng, Sihong He, and Roberto Trotta. 2024. Are large language models chameleons? *arXiv preprint arXiv:2405.19323*.

Mingmeng Geng and Roberto Trotta. 2024. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*.

Andrew Gray. 2024. Chatgpt" contamination": estimating the prevalence of llms in the scholarly literature. *arXiv preprint arXiv:2403.16887*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023a. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023b. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *ArXiv*, abs/2307.03838.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting machine-generated texts: Not just" ai vs humans" and explainability is complicated. *arXiv preprint arXiv:2406.18259*.

Dmitry Kobak, Rita González Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *arXiv preprint arXiv:2307.11729*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert West. 2024. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. 2024b. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Kentaro Matsui. 2024. Delving into pubmed records: Some terms in medical writing have drastically changed after the arrival of chatgpt. *medRxiv*, pages 2024–05.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

OpenAI. 2023. Introducing openai o1 preview. `https://openai.com/index/introducing-openai-o1-preview/`. Accessed: September 12, 2024.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-06-06.

Andrew J Peterson. 2024. Ai and the problem of knowledge collapse. *arXiv preprint arXiv:2404.03502*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *ArXiv*, abs/2212.04356.

Natália Resende and Andy Way. 2021. Can google translate rewire your l2 english processing? *Digital*, 1(1):66–85.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Zechang Sun, Yuan-Sen Ting, Yaobo Liang, Nan Duan, Song Huang, and Zheng Cai. 2024. Knowledge graph in astronomical research with large language models: Quantifying driving forces in interdisciplinary scientific discovery. *arXiv preprint arXiv:2406.01391*.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8384–8395.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*.

Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gašević, and Guanliang Chen. 2024. Towards detecting ai-generated text within human-ai collaborative hybrid texts. *arXiv preprint arXiv:2403.03506*.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436.

A Related Works

A.1 Impacts of LLMs

The storm of LLMs also crossed the boundaries of language, as researchers began to explore the wide variety of competencies and applications of LLMs in various disciplines: math (Ahn et al., 2024), physics (Sun et al., 2024), chemistry (Guo et al., 2023b), social science (Geng et al., 2024), psychology (Demszky et al., 2023), cognitive science (Chen et al., 2024), and almost all corners of science. In the process of interacting with LLMs, people's opinions may also change (Costello et al., 2024).

A.2 Measuring LLM Impact Through Word Frequency

We know that many scientists are using LLMs, but there are not many papers with quantitative estimates of the impact of LLMs in the scientific community.

As we mentioned before, there are also several papers that estimate the influence of LLMs based on word frequency, such as in academic writing (Liang et al., 2024b; Geng and Trotta, 2024; Kobak et al., 2024) and in peer review (Liang et al., 2024a). In addition, there are also some papers that simply observe changes (Gray, 2024; Matsui, 2024; Astarita et al., 2024) and some with more detailed analyses (Bao et al., 2024; Latona et al., 2024).

A.3 Machine Generated Text Detection

Detection of machine-generated text has also begun much earlier with different approaches (Gehrmann et al., 2019; Bakhtin et al., 2019; Uchendu et al., 2020). More have approaches have been proposed later, including metric-based methods (Mitchell et al., 2023; Yang et al., 2023; Bao et al., 2023), model-based methods (Guo et al., 2023a; Verma et al., 2023), and benchmarks (He et al., 2023). In addition, some papers also illustrate the ways to avoid LLM detection (Krishna et al., 2023; Koike et al., 2023), which makes this game harder.

Detecting a mixture of machine-generated text and human-written text is another challenge. Recent research on *"LLM-as-a-Coauthor"* (Zhang et al., 2024) explores the complex interactions between humans and LLMs in collaborative writing, investigating real-world applications of human-AI mixed text, moving beyond simple binary classification. In the face of mixed text (Zeng et al., 2024; Ji et al., 2024), more methods have also been proposed in the literature for detection and estimation, e.g., through words (Liang et al., 2024a) and styles (Gao et al., 2024).

But with implicit impact and ripple effects from LLMs, the line between human-written and machine-generated text is indistinct, and monitoring LLM-style text may be more meaningful than detecting who the author is.

B Data Collection and Experiments Setups

B.1 Data Collection

The abstracts for ICML 2021 and 2022 were collected from the official website, and all other data were scraped from the *OpenReview* platform. Since ICLR 2023 didn't distinguish *oral* and *spotlight*, we consider the *notable top 5%* papers and *notable top 25%* papers as them, respectively. As for NeurIPS 2022, accepted papers were not further classified. In addition, there are more than 10,000 rejected and withdrawn papers in ICLR from 2022 to 2024.

The videos of the oral presentation were collected from the official websites of ICLR, ICML, and NeurIPS. As NeurIPS 2022 did not feature oral papers, the videos of this year's NeurIPS were skipped. The videos of the ICLR 2024 oral presentation haven't been available online. To extract speech from the videos, we used the subtitle feature provided by each website. Using the Internet Download Manager (IDM) extension in Google Chrome, we downloaded all subtitle files in vtt format, which were converted into text via Python later.

B.2 Human Annotation Details

In this section, we provide details for human annotation. Dataset of oral presentation and abstract are annotated with four author of this paper, with two are female and two are male. As acknowledged, the diversity of annotators plays a crucial role in reducing bias and enhancing the reliability of the benchmark. These annotators have knowledge in this domain, with different genders, ages, and educational backgrounds. To ensure the annotators can proficiently mark the data, we provide them with detailed tutorials, teaching them how to evaluate model responses more objectively. Specifically, they are required to give judgments without considering answer lengths, and certain names or positions of the response. Besides, we implement cross-validation between different annotators and conduct continuous monitoring to ensure they are maintaining objectivity and fairness. We provide screenshots of the instruction and annotation in Figure 6.

B.3 LLM simulations

- model: gpt-3.5-turbo-0125
- temperature: 1
- seed: index of the abstract in the dataset
- top_p: 0.9

B.4 Estimations of LLM impact

The choice of words is based on the value of f^* and $\hat{r}$.

- $\frac{1}{f^*}$: 30, 40, 50, 60, 70, 80, 100, 150, 200, 500
- $\hat{r}$: 0.4, 0.5, 0.6, 0.7 (corresponding value of $\frac{\hat{r}+1}{\hat{r}^2}$)

C Supplementary Results

C.1 AIGC detector results

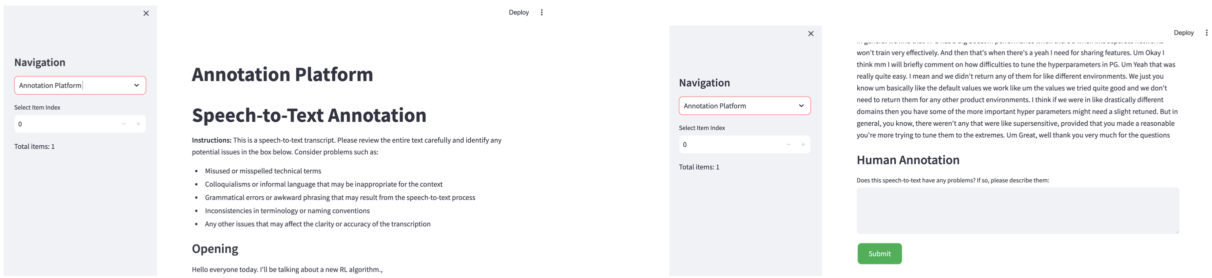
In this section, we present the detection results from Fast-DetectGPT (Bao et al., 2023), GLTR (Gehrmann et al., 2019) and RADAR (Hu et al., 2023). All experiments are conducted in a dual-4090 server, detailed as follows:

- **Fast-DetectGPT** serves as a coarse binary-classification, and the results are summarized in Figure 11 and detailed in Figures 12, 13, and 14. A higher criterion represents a greater probability of machine-generated text.
- **GLTR** works as a coarse four-class classification, and the results are summarized in Figures 15 to 18 and detailed in Figures 19 to 30. It analyzes GPT-2's predictions at each position in the text and calculates the rank of every word. Words ranked in the top 10 are classified as green, those in the top 100 as yellow, in the top 1000 as red, and the rest as purple.
- **RADAR** provides robust binary classification experiments, and the results are summarized in Figure 31 and detailed in Figures 32, 33, and 34. A higher threshold indicates an increased likelihood of identifying machine-generated text through adversarial learning techniques.

Consistent with previous analyses, there is no obvious effect of LLMs in the presentations of these oral presentations, demonstrating the need for fine-grained detection such as the word frequency experiments we presented before.

Table 1: Statistics and source data link of each conference analyzed in our paper. Missing data are marked with "*".

Conf.	Year	Oral	Spotlight	Poster	Source Link
ICML	2021	166	1017	*	https://icml.cc/Conferences/2021/Schedule
	2022	118	1115	*	https://icml.cc/Conferences/2022/Schedule
	2023	155	*	1673	https://openreview.net/group?id=ICML.cc/2023/Conference
	2024	144	191	2275	https://openreview.net/group?id=ICML.cc/2024/Conference
NeurIPS	2021	60	284	2286	https://openreview.net/group?id=NeurIPS.cc/2021/Conference
	2022	*	*	*2671*	https://openreview.net/group?id=NeurIPS.cc/2022/Conference
	2023	67	378	2773	https://openreview.net/group?id=NeurIPS.cc/2023/Conference
ICLR	2021	53	114	692	https://openreview.net/group?id=ICLR.cc/2021/Conference
	2022	55	175	864	https://openreview.net/group?id=ICLR.cc/2022/Conference
	2023	*90*	*281*	1201	https://openreview.net/group?id=ICLR.cc/2023/Conference
	2024	86	367	1807	https://openreview.net/group?id=ICLR.cc/2024/Conference



(a) Instruction for annotators.

(b) An example for human annotation.

Figure 6: Annotation details.

C.2 Issues in Speech-to-Text and Abstracts

In this section, we introduce the issues we discover in the speech-to-text process and abstracts, detailed as follows:

- There are words such as *"ok"* and *"okay"* which have identical pronunciations and meanings, but Speech-to-Text systems may exhibit a preference for one over the other. For instance, in the ICML oral presentations of 2021 and 2022, *"ok"* was predominantly used, whereas in 2023 and 2024, *"okay"* became more common. Such discrepancies introduce inconsistencies that can affect the accuracy of word frequency analyses.
- The conversion of technical terms by Speech-to-Text systems may not always be accurate. For example, in multiple oral presentations, the term *"LoRA"* (a well-known PEFT method) was incorrectly transcribed as the name *"Laura"* in the subtitles, which is evidently erroneous. Fortunately, the impact of such errors on the overall word frequency statistics appears to be minimal.
- While retrieving abstracts from ICML, ICLR, and NeurIPS conferences, it was observed that certain abstracts contained formatting elements such as italics or citations, e.g., `\textit{approximately valid}` and `\cite{chen2020learning}`. These formatting artifacts can distort the frequency analysis of certain words, such as *"cite."*
- A comparative analysis was conducted between the original subtitles of the oral presentations and the transcripts generated by Whisper's base model (Radford et al., 2022) from the audio files of the same presentations. This comparison, evaluating accuracy, consistency, and coherence, indicated that the original subtitles were significantly superior to the transcripts generated by Whisper. Therefore, the original subtitles are deemed more appropriate for use in word frequency analysis.

C.3 Word frequency

Figure 8 also shows the increasing trend in the frequency of these words.

D Case Study

In this section, we sample some oral presentations at each machine learning conference and visualize the MGT detection results. An example of the abstract is shown in Figure 35 and the presentation of an oral paper in ICML 2022 shown in Figure 36, 37, and 38.

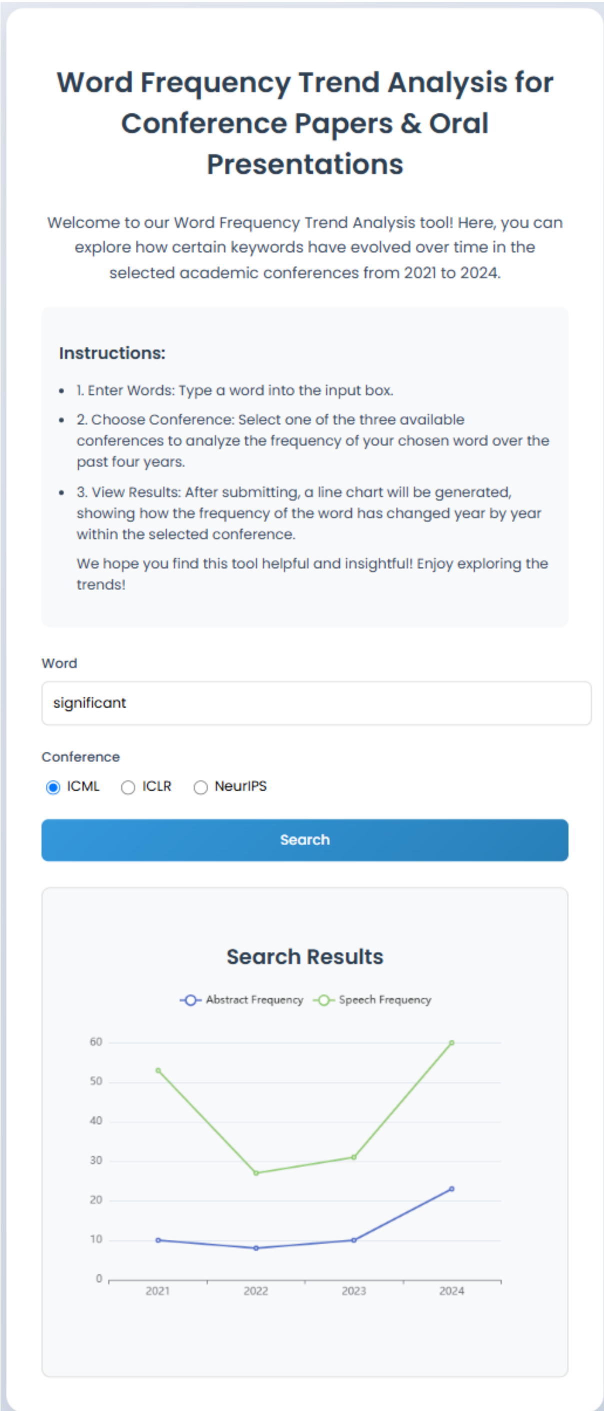


Figure 7: An example of our monitoring platform.

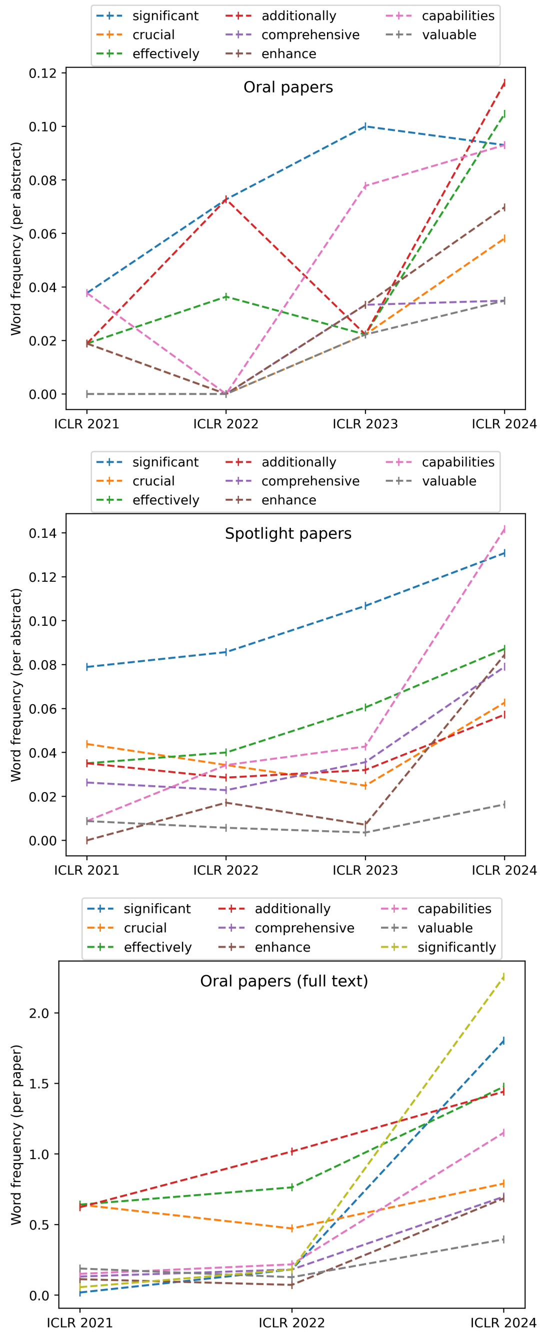
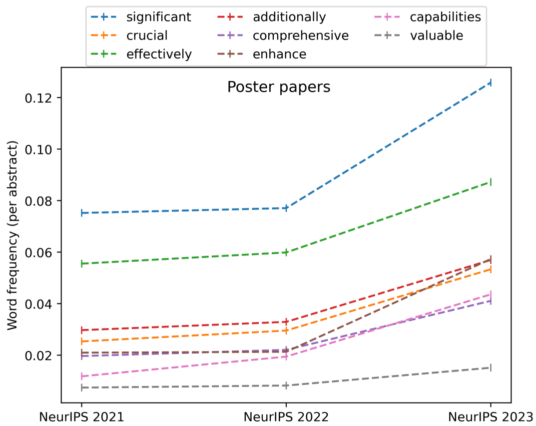
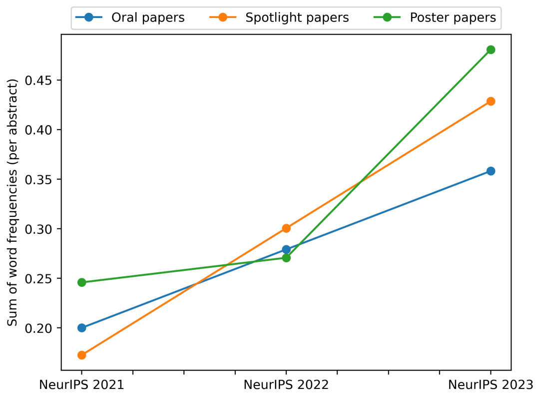


Figure 8: Word frequency *oral* papers and *spotlight* papers in ICLR.



(a) Word frequency in abstracts of *poster* papers.



(b) The sum of the frequencies of example words in the abstracts of different types of papers.

Figure 9: Word frequency in abstracts of NeurIPS papers. Papers accepted in 2022 are considered poster papers, and word frequencies for spotlight and oral papers were completed by interpolation.

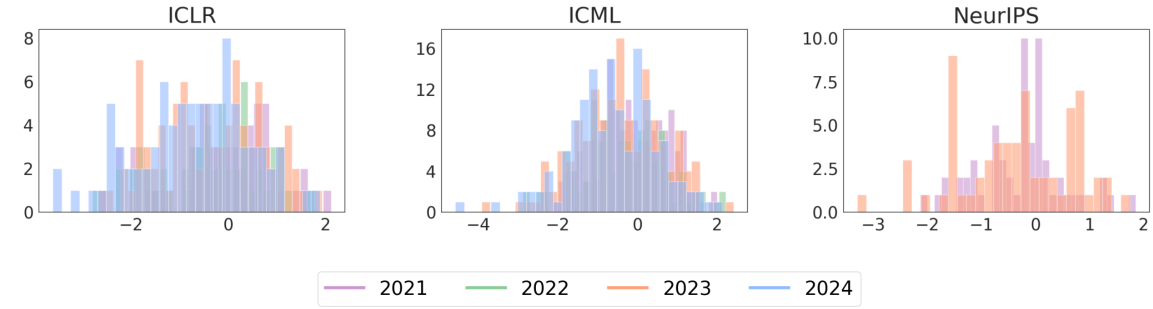


Figure 10: Criterion distribution in abstracts of oral works.

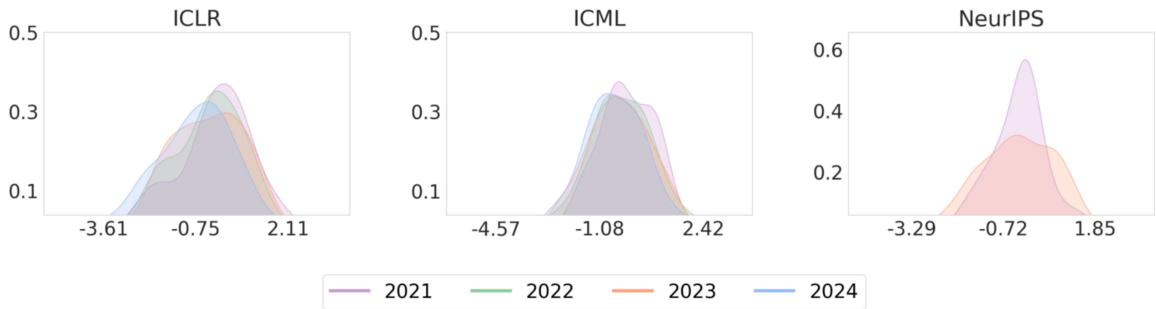


Figure 11: Criterion distribution in abstracts of oral works.

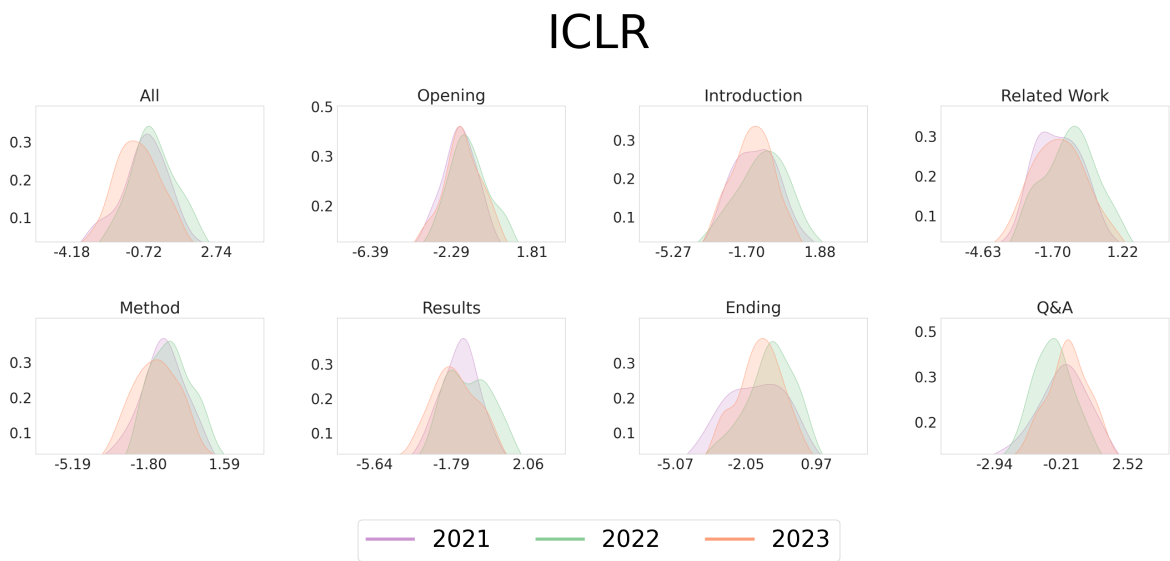


Figure 12: Criterion distribution in ICLR oral presentations.

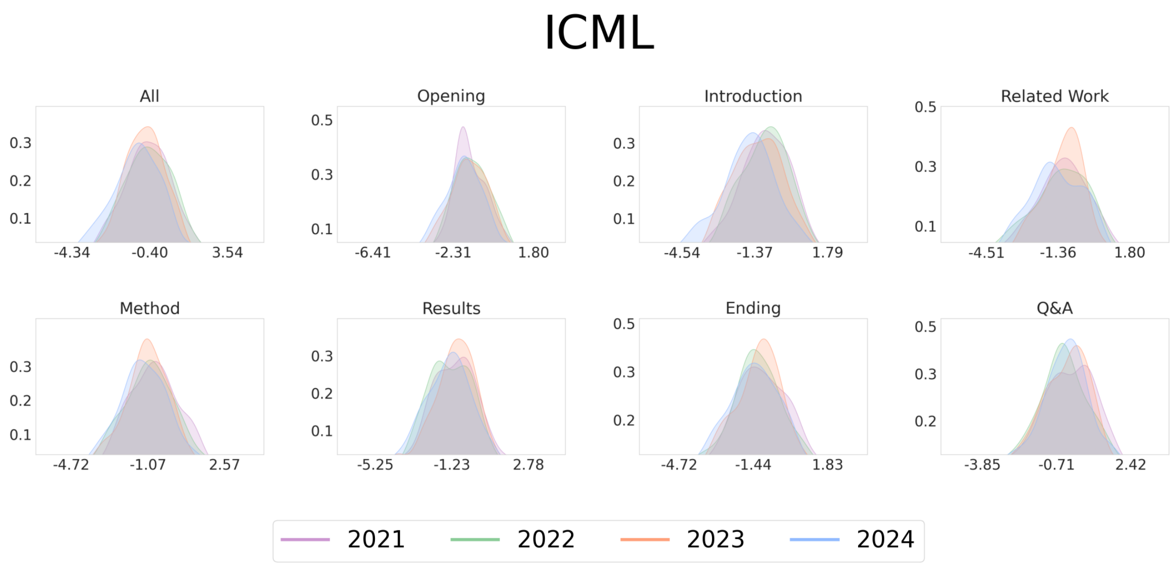


Figure 13: Criterion distribution in ICML oral presentations.

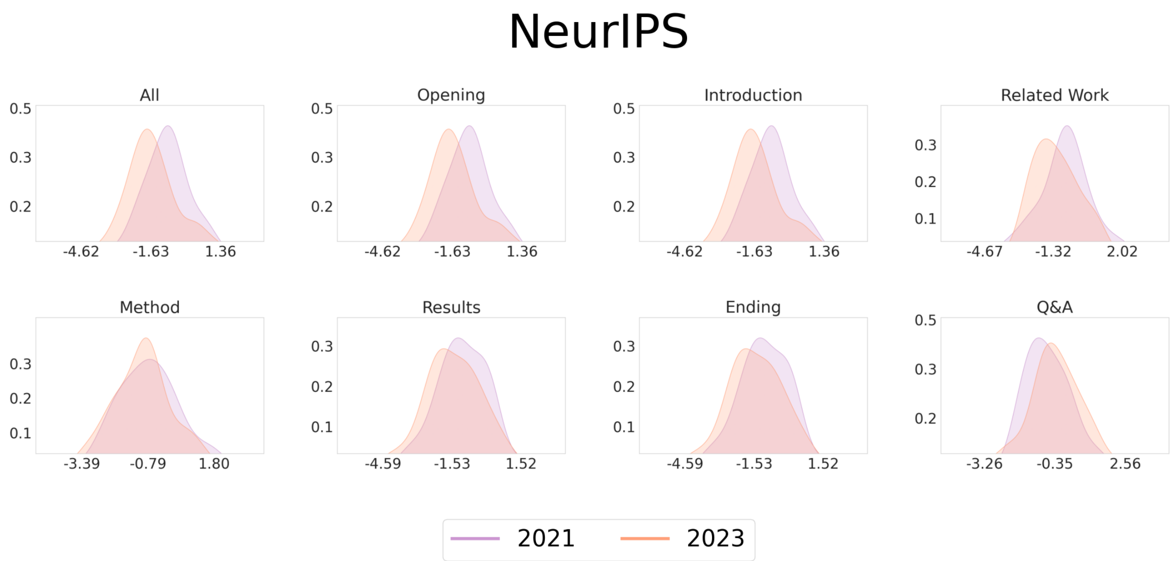


Figure 14: Criterion distribution in NeurIPS oral presentations.

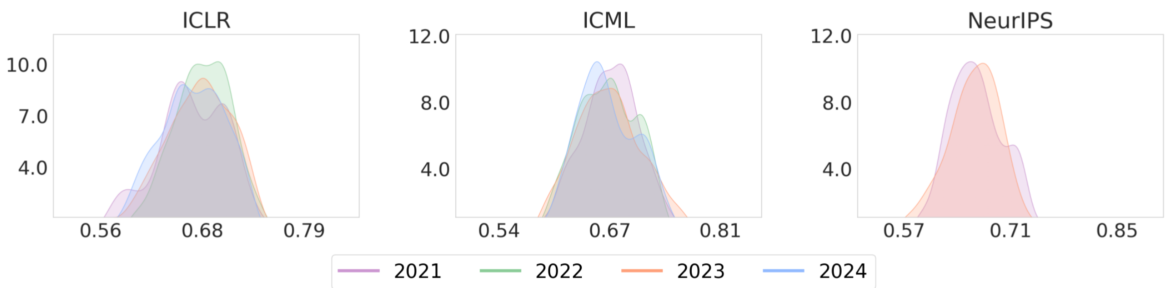


Figure 15: Green class of GLTR in abstract of oral works.

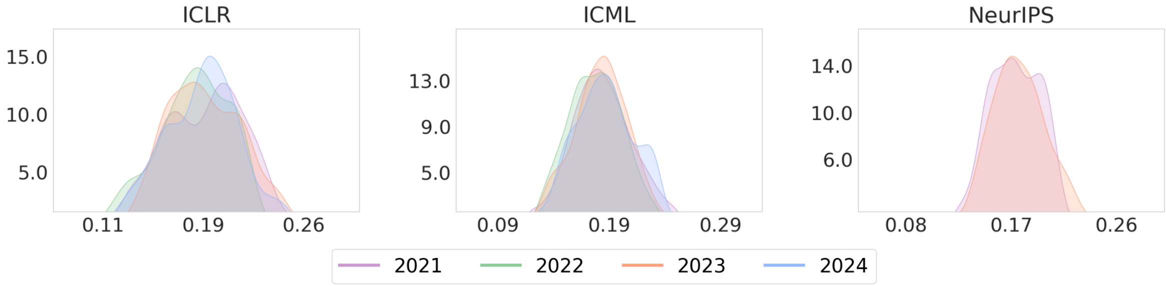


Figure 16: Yellow class of GLTR in abstract of oral works.

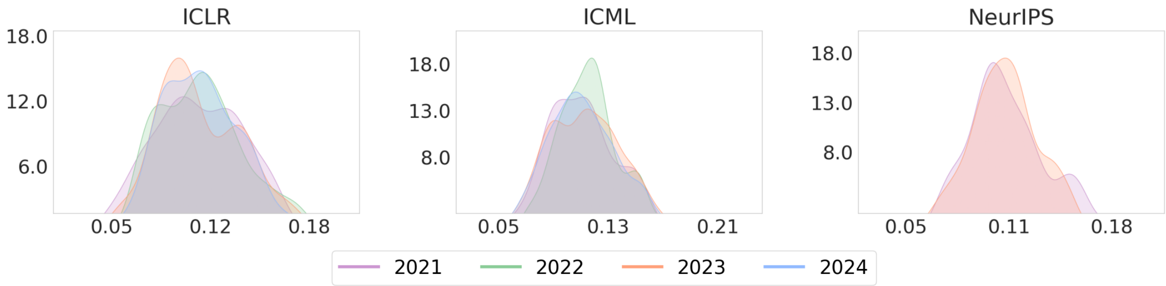


Figure 17: Red class of GLTR in abstract of oral works.

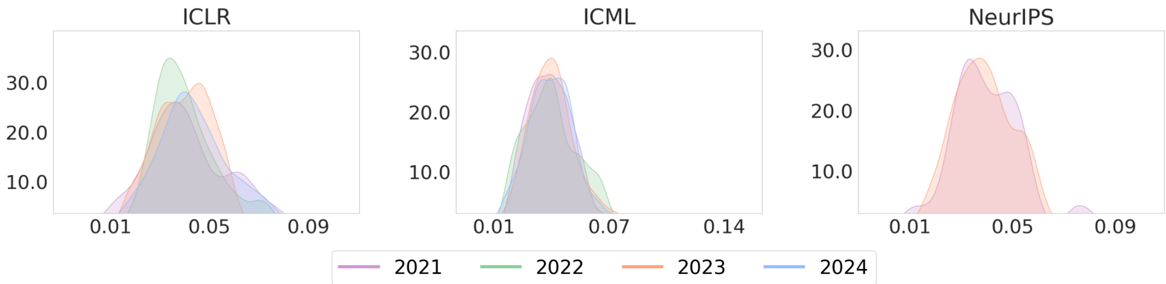


Figure 18: Purple class of GLTR in abstract of oral works.

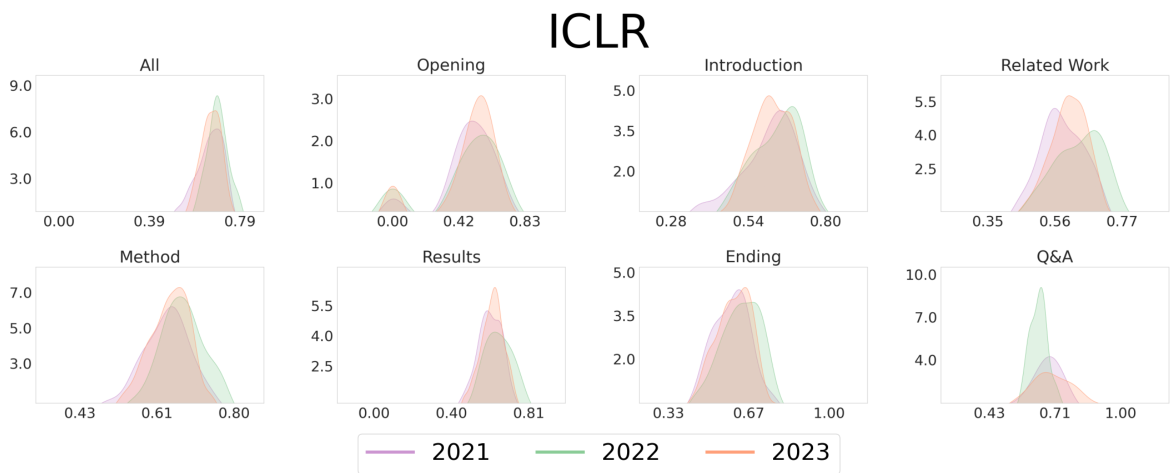


Figure 19: Green class of GLTR in ICLR oral presentations.

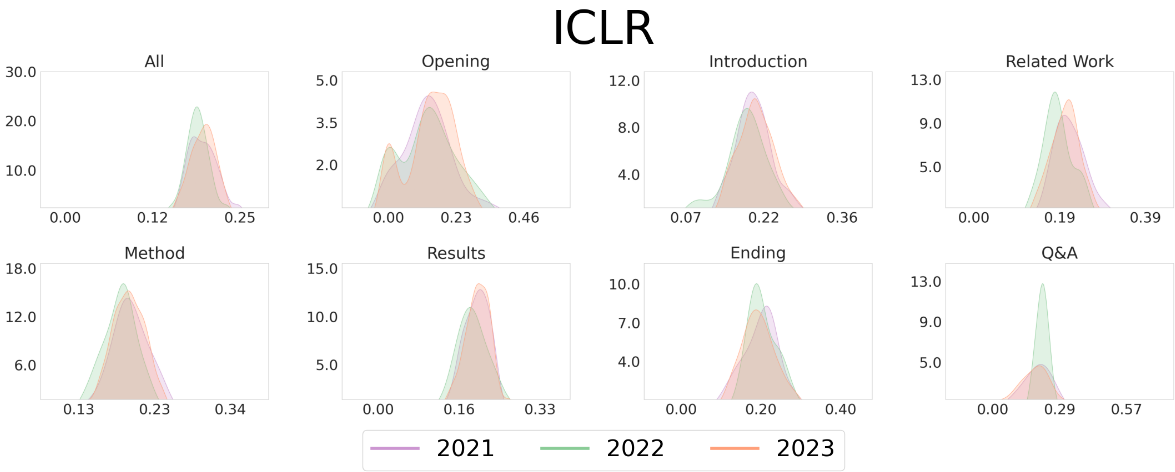


Figure 20: Yellow class of GLTR in ICLR oral presentations.

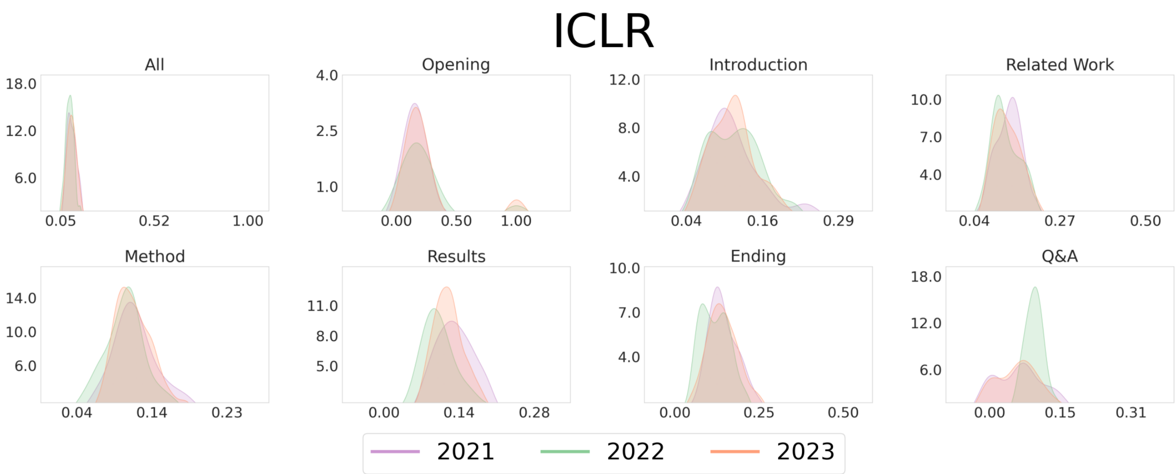


Figure 21: Red class of GLTR in ICLR oral presentations.

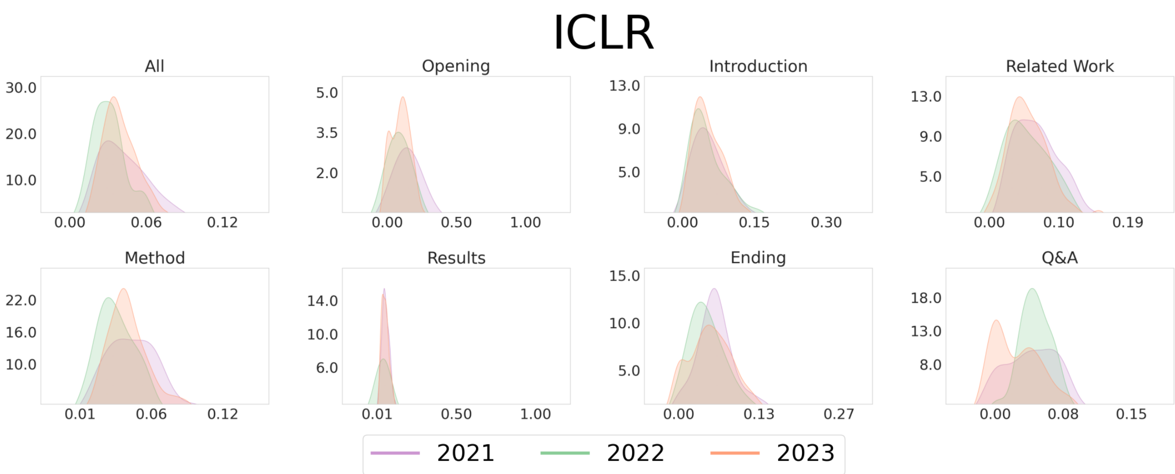


Figure 22: Purple class of GLTR in ICLR oral presentations.

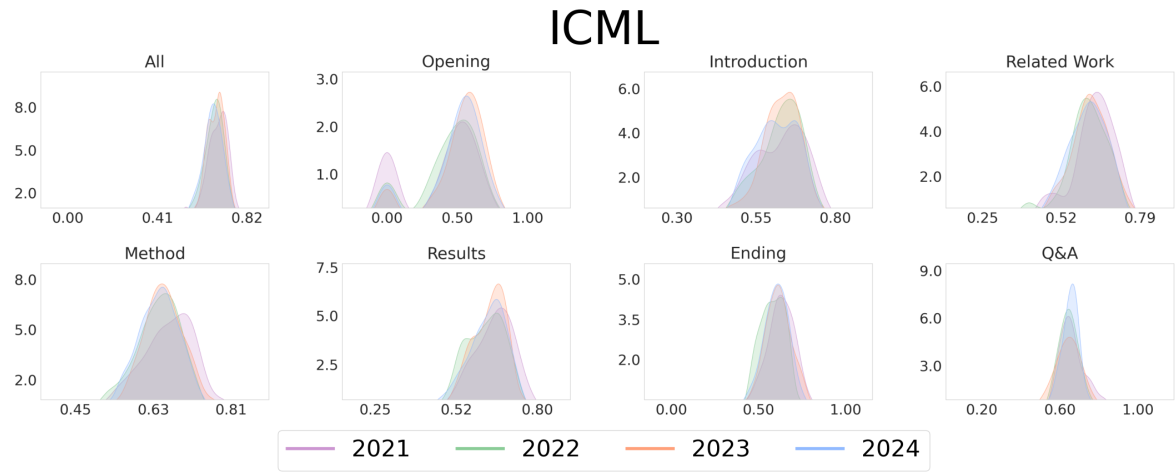


Figure 23: Green class of GLTR in ICML oral presentations.

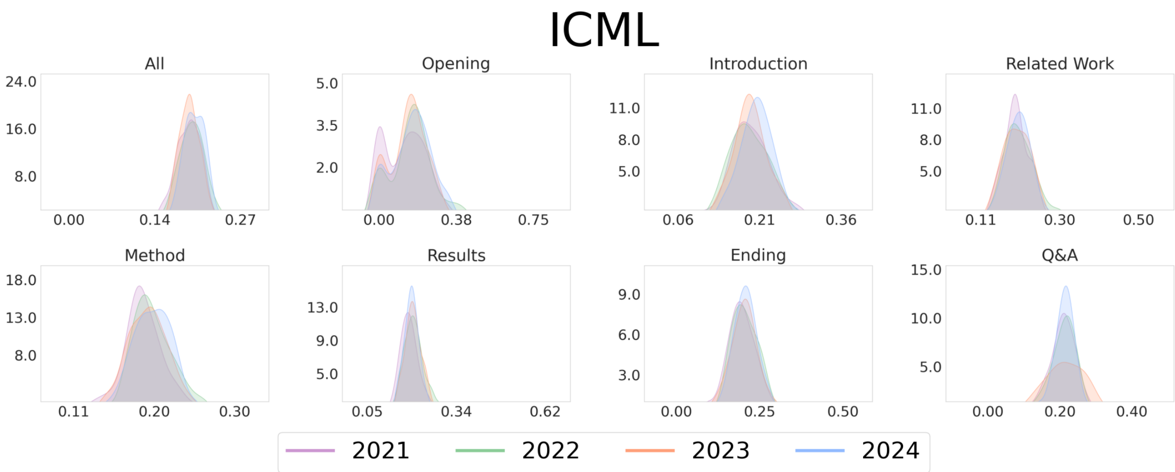


Figure 24: Yellow class of GLTR in ICML oral presentations.

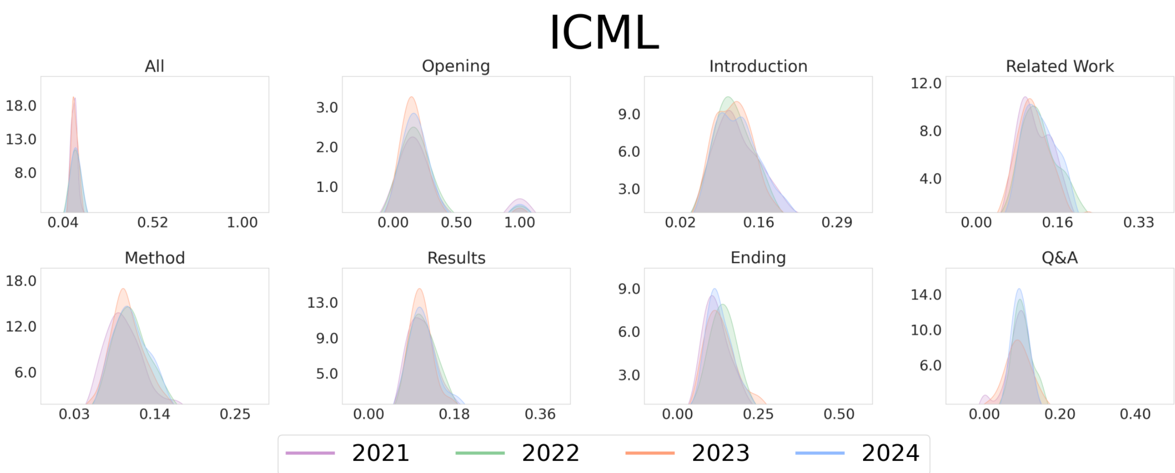


Figure 25: Red class of GLTR in ICML oral presentations.

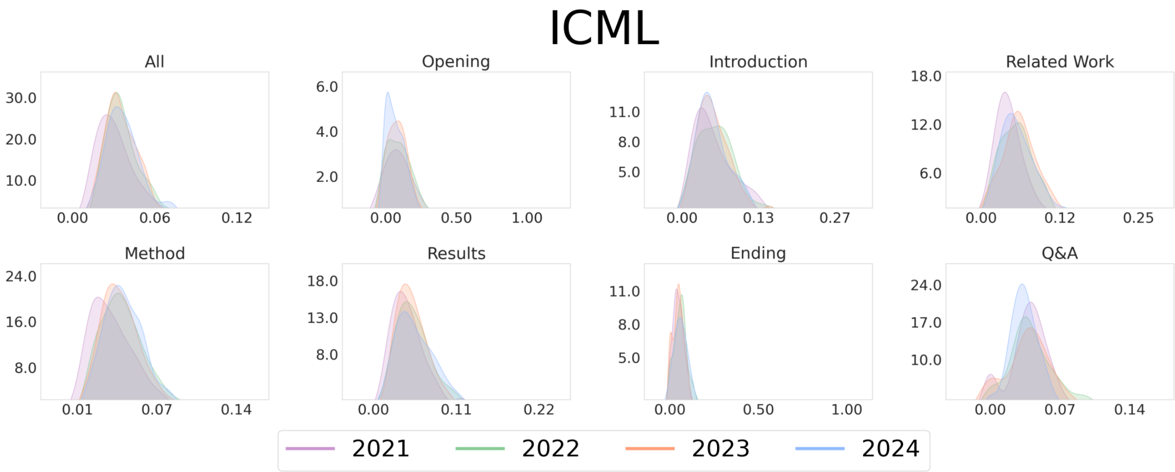


Figure 26: Purple class of GLTR in ICML oral presentations.

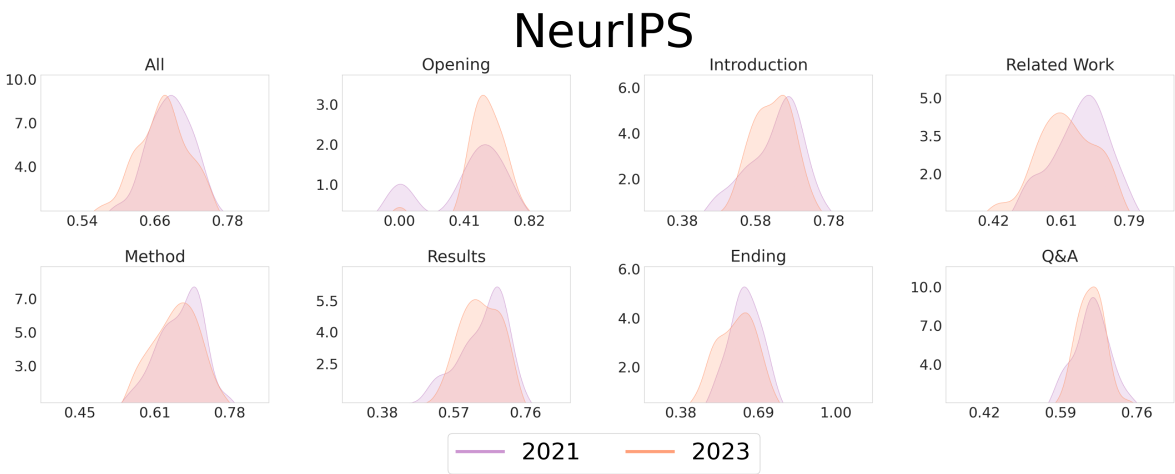


Figure 27: Green class of GLTR in NeurIPS oral presentations.

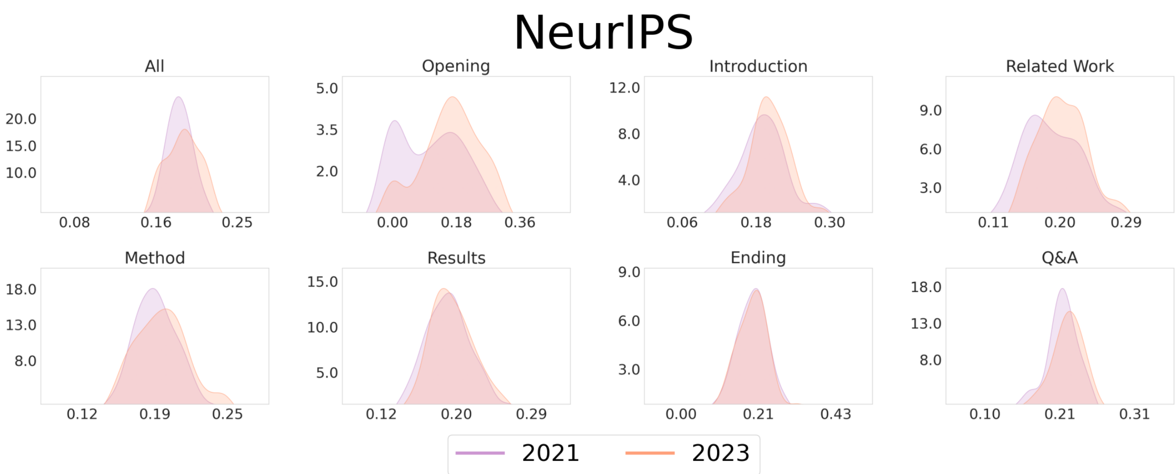


Figure 28: Yellow class of GLTR in NeurIPS oral presentations.

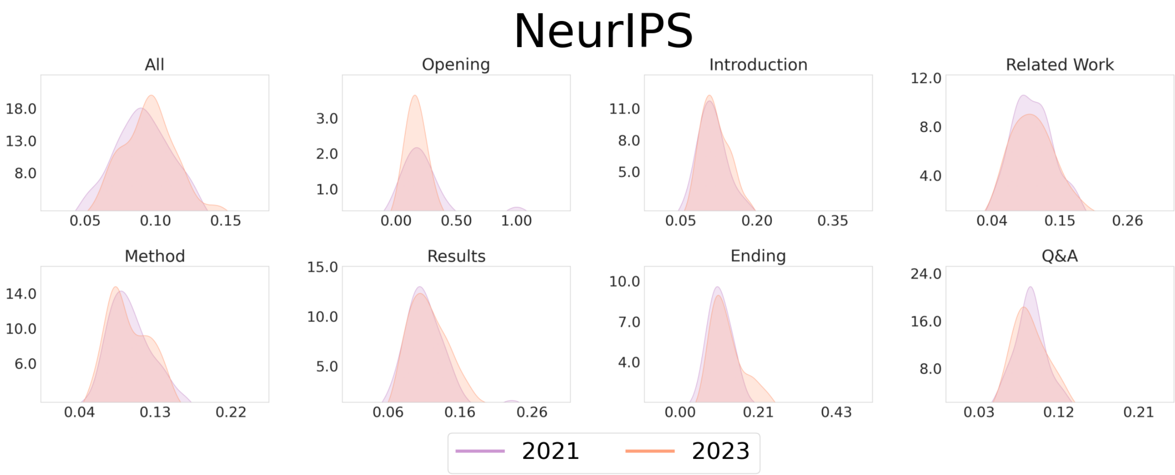


Figure 29: Red class of GLTR in NeurIPS oral presentations.

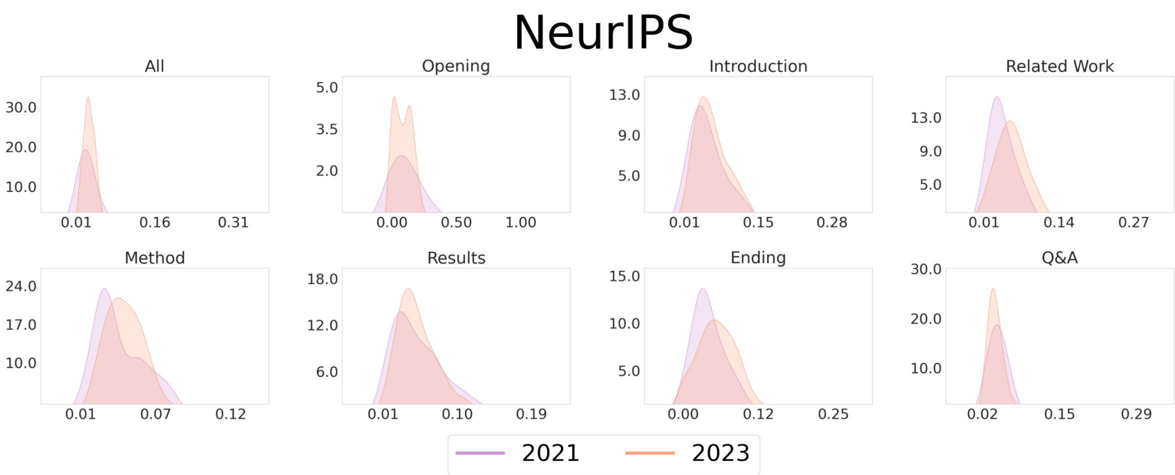


Figure 30: Purple class of GLTR in NeurIPS oral presentations.

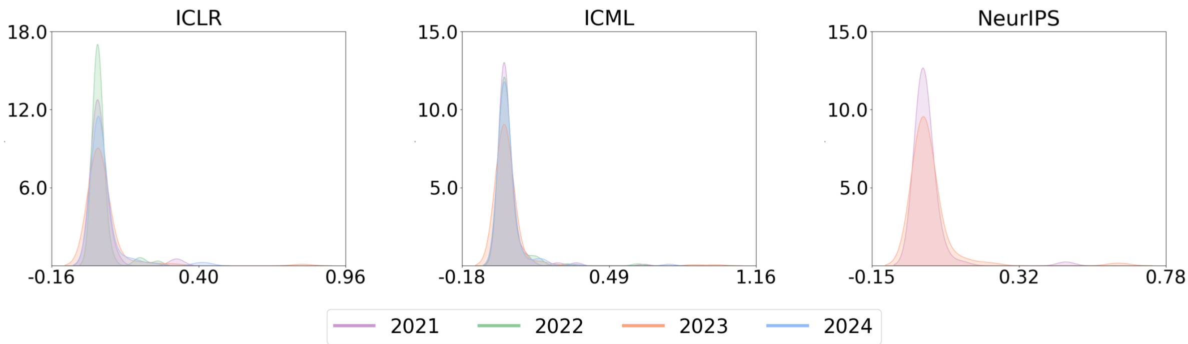


Figure 31: AI-Generated probability in abstracts of oral works.

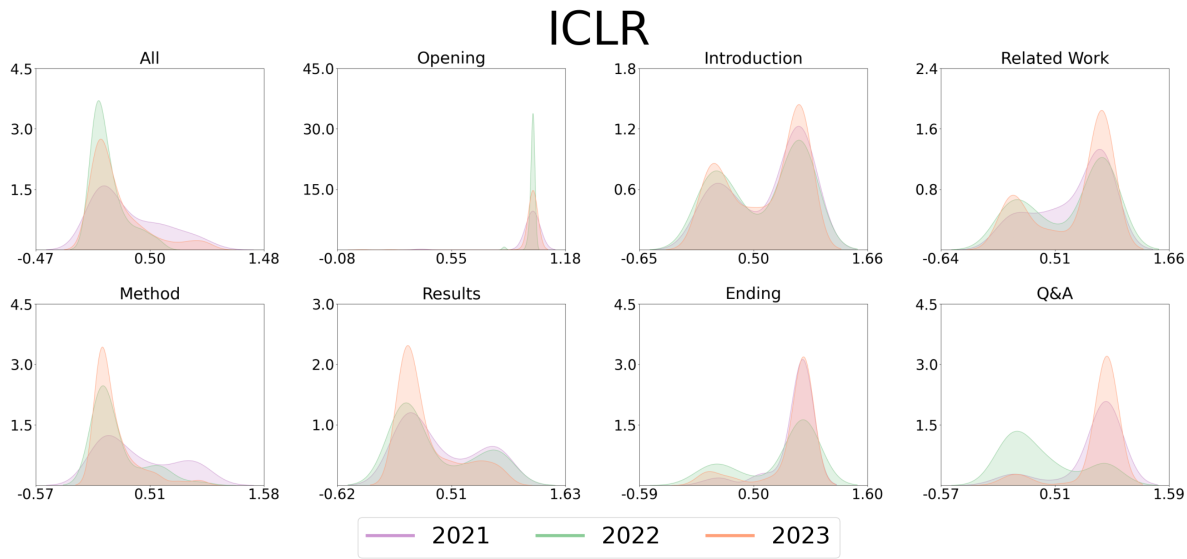


Figure 32: AI-Generated probability in ICLR oral presentations.

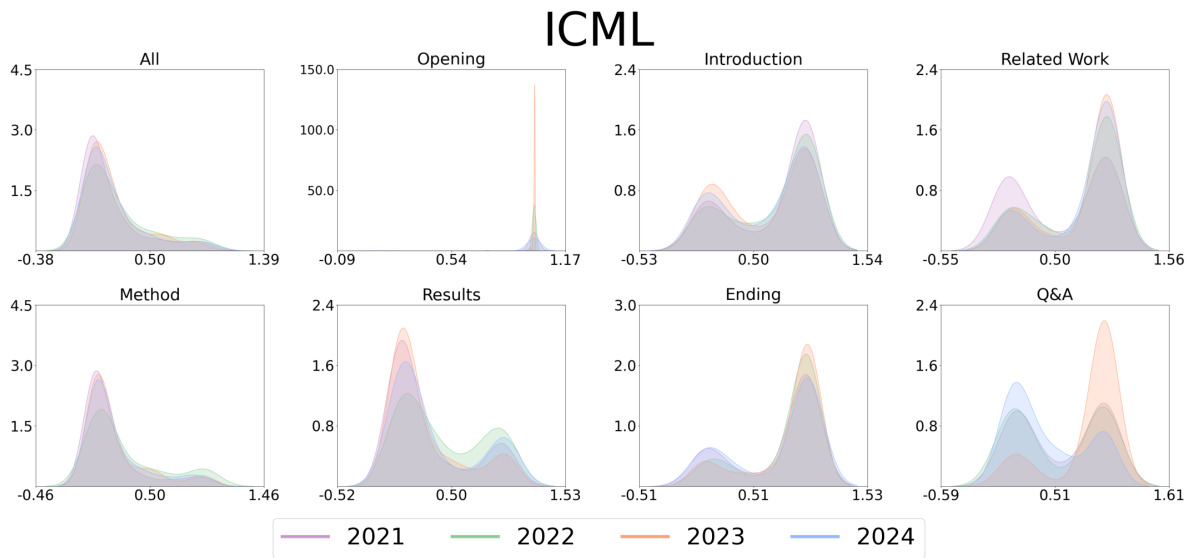


Figure 33: AI-Generated probability in ICML oral presentations.

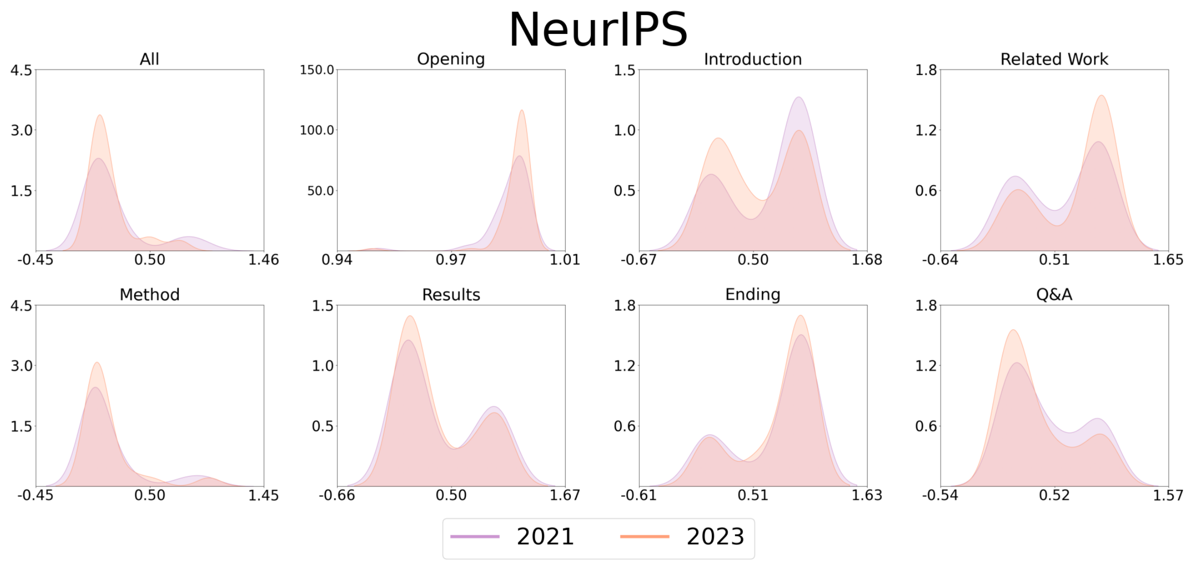


Figure 34: AI-Generated probability in NeurIPS oral presentations.

Abstract Example

Large language models are trained on vast amounts of internet data, prompting concerns that they have memorized public benchmarks. Detecting this type of contamination is challenging because the pretraining data used by proprietary models are often not publicly accessible.
We propose a procedure for detecting test set contamination of language models with exact false positive guarantees

and without access to pretraining data or model weights. Our approach leverages the fact that when there is no data contamination, all orderings of an exchangeable benchmark should be equally likely. In contrast, the tendency for language models to memorize example order means that a contaminated language model will find certain canonical orderings to be much more likely than others. Our test flags potential contamination whenever the likelihood of a canonically ordered benchmark dataset is significantly higher than the likelihood after shuffling the examples.

We demonstrate that our procedure is sensitive enough to reliably detect contamination in challenging situations, including models as small as 1.4 billion parameters, on small test sets with only 1000 examples, and datasets that appear only a few times in the pretraining corpus. Finally, we evaluate LLaMA-2 to apply our test in a realistic setting and find our results to be consistent with existing contamination evaluations.

Figure 35: Abstract Example.

Presentation Example - 1

'Opening': 'Hi Embassy Kuala',
'Introduction': 'And this presentation is about our clear paper titled complex query answering with neuron predictors. And this is joint work with Erik Daniel Helle',
'Related Work': 'So let's consider the setting. We have a knowledge graph graph structured knowledge base where knowledge about the world is represented in the form of relationships between entities. In the knowledge graph like this, one notes correspond to objects in the world or also upset concepts and I just correspond to relationships between these. In this example we have a knowledge graph about biomedical entities which tells us that paroxetine is used to treat anxiety and as bisimulation as a biological action. And that a pixel band treats deep vein thrombosis and one of its pharmacological effects is that it's an anticoagulant. Now one problem with real world where large scale knowledge graphs is that they are often incomplete. In this particular case, we are missing a link stating the topics abandons a medication that deep vein thrombosis is a disease and that a pixel band kind of side effects when taken together with oxygen. A very effective solution to identifying missing links in large knowledge graphs is via neural in production models in neural prediction. The underlying idea is that we can learn an embedding vector for all the nodes in the graph. For example, in this case we will learn an embedding back to for a pixel band proximity in bisimulation, Deep vein thrombosis and all other entities in the graph. Now assume we want to know the type of relationship if any between a pixel band and boxed in the likelihood that two entities in this case picks abandoned parks 18 are linked by a given type of relationship in this case interacts is a function of the embedding vectors of the source node. In this case a pixel man and embedding vector of the target node in this case, proximity in. And we can use this function for ranking missing links and find out that for example, a pixel bunnies are likely to interact with oxytocin. Even if this link is not directly available in the knowledge graph. Now consider the problem of answering complex queries on incomplete knowledge graphs. Here we have a query which medications have side effects when taken with drugs for treating anxiety and we want to have a list of medications that hands are query. This query can be formalized in logic form and it reads as follows, find M where M is available such that there is a D. Which is also variable, such that I am interacts with the and the tweets anxiety know that this is just an example and our method supports arbitrarily complex logic queries with conjunctions and dysfunctions. Now the best solution for solving this problem proposed so far is the following. First we automatically generate millions of complex query answer path and then we train a deep neural network to produce the correct answers given the creator, the neural model works as follows. First we represent the complex query as a graph. We were each note um corresponds either to a variable or to an entity in the query. Then the graph is passes through a deep neural network which will return ranking list of answers. Now that two main problems with this approach, one is that training is extremely expensive, since the models need to be trained on millions of genetic queries. Also, it's not really clear what happens if we evaluate on queries that differ from the queries that we used for training. Another issue is that there is no explanation for the reasons why a given answer was predicted to solve these problems.'

Figure 36: Presentation Example - 1

Presentation Example - 2

'Method': 'In this work, we propose a completely new paradigm for answering complex queries on incomplete knowledge graphs. We first train a model F. I, financing simple atomic queries like which drugs to it anxiety. And then we convert its query into an optimization problem where we need to identify the optimal values of the variables in this case, M and D. In the query that maximize the likelihood of both that M interacts with the and likely that the tweets anxiety. Then those two likelihoods are aggregated using a T norm, which is a continuous relaxation of the logical and from fancy logic. And depending on whether we search for the best values of M and the united creator continuous space, we can cast this problem as either as a continuous or discrete optimization problem. In the discrete case, we want to identify the best mapping from variables to entities that maximizes the score of the query. We experimented with a very simple greedy approach to solve this problem, we first start with the variable D. And search for the K most likely values for the by finding the top K. Treatments for anxiety according to the Newell in prediction model. Then for each of the candidate values of the we identify the most likely values for the problem and then we compute the query scores, originated parts of values for M and E, and which are the most likely values for MMD. Another approach we experimented with consists in directly optimizing the vector representations associated with the variables MMD. This can be done by a gradient based optimization. We first initialize the embeddings of M and the randomly. And then we optimize the embedding of them and the to maximize the score of the query. And finally replaced the embedding of them with the embedding of all the candidate entities and rank them using the corresponding query score',

'Results': 'Yeah. So we experimented on a variety of complex grid structures. And despite being only trained on simple creation, we can see that our model systematically generalize is too complex queries better than models trained on complex queries. In the first place, He had results on 3 - 15 K 237 for different complex great types. Here we can see the results on three basic 15 K. Any other results on 995 and here we have the average results between all types of complex queries that we considered. And we can see that our model produces significantly more accurate results on all the traii datasets. This improves of the existing models, both in terms of the data efficiency, because we only need to train our model on a much smaller dataset of simple queries and also in terms of out of out of distribution generalization, because we get better generalization accuracy on complex queries without having to train on them In the 1st place. Another really nice feature of this model is that it can provide explanations for its predictions. Other models in the space only returns a list of answers to the query. For instance, in this case, the answer is being produced at a pixel ban, I'm 15 block setting and others. Our model can also be used to provide the intermediate results associated with each query in the form of the variable assignments used to produce the answer. For example, here we can see that a pixel ban and I'm tellin were considered as answers because according to the model, they interact with oxytocin while dual oxygen was considered as as an answer because according to the model, it interacts with pregabalin and this allows us to check whether the results are being produced for the right reasons. In this case box setting and pregabalin are two possible treatments for anxiety. So the model is producing the correct answers for the correct reasons. This is not always the case. For example, consider the following query from feedback is 15 K 237. What international organizations contain the country of nationality of thomas Aquinas here, the model was able to return the correct set of answers NATO asI, D U and Wt O. However, this return for the wrong reasons. Um thomas Aquinas was mistakenly assumed to be from the US from the UK are from Germany. While the correct nationality of thomas Aquinas is italian. Our model enables us to detect such iris and possibly cracked them by refining the underlying newly prediction model.'

Figure 37: Presentation Example - 2

Presentation Example - 3

'Ending': 'So, to summarize in this paper, we propose a new approach France and complex creates on large scale and incomplete knowledge graphs. In our approach, we first train a neural prediction model on the task of answering simple atomic creation. And then we cast the problem of answering complex grace as an optimization problem where we need to find the mapping from variables, quantities that maximizes the score of the complex square. And we show that our approach generalizes extremely well to complex queries despite not having been trained on them in the first place, all the source code, the pre trained models and datasets online at this link. And if you want to collaborate on this topic or have any questions or comments, please feel free to reach out to us and thank you for listening. thank you very much So you may be. Now we will move to the next paper about complex query answering',

'Q&A': 'So we had one question from Cartier uh she mentioned that the work is supplied to relatively small data said she wants to know essentially what happens if you try to scale it up. And I follow with the second question quickly. I was wondering essentially because you have these different types of you know, complex queries and then you have two types of optimization. One is the continuous relaxation, another is the greedy discrete approach. Which one is better in which case is there any kind of ideal correlation between types of I don't know, complexity of the queries, how they are designed and so on or Yes, thank you. Thanks to the great questions. So about the first one about getting the method up to much larger quantities of data like the millions um of data points they use in other methods. I think it's super interesting and we can I think we might observe significant improvement for that from that. For example. One thing that we're doing at the moment is we are fixing how so we select one continuous relaxation of the logical end and are as an input parameter, we select the two norm and that economic beforehand as an input parameter and then we just execute the method but I think by scaling this matter to larger conditional data and to complex creates during training we can think of for example to um to train how we represent the logical and and or within the architecture and we can we would be able to use for example, paramedic economics and economies, which we are not touching at the moment because we don't use complexities, complexities during training. And also um the naturally predictors that we use for answering economic queries are not really trained. Um so um are not really trained on complex crises in in the first place that trained on atomic quiz. And I think that so the scores are not really trying to do interact together in some sense. So I think we might be able to observe a significant improvement from that as well. About the other question. So what we observed is that the discrete search seems to work consistently better than the continuous search um across all datasets and the types of complex queries. And we think that that might happen because the the continuous search for the entity representations might find some entity representations that do not do not correspond to any real entities in some sense. Or maybe the model might hallucinate some, I don't know, a dog with seven legs. Uh The other entities that are not really close in terms of representations to real entities in the knowledge base. So I think that's my that that might be a reason why um this could search works consistently better on uh continue such. Um Thanks. Okay, thank you very much. I think that answers the question. So I'll hand over tell me actually have another question about the linking paper. So you have a two step approach for your first predict links and then you run queries on top of that. Can you comment on? No, no, basically the links are predicted. Um so basically the links are predicted as kind of a part of the optimization problem. We translate the complex series into each complex query is translated into an optimization problem and the neural predictor. Um that makes whether there is a link between two entities is kind of a component of this optimization problem because correcting like all drinks beforehand, doesn't scale up because you can have billions of possible links uh in the name, like materializing all the links beforehand doesn't really uh scale also like you need to decide uh what whether to materialize a link or not. So basically the link prediction process is part of the complex variants diagnosis. Uh this map'.

Figure 38: Presentation Example - 3