Exploring the Limits of Prompting LLMs with Speaker-Specific Rhetorical Fingerprints

Wassiliki Siskou[1,2] and **Annette Hautli-Janisz**[2]
[1] University of Konstanz, Germany
[2]University of Passau, Germany
wassiliki.siskou@uni-passau.de
annette.hautli-janisz@uni-passau.de

Abstract

The capabilities of Large Language Models (LLMs) to mimic written content are being tested on a wide range of tasks and settings, from persuasive essays to programming code. However, the question to what extent they are capable of mimicking human conversational monologue is less well-researched. In this study, we explore the limits of popular LLMs in impersonating content in a high-stakes legal setting, namely for the generation of the decision statement in parole suitability hearings: We distill a linguistically well-motivated rhetorical fingerprint from individual presiding commissioners, based on patterns observed in verbatim transcripts and then enhance the model prompts with those characteristics. When comparing this enhanced prompt with an underspecified prompt we show that LLMs can approximate certain rhetorical features when prompted accordingly, but are not able to fully replicate the linguistic profile of the original speakers as their own fingerprint dominates.

1 Introduction

Recent research on LLM alignment shows that depending on the task, LLMs can mimic or imitate human language to an extent that the generated content is indistinguishable from or even surpasses the quality of human language. Mimickry is an intermediate step towards impersonation, the latter assuming that an agent not only copies general human behavior, but pretends to be a specific person and acts accordingly. In this paper we show that we can nudge LLMs towards impersonation, but that there remains a gap between actual human and generated content. We do so by crafting speaker-specific rhetorical fingerprints that we first use as prompt enhancements and then employ as means to identify the differences between human and generated content.

The setting in which we test this is sensitive: we use anonymized parole suitability hearing transcripts from California and task the model with generating the decision statement of the presiding commissioner. By distilling a rhetorical fingerprint of the commissioner across multiple hearings, we compare the effect of prompting several models with the fingerprint-enhanced prompt and their performance when prompted with a general prompt not containing the fingerprint. The experiments show that all LLMs seem to have their own linguistic fingerprint from which they do not deviate even if prompted so. Additionally, prompting the models to replicate the style they observe in a given text, does not succeed, as their own fingerprint remains more dominant.

2 Related Work

Recent studies explore how effectively LLMs can mimic human-like behavior in different aspects. For example, Milička et al. (2024) task different versions of OpenAI's GPT to impersonate children between two and six years old. Their findings show that the models are able to adapt their linguistic behavior to the developmental stage expected from them. Salewski et al. (2023) observe a boost in performance by prompting the LLM to act as a domain expert, but they also identify the reproduction of gender, age and racial biases in the model's output. Herbold et al. (2024) show that LLMs can impersonate politicians to the extent that the model responses are judged more authentic, relevant and coherent than the actual human responses.

To the best of our knowledge, there has been no work on how well LLMs perform in mimicking human-like speech with rhetorically enhanced prompts. Recently, several studies have focused on the capabilities of LLMs to emulate human writing styles by looking into coarse and more fine-grained linguistic analysis (Bhandarkar et al., 2024; Alhafni et al., 2024). Bhandarkar et al. (2024) test the performance of 12 pre-trained LLMs for stylistic rewriting, by instructing them to mimic the author's

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 143–154, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-014

writing style together with shallow guided instructions regarding different linguistic features. While their results show that current models are able to replicate author style to some extent, they are not capable of producing text that is fully indistinguishable from that of the original author.

In a more recent approach, Dinu et al. (2025) tested how good LLMs are at imitating writing styles by prompting it to complete an author's unfinished novel. While LLMs perform acceptably in mimicking the literary style, their quality was not assessed as being as good as the human written ones.

In this present study, we build on this line of research. We extend the focus from imitating writing styles to simulate spoken language, by using detailed rhetorical prompts. We distill the linguistic characteristics of each person and enhance the prompts dynamically to simulate specifically tailored spoken natural language dialogue.

3 The data

3.1 Parole Suitability Hearings

In California, an inmate's potential to be reintegrated into society despite serving a life sentence is assessed by one presiding and one deputy commissioner during parole suitability hearings (PSHs). After an hour-long interview with the inmate and their attorney, the presiding commissioner communicates the decision, taking into account the inmate's answers, a review of the rehabilitation plan, psychological assessments and disciplinary records.

Typically, decision statements follow a structured scheme, including an introduction, the announcement of the final decision, a discussion of the mitigating and aggravating factors, such as the institutional behavior of the inmate and the life crime itself. In case of a parole denial, commissioners may give recommendations for improvement. Additionally, they are required to set a denial length, which determines when the inmate is eligible to reappear before the parole board. While these elements are consistently covered by all commissioners, each commissioner may change the order of covering those parts in their statements or may choose to discuss one factor more in detail than others. Rhetorically, the decision statement has to establish authority by keeping a professional tone, at the same time signaling empathy and a reasoned judgment. We incorporated all these structural requirements in our prompt design to ensure alignment of the generated decision statements with the content observed in actual parole hearing decision statements.

3.2 The PSH v1.0 corpus

The dataset that underlies the present study, PSH v1.0, comprises 100 parole hearing transcripts that we requested from the California Department of Corrections and Rehabilitation (CDCR)[1]. We employ the anonymization model of Itani et al. (2024) to remove any instances of names, locations and age-related information to ensure no personal details of any individual involved in the parole hearings is leaked.

For PSH v1.0 we select two female and two male presiding commissioners with 25 transcripts per commissioner. The PDF files range between 37 and 162 pages (8,171 pages in total) and contain the verbatim transcripts of the hearing. The first section of the transcript contains all content said during the interview of the parole hearing. Altogether, this section amounts to 1,297,488 words in PSH v1.0 (excluding punctuation and numbers), with a range of 4,141 to 29,278 words per transcript.

The second section of each transcript contains the decision statements. As we are only interested in the statement provided by the presiding commissioner, we remove all utterances by other speakers. These include mainly interruptions by inmates, translations and supplementary remarks made by the deputy commissioners. The human presiding commissioner statements are between 890 and 4,049 words long and have not been shown to the LLMs tested in this study. We use those decision statements to first distill the rhetorical fingerprint of each commissioner and then to compare the original statements with the LLM-generated statements.

4 Rhetorical Fingerprints

4.1 The dimensions

To assess the relevant rhetorical characteristics of human decision statements, we conduct a manual analysis of 20 decision statements to identify key rhetorical features that are across all four presiding commissioners. This set of linguistic features represents the collective speech style of the presiding commissioners overall, as well as their individual

[1] https://www.cdcr.ca.gov/bph/psh-transcript/

speech style. Deriving both the collective and the individual fingerprint allows us to (1) create an individual linguistic profile to incorporate into the prompt and (2) to conduct a systematic comparison of authentic commissioner statements and the LLM-generated counterparts. The following features are taken into account:

Sentence complexity This feature gives a measurement of the syntactic complexity employed by the presiding commissioners when formulating the sentences. The score is calculated by counting the number of clausal modifiers, conjuncts, adverbial clauses, clausal complements, clausal subjects, and parataxes in each sentence, based on the dependency tag given to each token by SpaCy (Honnibal and Montani, 2017). We then average the complexity over all sentences.

Lexical diversity To assess the lexical diversity, e.g. how much variety and complexity there is in the statements, we use the measure of textual lexical diversity (MTLD) (McCarthy and Jarvis, 2010). For implementation we use the module provided by Shen (2022). Unlike Type-Token-Ratio (Chotlos, 1944), MTLD is length-independent and measures how many words are needed before the Type-Token-Ratio falls below a predefined threshold. Due to the difference in text length between original and AI-generated commissioner statements, we use MTLD for reasons of comparability. An MTLD score is calculated for each of the 25 decision statements and then averaged, resulting in an overall measure per commissioner.

Discourse markers We expect a coherent line of discourse and argumentation in a legal context such as parole hearings. Discourse markers such as *because*, *therefore*, and *however* help to link evidence and conclusions and contribute to the perception of fairness and transparency. We measure the construction of reasoned decisions by counting the occurrence of discourse markers listed in the PDTB resource (Prasad et al., 2008). For aggregating the information, we divide the number of discourse markers across all 25 decision statements by the total number of words spoken by each commissioner.

Nominalizations Nominalizations are known to abstract the responsibility and obscure agency (Fairclough, 2001). They are therefore attributed to an authoritative and bureaucratic tone. Although they are usually attributed to formal written language (Siskou et al., 2022), the manual analysis of the transcripts suggests that they are also relevant in the current context. We estimate the preference for nominalizations by counting nouns ending on *-tion*, *-ment*, *-ance*, etc. across all 25 decision statements and dividing them by the total wordcount per commissioner.

Modals Modal verbs like *must*, *should*, *could* encode power (Fairclough, 2001) and are often used by commissioners to frame parole decisions. Depending on the type and frequency of usage they may convey obligation and institutional authority or empathy. Building on the wordlist for modality used by Herbold et al. (2023), we added a few more modal verbs and adverbs to evaluate the degree of assertiveness in the commissioner's speech style. Modals are aggregated in the same way as nominalizations.

Pronoun usage Pronouns are an important linguistic feature, signaling how the presiding commissioners relate to the inmates and their role in the hearing. We distinguish two dimensions: First, the addressing of the inmate either with *you* versus the reference with *he* (there are no female prisoners in PSH v1.0), latter signaling are more distanced tone. Second, pronouns used when the presiding commissioners refer to themselves (e.g., the more personal *I*) versus a more collective reference (e.g., *we*). Each pronoun version is aggregated in the same way as nominalizations and modals are.

Jargon In institutional settings, such as parole hearings, legal jargon conveys authority, but does also exclude and confuse people who are not familiar with the domain. To compile a domain-specific wordlist of legal terms that are common in the context of parole hearings, we extracted all nouns in the corpus that occurred in at least 3 different decision statements. We then manually went through this list and selected only abbreviations and parole hearing specific and crime related terms. We proceeded in the same way for bigrams and trigrams of consecutive nouns. After the statement generation, we repeated this process to expand the list of jargon used by the LLMs. The final list mainly consists of abbreviations for rehabilitation programs, statutory references, as well as references to forms that inmates can request to file. We normalize jargon usage by dividing the frequency count by the total number of spoken words.

4.2 Initial findings

Overall, we attribute high sentence complexity, high lexical diversity, as well as a frequent use of nominalizations, modals, jargon, indirect references to the inmate (by using third person singular pronouns), and collective self-reference (first person plural) to an authoritative tone. Addressing the inmate directly and framing the decision in the first person singular are considered as empathetic language. The use of discourse markers indicates a reasoned judgment.

Commissioners might choose to alternate between direct and distanced references to the inmate. In our dataset, we see a frequent use of second-person singular pronouns (*you, your*), to directly address the inmate either throughout the entire decision statement or only when providing recommendations to the inmate directly for future hearings. However, some commissioners completely avoid direct engagement with the inmate. In this case, addressing the inmate by using third-person singular pronouns (*he, she, the inmate*) establishes a more distanced tone and enhances the power distance between commissioners and the inmate.

Similarly, depending on the rhetorical intent of the statement, they choose between pronouns that frame the decision as a collective agreement between commissioners or those that signal the commissioner's personal opinion. To emphasize the collective nature of the board's decision, most commissioners use first-person plural pronouns (*we, our*). Through point-of-view distancing ((Brown and Levinson, 1987, p. 204-206); (Locher, 2004, p. 130)) the speaker puts the focus on the idea that the decision resulted of panel deliberation and therefore can distance themselves from individual responsibility. Phrases like e.g. *"Subsequent growth [...] and increased maturity, um, while incarcerated, as we've reflected on this, we didn't find much."* highlight the institutional nature of the decision and signal unity of the commissioners in decision-making. Some commissioners might use first-person singular (*I, my*) pronouns when announcing the decision. As this breaks from the institutional neutrality, it is rather unusual. When it does appear, it typically is used to signal strong personal conviction, as in *"You had a rule violation, a pattern, um, back in and, uh, as I looked through your history [...]"*.

4.3 Standardization and visualization

To normalize the linguistic features, we first calculate their relative frequency for each presiding commissioner across all of their 25 original decision statements. Frequency counts of each feature are aggregated and divided by the total word count per commissioner.

As the features selected for the linguistic evaluation do not share the same distribution, we standardize the normalized feature values with z-scores for better comparability using the `pandas` library. This allows us to observe differences in language use and in units of standard deviations from the mean use across all commissioners, giving insight into how individual presiding commissioners differ from the group norm in their feature use. Figure 1 shows an example of a fingerprint visualization.

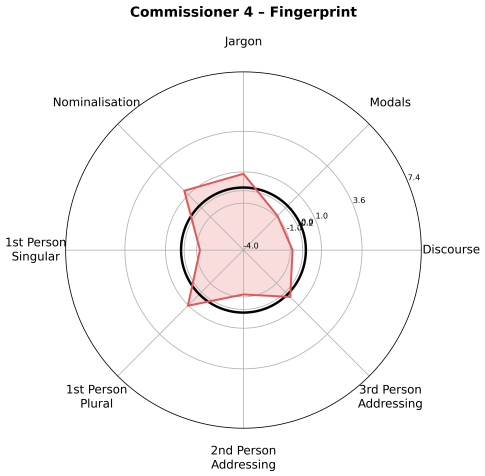

Figure 1: Rhetorical fingerprint across eight dimensions. Lexical Diversity and Sentence Complexity were excluded from this visualization.

The axes of the radarplots represent the rhetorical features. The grey lines indicate the z-score scale. The thick grid line indicates a z-score of 0. Values below or above the thick grid line indicate a lower or higher usage of this particular feature compared to the average usage across all presiding commissioners. For instance, according to Figure 1, commissioner 4 uses more nominalizations, first person plural and jargon than their colleagues, with a z-score of 1 (or higher) – indicating that their jargon usage is at least one standard deviation above the group average. In contrast, they are using less modals, discourse markers and first person singular than their colleagues. Although not included in the Figure above, the z-scores from Table 1 show that compared to their colleagues, commissioner 4

prefers statements with a higher lexical diversity as well as more complex sentences.

This information is used for an enhanced prompting of the models with a speaker-specific rhetorical fingerprint, turning the z-scores into natural language text. The details are discussed in the following section.

5 Prompt engineering

5.1 Assembling the system prompt

In the system prompt, we instruct the models to impersonate an experienced presiding commissioner in a Californian parole hearing and to generate a decision statement about whether to grant or deny parole to the inmate. The decision must be based on the information given in the transcript (provided in the user prompt) and on California state laws and policies. A description of the parole process that is publicly available on the official website of the Board of Parole Hearings[2] and that explains the general factors that need to be considered to assess the risk of reintegration into society (c.f., Section 3.1) is also included in the system prompt. To ensure realistic output, we instruct the models to deliver a spoken statement. We also emphasize the importance of professionalism and factual grounding to prevent the LLMs of inventing details to the case. We explicitly prohibit headings and bullet points. The exact system prompt can be found in Appendix A.1.

5.2 Assembling the user prompt

While the system prompt provides the more general information about how we expect the LLMs to behave as presiding commissioners, the user prompt provides more detail on the linguistic characteristics expected in the outputs as well as some structural guidance (e.g., by providing introductory phrases for the decision statements and instructions on what to discuss in the statement itself).

The rhetorical fingerprint is assembled in a building-block manner with static feature descriptions and commissioner-specific prompt sections. First, we define the usage categories. Previous studies (Sun et al., 2023) show that LLMs tend to underperform when prompted to use specific features with a hard-restricted frequency. We therefore turn the z-scores from the rhetorical fingerprint into natural language sentences and provide those in the user prompt. The z-scores are converted into four categories, namely 'strong', 'frequent', 'rare', and 'avoided' feature usage by the following heuristics:

- **Strong usage**: if $z_f > 1$
- **Frequent usage**: if $0 < z_f \leq 1$
- **Rare usage**: if $-1 < z_f \leq 0$
- **Avoided usage**: if $z_f \leq -1$

Second, we add static explanations for each linguistic feature in the fingerprint and add a usage instruction based on the previously distilled fingerprints for each commissioner. The user prompt additionally provides the transcript and placeholders for metadata concerning age and gender of the inmate for each case. We also provide two typical opening lines that we take from the original transcripts and instruct the LLMs to not use section headings or bullet points. To avoid hallucinations, we include a section that demands the models to only rely on facts given in the transcript. In the end we arrive at a commissioner-specific impersonation prompt, an example of which can be found in Appendix A.2.

To test whether these precise linguistic instructions improve the rhetorical alignment, we mirror the fingerprint prompt with a simplified version of the user prompt, including only the general information about parole hearings. Under this condition, we remove the information about the speaker specific fingerprints and task the LLMs to mirror the language style of the presiding commissioner by drawing on the language patterns in the transcript without any guidance on linguistic features and style. We refer to this condition as primed-by-corpus.

5.3 Prompting parameters

During the prompt engineering phase, we test the performance of all LLMs on a transcript that is not included in the final corpus. We tested the performance of user and system prompts multiple times in an iterative way. Swapping information between the system and the user prompt did not result in any notable difference in response quality. The temperature is set to 0.3 for stylistic consistency after testing for multiple other temperature settings. The three state-of-the-art models, namely GPT-4o, GPT-4.1 and DeepSeek R1, were accessed and prompted via their respective API by using the same system

[2] https://www.cdcr.ca.gov/victim-services/parole-process/

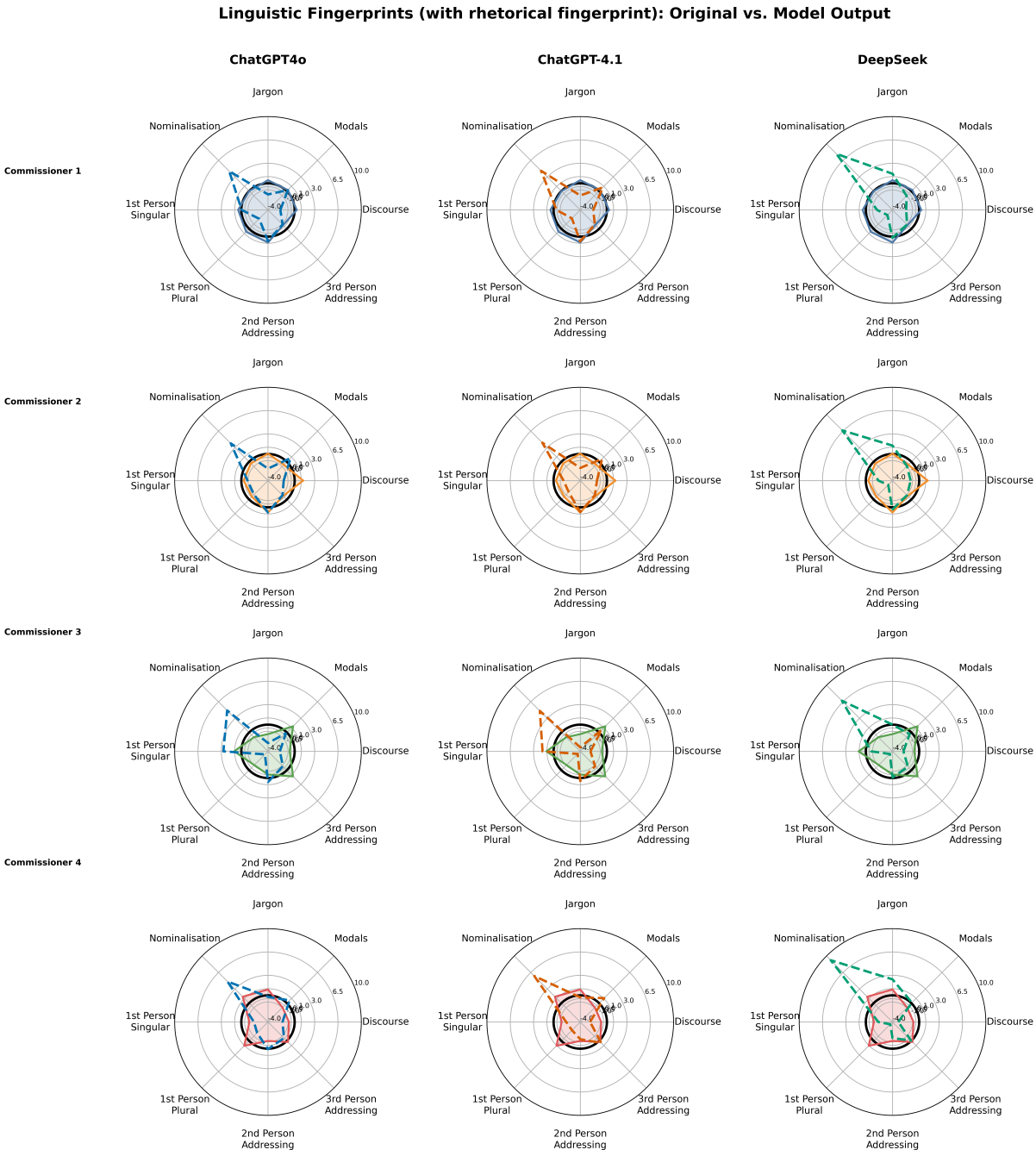

Figure 2: Comparison of rhetorical fingerprints when prompted with a commissioner's rhetorical fingerprint. Original vs. Generated statements. Solid lines indicate original commissioner fingerprints. Dotted lines indicate the fingerprint of the respective models. The thick grid line indicates a z-score of 0.

and user prompts, with and without commissioner-specific rhetorical fingerprints. Despite setting the output token parameter to the highest possible for each model, we observe that all three models give responses that are way below their maximum token output limit.

6 Results

The comparison in this paper is two-fold: First, we identify the rhetorical differences that hold between human and generated, impersonated content. Second, we investigate whether an enhanced prompt with a rhetorical fingerprint yields responses with a higher level of impersonation than a 'plain' prompt with general instructions.

Regarding the first question, we compare the rhetorical approximation of the generated statements with the original commissioners' rhetorical patterns. To this end we calculate the z-scores for each of the selected linguistic features between

Comm.	Original				ChatGPT-4o								ChatGPT-4.1								DeepSeek							
					byCorpus				fingerprint				byCorpus				fingerprint				byCorpus				fingerprint			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Lexical Diversity	-0.55	-0.32	-0.61	1.48	16.92	17.71	16.21	16.39	21.32	22.80	21.25	21.88	5.82	6.06	6.29	5.76	16.57	17.36	16.56	15.67	37.80	38.35	40.20	38.68	40.49	40.28	39.52	36.46
Sentence Complexity	0.04	0.16	-1.31	1.10	25.32	27.67	26.35	27.61	30.02	31.76	33.30	33.66	17.52	19.02	18.24	16.52	23.25	27.20	25.26	27.13	28.23	30.18	28.64	29.96	32.09	30.48	31.19	34.91
Discourse Markers	0.31	1.27	-0.73	-0.85	-2.23	-2.29	-2.09	-2.35	-2.14	-1.65	-2.21	-1.83	-0.93	-0.66	-1.00	-0.73	-2.05	-1.49	-2.47	-2.52	-1.95	-1.62	-2.72	-2.16	-1.95	-1.28	-2.36	-3.13
Modals	0.22	-0.62	1.31	-0.90	-0.18	0.28	-0.09	-0.30	0.22	0.78	-0.20	0.57	-0.83	-0.44	-0.41	-1.10	0.77	0.61	0.56	1.07	-1.06	-1.03	-0.23	-1.20	-1.01	-0.92	-0.36	0.12
Jargon	0.46	0.06	-1.41	0.88	-2.26	-2.75	-2.50	-2.03	-1.68	-2.19	-2.76	-0.21	0.49	-0.62	-1.33	0.27	-1.83	-2.10	-3.41	-0.30	3.16	2.66	1.01	3.18	1.43	1.26	0.01	2.40
Nominalization	-0.008	-0.35	-1.00	1.36	3.90	3.61	4.11	4.20	4.24	3.91	4.66	4.59	3.28	3.37	4.03	3.43	4.30	4.04	4.57	5.75	7.42	6.73	7.03	6.47	7.90	6.64	6.73	9.10
3rd Person Addressing	-0.79	-0.76	1.31	0.24	-0.91	-0.92	-0.91	-0.91	-0.91	-0.91	-0.84	-0.60	-0.90	-0.91	-0.83	-0.88	-0.90	-0.90	-0.85	0.45	-0.91	-0.89	-0.88	-0.83	-0.91	-0.91	-0.72	-0.05
2nd Person Addressing	0.88	0.78	-0.50	-1.15	0.77	0.83	0.57	0.71	0.65	0.77	0.61	0.07	1.22	1.25	0.97	1.12	0.70	0.74	0.44	-1.43	0.22	0.34	0.21	0.04	0.18	0.32	-0.08	-1.60
1st Person Plural	0.63	-0.58	-1.08	1.04	-1.37	-1.13	-1.54	-1.36	-2.13	-1.03	-3.35	-1.69	-2.74	-2.75	-2.73	-2.60	-2.22	-1.56	-3.45	-2.25	-3.08	-3.21	-3.16	-3.01	-2.92	-3.09	-3.44	-3.52
1st Person Singular	0.48	-0.35	1.08	-1.21	-1.48	-1.47	-1.50	-1.47	-0.05	-0.94	2.69	-1.90	-0.85	-1.01	-0.61	-0.94	-0.43	-1.50	1.66	-1.90	-1.81	-1.81	-1.76	-1.79	-1.68	-1.90	-0.79	-1.87

Table 1: Comparison of z-scores for all features across original and LLM-generated outputs

original and generated statement by first normalizing the frequency count for each feature in the generated statements. To calculate the z-scores we use the mean and standard deviations of each feature calculated from the original commissioner statements. Using the original mean and standard deviation metrics establishes the baseline against which the generated decisions statements are compared. The resulting z-scores show to which degree the generated statements deviate from the original statements: Positive z-scores indicate a stronger usage compared to the commissioner average, negative values reflect underuse (in terms of standard deviations). An overview of the performance of each LLM for the primed-by-corpus and rhetorical fingerprint scenario can be found in Table 1.

Figure 2 shows the resulting radar plots of the rhetorical fingerprint prompts in comparison to the original fingerprint visualization[3]. The axes of the radarplots represent the individual features. The grey lines indicate the z-score scale. The thick grid line indicates a z-score of 0. Columns represent LLMs, while lines represent the individual commissioners. Each LLM and commissioner is color-coded. Solid lines represent the scores in the rhetorical fingerprint of the original commissioners, while the dotted lines show the rhetorical approximation of the generation models.

In the following we discuss the dimensions in the fingerprint in terms of how the models deviate rhetorically from the original commissioners rhetorical patterns.

6.1 Lexical diversity and sentence complexity

From a procedural point of view, high lexical diversity and/or sentence complexity makes parole hearing decision statements difficult to understand, going against the guideline that the hearings should be accessible and easy to understand by the inmates. The original decision statements exhibit a relatively stable usage of minimal lexical diversity and simple sentence structures, indicating that commissioners are mindful about making their statements accessible. In our dataset, Commissioner 4 is the only one showing an elevated z-score for lexical diversity (1.48) and sentence complexity (1.10).

The analysis of the LLM-generated statements shows that all three LLMs highly deviate in lexical diversity and sentence complexity compared to the original decision statements (see Table 1), across impersonated commissioners and conditions in the user and system prompt. GPT-4.1 shows the lowest z-score for lexical diversity (5.76) and sentence complexity (16.52) for Commissioner 4 in the primed-by-corpus condition. All other models exceed these z-scores substantially (z-scores range from 5.76 to 40.20 for Lexical Diversity and 16.52 to 34.91 for sentence complexity). DeepSeek demonstrates the highest z-scores for both features, indicating that even elaborate prompting does not help to mitigate this behavior. Taken together, this indicates that LLMs are insensitive to prompts when it comes to aligning spoken content in terms of its lexical diversity and sentence complexity. This is probably due to the underlying training data being mostly written language.

Due to the emerged non-alignment in terms of lexical diversity and sentence complexity, we exclude both dimensions from the radar plots in Figure 2 to prevent those features from skewing the plots.

[3]The visualizations for the primed-by-corpus condition can be found in Appendix B.

6.2 Nominalizations and jargon

Nominalizations and a frequent use of domain-specific jargon are attributed to written communication and make the content of the statement inaccessible to individuals who are not familiar with legal language. In the original statements we observe a variety of jargon and nominalization preference patterns. What we observe consistently is that commissioners who are using more jargon also use more nominalizations than their colleagues and vice versa. The plots in Figure 2 indicate a consistent underuse of jargon in both GPT models prompted with rhetorical fingerprints, unless they are prompted to use jargon strongly (Commissioner 4). When prompted with the more general prompt without the rhetorical fingerprint, GPT-4o continues to underperform, while GPT-4.1 seems to approximate the linguistic behavior of the original commissioner and thus infers the degree of usage of this feature. DeepSeek consistently overuses jargon.

Additionally, all three models show a strong preference for using nominalizations, suggesting a strong bias towards formal and written language, likely due to their training data. A similar observation has been made by McGovern et al. (2025), who show that LLMs exhibit a high usage of nouns in their responses. This behavior cannot be mitigated by prompting and holds across all models and conditions. Prompting for strong usage of nominalizations even triggers the models to use more nominalizations than they already do (see DeepSeek for Commissioner 4, where z-score was 6.47 for primed-by-corpus and 9.1 for fingerprint condition).

6.3 Modal verbs and discourse markers

Modal verbs and discourse markers are important features for parole decision statements as they help to convey authority and coherence by expressing obligations and transparency about the reasoned judgment. The commissioners in our dataset either use modal verbs frequently or rarely. The same applies for discourse markers, but Commissioner 2 is the only one showing a strong preference to use them. The analysis of the generated statements shows that all models tend to underuse modal verbs. Only GPT-4o fully replicates the behavior of Commissioner 1 when prompted with the rhetorical fingerprint instruction to frequently use modal verbs.

Discourse markers are consistently underused by all models and all commissioners across prompting conditions. This indicates that LLMs show limited sensitivity to these features. We attribute this to the fact that LLMs might interpret discourse markers as filler words which can be dropped without affecting the semantic structure of the generated text. Overall, we can conclude that even when explicitly prompted with specific usage instructions all LLMs show limitations in their ability to replicate reasoning structures and modality for most cases.

6.4 Pronouns

Addressing the inmate in the third person singular ('he', 'she',) even if they are present in the same room conveys authority and manifests power. We only see a preference for third person singular addresses with Commissioners 3 and 4, whereas Commissioners 1 and 2 prefer to address the inmates directly by using the more personal 'you'. By looking at the radar plots for all commissioners in Figure 2, we see a pattern of preferring second person singular addresses across all models. Prompting them for strong indirect inmate addressing does not yield the expected results (see model performance for Commissioner 4). GPT-4.1 and DeepSeek follow this instruction to a very small extent, but only if prompted for strong usage. We attribute this behavior to our prompt context which explicitly asks for conversational tone. The underlying dialogue training data for each model is very likely coming from written online communications (e.g. Reddit), where indirect addresses are uncommon. LLMs may therefore infer that they are speaking with the inmate, instead of about them.

Regarding the use of pronouns when referencing either themselves as individuals or as a collective, we observe that all models underuse the first person plural. We suspect here that our prompts are being misinterpreted by the models, which are unaware of the fact that the presiding commissioner is the representative of the parole hearing panel. They therefore default to the individual self-referencing pronoun 'I' (which is also more likely to be overrepresented in their training data).

6.5 LLM-specific fingerprint

All models exhibit their own model-specific linguistic characteristics, independent of the prompt. This suggests that some features are inherent to the model's own rhetorical style and are therefore not adjustable by prompting at all or only to a small degree. This is particularly evident when comparing

the radar plots of GPT-4o and GPT-4.1 under the rhetorical fingerprint condition: The overall shape of the fingerprint remains nearly identical across these two model versions when prompted with the speaker-specific rhetorical fingerprints, suggesting that linguistic characteristics are inherited across model versions. For the DeepSeek model we barely see any change in linguistic behavior between the two prompting conditions. Only when prompted for strong usage of nominalizations we see a minor adjustment in feature intensity. Nevertheless, the instruction of using first person plural pronouns gets ignored completely, which reinforces our suggestion of model-specific rhetorical fingerprints.

7 Conclusion

In this study we test whether state-of-the-art LLMs can be pushed to impersonate humans when prompted with linguistically informed fingerprints. Our findings from testing three off-the-shelf models in generating parole hearing decision statements, a high-stake setting, underscore the current rhetorical limits of LLMs in mirroring human-like behavior as they fail to deviate from their own model-specific rhetorical fingerprint.

The next step in this endeavor is a more detailed investigation of the effect of rhetorically enhanced prompting and the outcome of these hearings (whether parole is granted or not and under which conditions). A significant amount of more detailed analyses is required in order to show the limits of applying LLMs in sensitive and high-stake settings like the present one.

Limitations

Feature selection. In this study we only consider a limited number of linguistic features, which does not cover the full complexity of human rhetorical characteristics. We restricted the prompts to stylistic characteristics, which may oversimplify human language.

Domain. We test the ability of three LLMs on a very specific domain. Future work should look into the performance of LLMs when mimicking human-like speech in different domains.

Model selection. At the time of writing, we covered the three most popular LLMs, all of which do not openly disclose their training data. Therefore, we do not know whether parole hearing transcripts were included in the model's training data. More recent models may differ in performance.

Ethics Statement

While this paper focuses on the linguistic capabilities of LLMs to reproduce a certain rhetorical fingerprint, we are aware of the potential risks associated with generating human-like institutional dialogue. This study intends to assess the stylistic approximation capabilities of LLMs within a controlled research setting. It is not intended to support or encourage the use of LLMs for deceptive or harmful applications, especially in legal settings. We also do not propose to actually use them in parole hearing evaluations. For reasons of data protection, we will not publish the original, nor the anonymized transcripts. However, we will provide a list of the individual hearings upon request, making it possible to interested researchers to request the exact same files from the CDCR for replication purposes.

Acknowledgments

The work reported on in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379 as part of the project "Inequality in Street-level Bureaucracy: Linguistic Analysis of Public Service Encounters".

References

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. 2024. Emulating author style: A feasibility study of prompt-enabled text stylization with off-the-shelf LLMs. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 76–82, St. Julians, Malta. Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

John W. Chotlos. 1944. IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75–111. Place: US Publisher: American Psychological Association.

Anca Dinu, Andra-Maria Florescu, and Liviu Dinu. 2025. Analyzing large language models' pastiche ability: a case study on a 20th century Romanian author. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 20–32, Albuquerque, USA. Association for Computational Linguistics.

Norman Fairclough. 2001. *Language and Power*. Language in social life series. Longman.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1):18617.

Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. 2024. Large language models can impersonate politicians and other public figures.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Abed Itani, Wassiliki Siskou, and Annette Hautli-Janisz. 2024. Automated anonymization of parole hearing transcripts. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 115–128, Miami, FL, USA. Association for Computational Linguistics.

Miriam A. Locher. 2004. *Power and Politeness in Action*. De Gruyter Mouton, Berlin, New York.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2025. Your large language models are leaving fingerprints. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 85–95, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS ONE*, 19(3):1–25.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Wassiliki Siskou, Laurin Friedrich, Steffen Eckhard, Ingrid Espinoza, and Annette Hautli-Janisz. 2022. Measuring plain language in public service encounters. In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022) Potsdam, Germany*.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

A Prompts

A.1 System prompt

This is the general system prompt that we used for all 4 commissioners under all conditions:

```
   You are an experienced Parole Commissioner from
the Board of Parole Hearings in California, deciding
whether to grant or deny parole to inmates.  You
need to make informed parole decisions based on
California state laws and policies, including
the guidelines of the California Board of Parole
Hearings.
Parole proceedings are not to decide guilt or
innocence.  The Board of Parole Hearings accepts as
fact the guilty verdict imposed by the courts.  The
purpose of a parole proceeding is to determine if or
when an inmate can be returned to society.  Under
normal circumstances, the panel or the Board shall
set a release date unless it determines that the
gravity of the crime (offense), or the timing and
gravity of current or past convictions, requires
a more lengthy period of incarceration to ensure
public safety.

In general, some of the factors considered by
the panel and which are discussed in the proceeding
include:
    • counseling reports and psychological
      evaluations
    • behavior in prison (i.e., disciplinary notices
      or laudatory accomplishments)
    • vocational and educational accomplishments in
      prison
    • involvement in self-help therapy programs that
      can range from anti-addiction programs for
      drugs and alcohol to anger management
    • parole plans, including where an inmate would
      live and support themselves if they were
      released
   After reading the transcript provided by the
user, your task is to decide whether the inmate
should be granted or denied parole and deliver your
decision as spoken dialogue, mirroring a natural,
ongoing conversation in the hearing room.  Do not
use headings or bullet points in your statement.

Remain professional and consistent with the tone
and format expected of official parole decision
statements.  However, because you will be delivering
this decision as spoken dialogue, adapt the
formality to reflect a real parole hearing's
conversational flow.  You are not allowed to use
bullet points or headings in your statements.  You
must provide detailed, professional, fair, and
well-reasoned responses.  Avoid bias, stereotypes,
prejudice, or speculation.  Refer only to the facts
```

and background details included in the transcript. If some details are missing, acknowledge them rather than inventing information.

A.2 User prompt

This is an example of a commissioner prompt (commissioner 4). Passages written in blue were only included under the rhetorical fingerprint condition. Passages written in violet were only included under the primed-by-corpus condition. Passages written in black appear in both conditions.

You will read a transcript of a Californian parole hearing and act as the presiding commissioner.

After reading the transcript, your task is to decide whether the inmate should be granted or denied parole and then deliver your decision as spoken dialogue, mirroring a natural, ongoing conversation in the hearing room.

These are the overall style requirements:

- Conversational tone: avoid enumerations, headings, or overly formal written structures. Instead, formulate your response as if it were spoken in a parole hearing. You may use occasional pauses and conversational transitions to make it flow naturally.
- Commissioner style: You are speaking as the presiding commissioner. Please adapt your response to reflect the commissioner's typical language style, including tone, sentence structure, level of formality, and use of hesitation markers. Your statement should feel authentic to a commissioner's usual way of delivering decisions.

Do not label sections in your final text, but address these points in a conversational and detailed manner:

- Introduction: Set the context of the hearing. You can use these opening lines to do so: "Today's date is [MONTH] [DAY], [YEAR]. The time is approximately [TIME] AM. All parties who were present before have returned." or "Today's date is [MONTH] [DAY], [YEAR]. The time is approximately [TIME] PM. We're back in the matter of Mr. ..."
- Decision: Clearly state whether the inmate is granted or denied parole.
- Evaluation and Reasoning: Discuss both aggravating and mitigating factors that influence your decision.
- Recommendations (if parole is denied): Specify the denial length and explain the reasons for setting that length. You can set 3, 5, 7, 10 or 15 years of length, depending on the severity of each case. Offer detailed suggestions for what the inmate could do to improve the likelihood of a positive outcome at a future parole hearing (e.g. additional programming, self-improvement efforts, insight development).
- Clarify (only if parole is granted): Clarify that this decision is not final and will be subject to further review by the Governor. Explain that the inmate will be formally notified in writing once a final decision is made.

After reading the transcript, your task is to decide whether the inmate should be granted or denied parole and deliver your decision as spoken dialogue, mirroring a natural, ongoing conversation in the hearing room. Do not use headings or bullet points in your statement.

Below are the key linguistic features you may use, along with usage instructions. Each feature includes a Usage Category that can be set to any of the following:

- avoid: Do not use this feature.
- rarely: Use this feature only a few times.
- frequently: Use this feature regularly, but do not overuse it.
- strong: Use this feature a lot.

In your spoken statements, you are required to use the following linguistic features with the indicated frequency: In your spoken statements, you are required to use the following linguistic features with the indicated frequency:

- Lexical Diversity to express nuanced viewpoints and considerations. Use a wide-ranging vocabulary by using synonyms and varied expressions throughout your statements. This corresponds to the usage category "strong".
- Sentence complexity: Use a lot of complex and long sentences. This corresponds to the category "strong".
- Discourse markers (e.g., "because", "however", "while") to indicate causal reasoning, contrasts, or transitions. Use these words rarely.
- Modals: Words like "could", "should", "would", "may", "might" are modal verbs and are used to convey obligations or possibilities. Use these words rarely.
- Nominalizations: Nominalizations are verbs that are turned into nouns, like e.g. "the denial", "the recommendation", "the rehabilitation". This is a strong feature. Use nominalizations very often.
- Jargon: Strongly use legal terms legal terms like "recidivism", "suitability", "mitigating factors" and other technical terms that are typically used in the context of parole hearings.
- First-Person Singular: Avoid using first-person singular pronouns in phrases like "I reviewed", "I find" to refer to the presiding commissioner's decision.
- First-Person Plural: Strongly use phrases like "we reviewed", "we find" to refer to the panel's collective voice.
- Second-Person Singular References: Avoid to directly address the inmate by using second-person singular pronouns.
- Third-person singular when referencing the inmate in a detached or formal sense (e.g., "he is not suitable for parole", "the inmate has demonstrated insight"). Frequently refer to the inmate by using third-person singular pronouns to address them in a more detached way.

At the very end, include one of the following lines:

- Decision: granted
- Decision: denied

If you deny parole, also add:

- Denial length: X years You can set 3, 5, 7, 10 or 15 years of denial, depending on the severity of each case.

Use the following details:

- Inmate ID: inmate_id
- Gender: gender
- Current Age: current_age
- Age of Imprisonment: age_of_imprisonment

Main Part of the Hearing Transcript:

- {transcript}

Base your decision solely on the facts provided. Write your response as one continuous speech, providing detailed reasoning. Your statement must be very long and detailed.

B Fingerprints primed-by-corpus condition

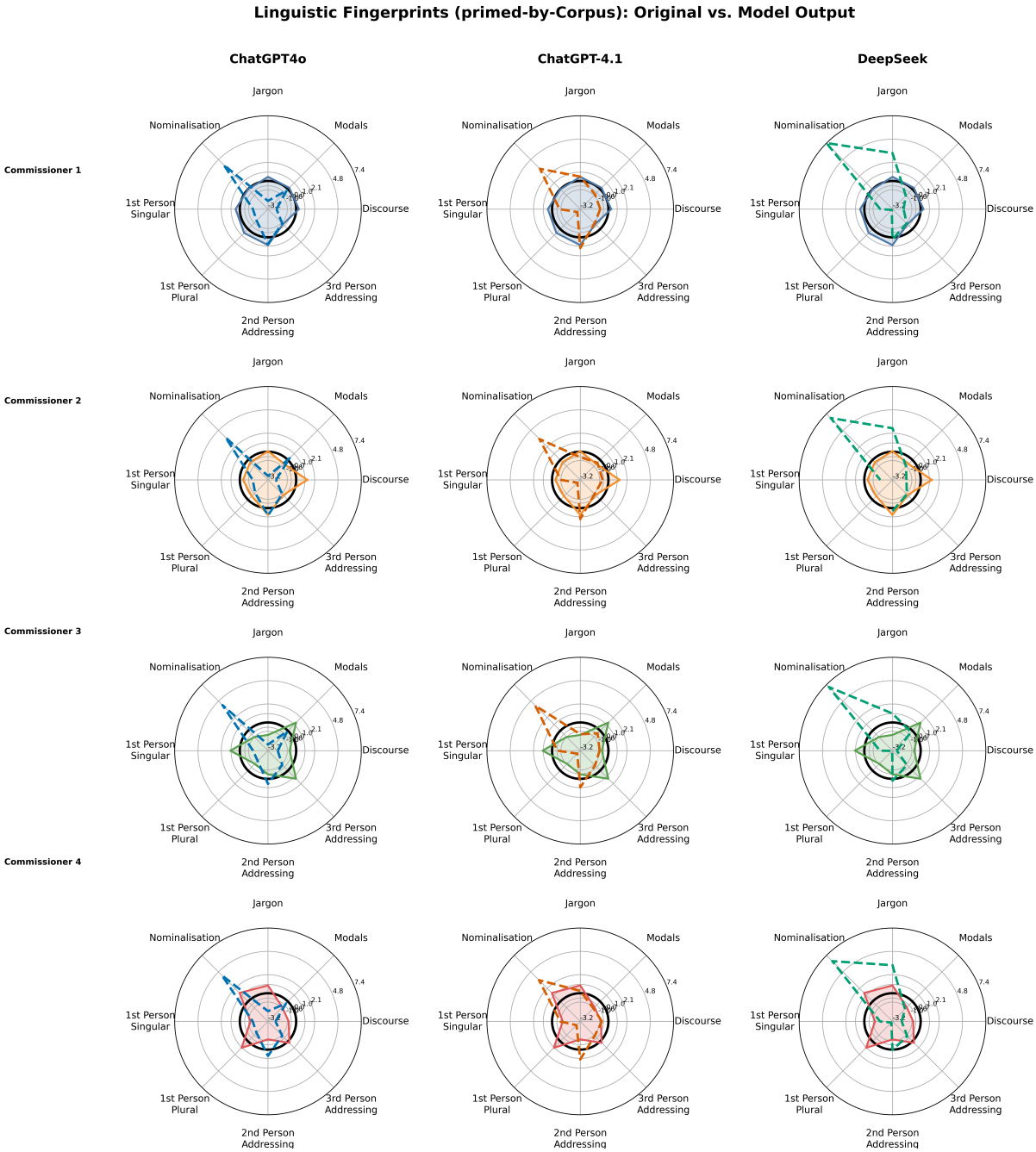

Figure 3: Comparison of linguistic fingerprints when primed-by-corpus. Original vs. Generated statements. Solid lines indicate original commissioner fingerprints. Dotted lines indicate the fingerprint of the respective models. The thick grid line indicates a z-score of 0.