”So, are you a different person today?” Analyzing Bias in Questions during Parole Hearings

Wassiliki Siskou[1,2] and **Ingrid Espinoza**[1]
[1]Cluster of Excellence ”The Politics of Inequality”, University of Konstanz
[2]University of Passau
firstname.lastname@uni-konstanz.de

Abstract

During Parole Suitability Hearings commissioners need to evaluate whether an inmate's risk of reoffending has decreased sufficiently to justify their release from prison before completing their full sentence. The conversation between the commissioners and the inmate is the key element of such hearings and is largely driven by question-and-answer patterns which can be influenced by the commissioner's questioning behavior. To our knowledge, no previous study has investigated the relationship between the types of questions asked during parole hearings and potentially biased outcomes. We address this gap by analysing commissioner's questioning behavior during Californian parole hearings. We test ChatGPT-4o's capability of annotating questions automatically and achieve a high F1-score of 0.91 without prior training. By analysing all questions posed directly by commissioners to inmates, we tested for potential biases in question types across multiple demographic variables. The results show minimal bias in questioning behavior toward inmates asking for parole.

1 Introduction

During Parole Suitability Hearings commissioners need to weigh different factors to evaluate whether an inmate's risk of reoffending has decreased sufficiently and therefore justify their release from prison before completing their full sentence. The interaction between the commissioners and the inmates is the key element of such hearings, which are largely driven by question-and-answer patterns. Questions asked by commissioners to inmates, in particular, shape the entire conversation and guide the inmate's responses, demonstrating that these questions are far more than a mere formality. This is because different types of questions open spaces for different types of answers. While open questions give room for elaboration and justification, closed questions limit the answer space to either one of the alternatives given or ”yes” or ”no” in the case of polar questions. Given this, we expect a positive relation between positive parole outcomes and the share of open questions in the hearings analysed.

While biased outcomes in parole hearings have been studied by various disciplines, mainly focusing on the relationship between demographic variables and parole outcomes (Huebner and Bynum, 2008; Morgan and Smith, 2008; Young et al., 2015; Young and Pearlman, 2022; Hail-Jares, 2019), only a few have looked further into the linguistic particularities of this kind of dialogue (Cochran and Comeau-Kirschner, 2016; Todd et al., 2020). To our knowledge, no previous study has investigated the effect of commissioners' questioning behavior during parole hearings on the reproduction of social inequality, despite their crucial role in the delivery of procedural justice.

This study aims to contribute to this research gap by analyzing the questions posed to inmates during parole hearings from California. First we show that with the help of Large Language Models, we are able to generate linguistically informed annotations automatically, allowing us to identify patterns in questioning style of parole board members, while simultaneously minimizing annotation cost and effort. Drawing upon these annotations, we examine the relationship between question types and demographic variables collected from a sample of 102 parole-seeking inmates.

The key questions of this study are the following: How do the types of questions asked during parole hearings relate to their outcomes? What is the relation between the types of questions posed and an inmate's racial background? Are there any disparities in the share of questions posed related to an inmate's racial background and does this impact their chances of being released on parole? Does the commissioners' gender bias their questioning behavior towards inmates of different races?

Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024), pages 116–128

2 Parole Suitability Hearings in California

The decision about whether an inmate no longer poses a severe risk to the public is based on documented information about the inmate's behavior, rehabilitation progress, and potential for successful reintegration into society. All documents relevant to the process are collected and reviewed by both the Board of Parole Hearings and the inmate before the hearing.

The hearings consist of an hour-long interview between both commissioners and the inmate, and are used to clarify issues that need to be addressed. Moreover, it also allows the commissioners to assess the inmate's credibility and rehabilitation by interrogating the inmate directly. At the end of this interview, the commissioners make a decision on the inmate's parole eligibility. The committee might grant parole, deny it or postpone the decision to a later date. During the parole interview, the inmate is required to answer questions about various aspects of their life. The topics covered are their life prior to the crime that led to their life sentence, their criminal record, the crime itself, as well as their behaviour since incarceration and possible parole plans (such as job opportunities, housing, and community support).

Typically, the presiding commissioner conducts the initial and concluding parts of the hearing, addressing the pre-commitment factors and the decision, while deputy commissioners usually cover the post-commitment factors and parole plans.

Despite the structured nature of the hearing process, there is room for potential biases stemming from the different types of questions asked, consequently limiting the answer space available to the inmate. By analyzing these subtle linguistic strategies, we aim to shed light on this understudied area.

3 Related Work

3.1 (Computational) Linguistic Background

Questions are fundamental constructs in pragmatics and discourse analysis. In parole suitability hearings in particular, they play a crucial role in shaping the dialogue dynamics, as they guide the direction of the interview and influence the flow of information.

Previous research, such as Kalouli et al. (2018) has focused on the pragmatic function of questions. In their study, they adapt linguistically informed heuristics to classify questions into information-seeking and non-information seeking questions on a parallel Bible corpus. Hautli-Janisz et al. (2022a) and Hautli-Janisz et al. (2022b) follow a similar approach and propose a taxonomy to categorize questions into pure, assertive, rhetorical and challenge questioning, depending on their pragmatic function in argumentative dialogue.

In addition, Stivers and Enfield (2010) propose a question-response coding scheme for dyadic and multi-party interactions, based on their logical semantic structure. They categorized questions into three primary types: polar questions, requiring a "yes" or "no" as response, alternative questions, which offer a set of predefined choices as answer, and content questions (wh-questions), which seek more elaborate answers from the respondent. Moreover, this coding scheme opens the possibility to label questions according to their pragmatic function, which they call 'social actions' (such as requests for information, requests for confirmation, etc.). The categorization from Stivers and Enfield (2010) allows further classification based on the depth of responses they elicit into open and closed questions, which suits our research purpose best, since it facilitates the automatic classification and allows us to thoroughly analyse any patterns in questioning that potentially impact parole outcomes.

Research from forensic linguistics (Cochran and Comeau-Kirschner, 2016) investigated the linguistic strategies convicted sex offenders use during their parole hearings, finding differences in language use between those granted parole and those denied. Todd et al. (2020) applied Language Models to detect anomalies in Californian Parole Hearing transcripts, providing a method to review the hearing process.

In the realm of machine learning, approaches for automatic question annotation were mainly used to improve answers given by question answering systems (Bullington et al., 2007). A machine-learning approach that uses fine-grained taxonomy was introduced by Li and Roth (2002) to further categorize open-ended wh-questions by considering the semantic constraints of the expected answer. Recent studies (Gweon and Schonlau, 2023) have focused on the classification of answer types to open-ended questions in surveys using Large Language Models.

With the rise of Generative AI in recent years, researchers from different disciplines have used Large Language Models such as GPT-3.5 and

GPT-4 to generate annotated text data and explore whether the annotations match human judgement. Gilardi et al. (2023) explore if ChatGPT can perform high-quality annotations by only giving the model the coding instructions they would give to human annotators. Compared to crowd workers from MTurk and trained annotators, annotations generated via zero-shot prompting ChatGPT-3.5-turbo were found to have a higher accuracy when labelling tweets and news articles. Similar results were reported in Mens et al. (2023) for measuring semantic similarity with GPT-4, achieving state-of-the art results without requiring training. These results show that AI generated annotations often match or even exceed human judgement and therefore save not only time but also financial resources.

3.2 Background from Legal Studies and Public Administration

Racial disparities in parole hearing decisions have been the subject of numerous previous studies. Findings indicate that even after adjusting for criminal severity and rehabilitation efforts, Black parole candidates had a much lower chance of being granted parole than White candidates (Young and Pearlman, 2022). As a consequence, Black prisoners experience a noticeable lengthier parole delay compared to White prisoners (Huebner and Bynum, 2008). The same finding holds after accounting for various legal and demographic factors.

One of the main explaining factors for this racial inequality is street-level bureaucrat's biased decision-making (Hertogh, 2018; Lotta and Pires, 2019; Raaphorst, 2022). Bureaucrats working at the frontline of public policy implementation, such as parole hearing commissioners, interact with citizens from positions of power while being pulled along by state institutional forces that hold sway over them, cultural renderings of worthiness they carry with them, and collective guidance communicated through the exchange of stories among them (Maynard-Moody and Musheno, 2012). All these factors can influence and bias the way they interact with inmates, as well as the decisions they make upon these interactions. In making a decision for or against the parole suitability of an inmate, parole board members have to heavily weigh factors such as institutional misconduct, educational attainment, the nature of the offense, psychological evaluations, and disciplinary reports. Even though parole board members are bound by a set of strict guidelines in their decision-making, evidence shows that reliance on subjective judgments creates a "decisional scaffolding" that reinforces racial disparities (Huebner and Bynum, 2008; Young et al., 2015; Young and Pearlman, 2022), suggesting that social influence does play a role in how board members shape their decision.

Besides racial disparities, there is also evidence that certain socio-demographic factors such as community context have a negative effect on parole decisions. Huebner and Bynum (2008) found that Hispanics who were living in more disadvantaged neighborhoods had smaller chances of being granted parole by the Board. Young et al. (2015) found that both, older inmates, and inmates who were young at the time of the crime commitment, had an increased likelihood of being released for parole. Both of these findings are in line with guidelines the Board has to bound by, such as the Elderly Parole Program[1] and the Youth Offender Parole Program[2]. Other factors, such as substance abuse programming participation, and low-risk scores in psychological evaluations are positively associated with release (Young et al., 2015), demonstrating key points the Board considers when deciding whether an inmate is ready to be released.

There is little research within the field of policy implementation and public service delivery at the intersection of language and power. Most of it addresses how speaking the language of minority groups can enhance or diminish responsiveness toward citizens (Scheibelhofer et al., 2021; Holzinger, 2020), or how (written) bureaucratic jargon can emerge as a burden for citizens unfamiliar with administrative language (Fisch and Burkhard, 2014). Only a few scholars have recently analyzed interaction in public service encounters using language as an indicator of behavior in the way we do in this paper (Siskou et al., 2022; Espinoza et al., 2024; Eckhard and Friedrich, 2022).

The relationship between gender and parole hearings has so far mainly been studied looking at the effect of an inmate's gender on parole outcomes. Findings suggest differences when it comes down to the type of commitment offenses, prior prison sentences, age at admission to confinement from which paroled, as well as alcohol and drug involvement (Moseley and Gerould, 1975; Silverstein, 2006).

[1] https://www.cdcr.ca.gov/bph/elderly-parole-hearings-overview/

[2] https://www.cdcr.ca.gov/bph/youth-offender-hearings-overview/

In light of these findings, we expect to see racial disparities in the positive relation between positive parole outcomes and open questions. Regarding gender, given our sample, we can only assess if the commissioner's gender biases their questioning behavior (see Section 4.2).

4 Data

4.1 Parole Hearing Transcripts

Transcripts of Parole Suitability Hearings are available to the public and serve as official records of the proceedings, as they include all verbatim statements made during the hearing by parole board members, the inmate, their attorney, and any other individual present (like e.g. district attorney, victims or victim's next of kin). We obtained a total of 283 of such parole hearing transcripts in PDF format, which we officially requested from the California Department of Corrections and Rehabilitation (CDCR)[3]. The hearings used for this analysis took place between August and September 2021 and were conducted via video conference due to the COVID-19 pandemic. Official transcriptions of the video recordings were made by authorized transcribers hired by the U.S. authorities.

4.2 The metadata

Given the lack of metadata, we manually redacted the transcripts and decided to extract a total of 13 different variables per transcript, based on the mentioned previous findings from research on parole hearing outcomes. Inmate related metadata include their gender, race, age, age at the time of the crime, type of life crime (violent, non violent or sexual offense), years served in prison, education background, third-striker status[4], gang affiliation, total number of pages of the transcript, and the outcome of the parole suitability hearing. We additionally extracted the gender of the presiding and deputy commissioner present in the hearing. Due to the fact that not all corresponding information for the selected variables was consistently mentioned during the hearings, some gaps in the dataset are unavoidable.

Our final dataset consists of metadata for 102 parole candidates, comprising 100 males and 2 females. We successfully extracted 21 transcripts with corresponding metadata for white inmates, 20 for Hispanic inmates, 20 for black inmates, and 7 for inmates of other races (mainly Asian). The racial composition of our final sample was intentionally balanced, despite the general overrepresentation of Black and Latino men among prisoners in California. It is important to note that there is no publicly available data about the racial demographics of individual parolees. However, the CDCR has reported minimal disparities in parole grant rates across different racial groups. In 34 transcripts, race was not mentioned, but the majority of the remaining 12 variables could still be extracted from the text data.

According to the official CDCR report 8,717 parole hearings were held in the year of 2021. Out of 4,188 Hearings with an outcome 1,424 inmates were granted parole (34% grant rate), while 2,764 were denied (66% denial rate). The remaining results were either postponed, voluntarily waived, stipulations or cancellations. The majority of parole suitability hearings (97%) were held for male and only 256 for female parole candidates. The statistical report is publicly available[5] [6]. We see the same distribution of grant rate (32 hearings) and denial rate (70 hearings) in our selected sample, which also replicates the gender distribution (100 male vs. 2 female) observed in the official report.

Out of the 102 parole hearings in our dataset, 32 were presided over by a female commissioner and 70 by a male commissioner. For deputy commissioners, 43 were female, and 59 were male.

4.3 The final Corpus

The PDF transcripts of our final dataset, range from 37 to 164 pages, with an average length of 85 pages. The dataset comprises 48,478 thousand utterances, 142,540 thousand sentences and 21,122 questions in total, of which 16,039 were directly asked by commissioners to inmates. While questions constitute only 12% of the total sentences in the corpus, 76% of the questions asked in the corpus are directed to the inmate. The low percentage of questions in the corpus is due to several factors. First, each side is entitled to make lengthy closing state-

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

[4] "Third-striker" refers to an individual convicted under California's Three Strikes Law, which mandates severe penalties for those convicted of three or more serious or violent felonies. Typically, third-strikers face a life sentence and can only be eligible for (not granted) parole after serving a minimum of 25 years in prison.

[5] https://www.cdcr.ca.gov/bph/2021/03/15/calendar-year-2021-suitability-results/

[6] https://www.cdcr.ca.gov/bph/wp-content/uploads/sites/161/2023/05/pv-2021-Significant-Events.pdf

ments. Second, the inmate's answers typically do not include questions. Finally, the commissioners' decision statements are entirely assertive and very long, as they provide detailed justifications for their decisions about granting or denying parole.

We used Python to process the PDF files, converting them to text format and extracting utterances based in speaker tags. Questions were identified by searching for questions marks in utterances attributed to the "Presiding Commissioner" or "Deputy Commissioner", ensuring they were immediately followed by an inmate reply.
In terms of data acquisition, we encountered several challenges resulting in only including 102 transcripts of the 283 initially obtained to the final corpus. The excluded transcripts were omitted for various reasons, including corrupted files that could not be opened, hearings held *in absentia* (where the inmate was not present), or hearings that resulted in a waiver, postponement or stipulation rather than clear parole decisions. Additionally, the process of manually retrieving metadata from the transcripts is a time intensive task, as it requires to thoroughly read through each file to ensure the accurate extraction of relevant information. This labor-intensive approach is the primary reason for the relatively small sample size, but guarantees the reliability of the metadata used in our analysis.

Due to data privacy concerns we will not publish the unanonymized dataset, but can provide a list of the requested transcripts upon demand.

5 Methods

5.1 Question Taxonomy

To investigate question-asking patterns of commissioners towards inmates, we adopted the approach proposed by Stivers and Enfield (2010) and decided to use a taxonomy for question classification that is intentionally under-specifying. Specifically, we examined polar, alternative, and wh-questions, which we distilled into two broader categories: closed-ended and open-ended questions (see Section 3.1).

Open questions, typically referred to as content or wh-questions, are intended to require detailed answers and give the interviewees the freedom to decide for themselves how detailed they want to answer. In the context of parole hearings, in particular, open questions allow the inmate to elaborate, explain their actions, and give insight into their personal growth, which is ultimately crucial for the final decision. Example (1a) illustrates an open question found in our dataset.

In contrast, closed questions, including polar and alternative questions, are designed to elicit specific, limited responses that often require a 'yes' or 'no' answer (see Example (1b)) or a selection from predetermined options. As a consequence the degree of information given by the inmate is restricted and the control of the conversation content is in the hands of the commissioners. From the commissioners' perspective, these questions help to verify specific details during the hearing and ensure clarity and accountability of responses, while simultaneously helping to reduce the likelihood of evasive answers. We chose to collapse polar and alternative questions into one category to reduce the number of labels and therefore simplifying the classification scheme. This ensures more reliable labeling and avoids unnecessary complexity. We also included a category labeled "other" to capture any questions that do not clearly fit into either of the two categories (see Example (1c)).

(1) a. **open question:** Why aren't you doing something besides sitting in prison?
b. **closed question:** Were you under the influence when you shot the kid?
c. **other:** Pardon me?

Our motivation for focusing on open vs. closed-ended questions stems from their central role in managing conversation dynamics (Kikteva et al., 2022) and the asymmetry of power inherent in parole hearings. While there are alternative schemes, we chose this taxonomy to capture the essential contrast between open and closed questions and directly relate to the control of conversation content and inmate participation through the commissioners. Ultimately, the types of questions asked determine the degree of information elicited in the answers, which in turn contribute to the decision the commissioners will make at the end of the hearing.

5.2 Gold standard annotation

To create the gold standard annotation for questions in parole hearings, a total of 750 randomly selected questions posed directly by either the presiding or deputy commissioner to the inmate were extracted from a smaller subcorpus. Our student assistant (a master's student in computational linguistics) and one of the authors were tasked with annotating the questions independently according to the

question taxonomy described above. With 84% of the data coded identically and a Cohen's kappa of 0.67, the initial inter-annotator agreement was substantial. However, in a review process, questions with diverging labeling were re-evaluated by both annotators. As most disagreements were observed in the "other" category, this step was used to refine its application. Many of these disagreements involved clarification questions that were difficult to categorize, such as *"Beating somebody up?"* or *"After the 2015 write-up?"* which sometimes led to ambiguity about whether these should be labelled as closed question or "other". After engaging in discussions and reaching consensus, the questions were relabelled accordingly. In a subsequent step, 1250 more randomly extracted questions were annotated.

The final gold standard (12% of the entire corpus) comprises 1193 closed questions (60%), 667 open questions (33%) and 140 (7%) questions that were annotated as "other".

5.3 Model evaluation

To compare traditional linguistic analysis techniques with cutting-edge AI approaches, we used one rule-based and one LLM-based annotation method. The purpose of this preliminary evaluation was to compare the efficiency, accuracy, and consistency of each method, providing insights into their effectiveness and suitability for the large-scale annotation task. We evaluated the performance of a rule-based system against annotations generated by different models of ChatGPT on the same subset that our human annotators had used.

For rule-based annotations we adapted the English version of the NLP pipeline *LiAnS* (Linguistic Annotation Service), which was originally designed to analyse spoken dialogues in English and German using linguistic features (Gold et al., 2015). We tailored a set of linguistic cues and disambiguation rules specifically to annotate questions according to their question type.

Following the instructions of Törnberg (2023), we additionally prompted ChatGPT-4o, ChatGPT-4o-mini and ChatGPT-3.5-turbo via the OpenAI API[7] with the following zero-shot prompt using Python:

"Classify the following question as "open" (wh-questions), "closed" (yes/no or alternative questions), or "other". Provide the classification followed by the probability with two decimal points. The response should consist of the classification ("open", "closed" or "other") and the probability only, with no additional text.
Question: 'question' "

We designed the prompt to clearly specify the annotation criteria and question types, ensuring that the model generated annotations aligned with our question taxonomy. We also required the model to provide a probability for label assignment, giving us the possibility to monitor its annotation confidence. Following findings from previous research, the temperature was set to 0, in order to keep the annotations deterministic and consistent (Gilardi et al., 2023). After comparing the accuracy scores of annotations generated by ChatGPT-3.5-turbo (0.72), ChatGPT-4o-mini (0.84) and ChatGPT-4o (0.91) the latter was chosen for the automatic annotation. We additionally tested ChatGPT-4o's annotation performance using a similar few-shot prompt (see Example 2 and Table 3 in appendix A), which did not improve results compared to the zero-shot prompt. Upon examining the model's reported probabilities, we observed values between 0.70 and 1.0 for the zero-shot prompting, with only eight questions receiving a confidence score below 0.85.

Metric	Model	Precision	Recall	F1
open	ChatGPT4o	0.93	0.96	0.94
	LiAnS	0.79	0.98	0.87
closed	ChatGPT4o	0.94	0.93	0.94
	LiAnS	0.96	0.62	0.75
other	ChatGPT4o	0.55	0.52	0.53
	LiAnS	0.24	0.70	0.36
macro avg	ChatGPT4o	0.81	0.80	0.81
	LiAnS	0.66	0.77	0.66
weighted avg	ChatGPT4o	0.91	0.91	0.91
	LiAnS	0.85	0.74	0.77
accuracy	ChatGPT4o			**0.91**
	LiAnS			**0.74**

Table 1: Comparison of rule-based question classification model *LiAnS* and ChatGPT-4o based on Precision, Recall, and F1-Score.

Table 1 shows the overall performance metrics for ChatGPT 4o and the rule-based model across the three categories compared to the gold standard annotation: open, closed and other questions. Compared to the rule-based model, ChatGPT 4o demonstrated better performance in all three categories, with an overall accuracy of 0.91 compared to the

[7] https://platform.openai.com/overview

rule-based model's 0.74. Specifically, ChatGPT-4o achieved higher F1-scores for open (0.94 vs. 0.87), closed (0.94 vs. 0.75), and other questions (0.53 vs. 0.36). The low scores for "other" do reflect the disagreement encountered for human annotators in the creation of the gold standard.

Based on these results, ChatGPT 4o is the preferable choice for the annotation of the full 16,039-question dataset.

6 Results

We prompted ChatGPT-4o to annotate all questions directly posed to an inmate by any of the commissioners. Out of 16,039 questions 9990 were annotated as closed (62%), 5385 as open (34%), and 664 as "other" (4%), typically consisting of clarification requests or cut-off questions. This distribution indicates that approximately two-thirds of the questions asked during the 102 selected parole hearings constrain the inmate's response to a predetermined format, while only one-third allow for a more open-ended reply.

Using the fully annotated dataset, we conducted a statistical analysis in order to answer the research questions posed in Section 1. Our first question addresses how the types of questions asked during parole hearings relate to the outcomes of those hearings. Overall, more closed questions were asked in hearings where inmates were found to be eligible for parole ($\hat{\mu}_{\text{granted}} = 94.5$ vs. $\hat{\mu}_{\text{denied}} = 89$). In the case of open questions, the proportion was slightly higher in denied hearings ($\hat{\mu}_{\text{denied}} = 54$ vs. $\hat{\mu}_{\text{granted}} = 50.5$). Nevertheless, the results of a Mann-Whitney Test showed a statistically insignificant relationship between question types and parole hearing outcomes.

The second research question aims to analyze the relationship between the types of questions posed and an inmate's racial background. We are especially interested in whether there are any disparities in the share of questions posed related to an inmate's racial background and whether this impacts their chances of being released on parole. According to our dataset, the share of open and closed questions was higher for Black inmates in comparison to White, Hispanic, and inmates of other ethnic groups (see Figure 1). Given that the distribution of open and closed questions followed a normal and homogeneous distribution, we calculated an F-test to test for significant differences among racial groups and the share of posed questions. Figure 1 shows these differences were not statistically significant, for either open or closed questions. Given the non-normal distribution and heterogeneity of "other" questions, we calculated a Kruskal Wallis Test. This test yielded results at the 0.1 significance level, meaning that the share of "other" questions asked to Black ($\hat{\mu}_{\text{Black}} = 7$) inmates was significantly higher in comparison to all other racial groups ($\hat{\mu}_{\text{White}} = 4$, $\hat{\mu}_{\text{Hispanic}} = 5$, $\hat{\mu}_{\text{Other}} = 3$; p = 0.06, see Figure 2 in appendix B). Upon manual examination of the questions labelled as "other", we found that they primarily consisted of cut-off questions. The majority was incomplete utterances, due to inaudible content and interruptions (as marked and transcribed in the PDF files), or one-word clarification requests. For example, we found instances like *"– know about the fight in November?"* or *"<inaudible>?"*, which are challenging to be interpreted in isolation, as they heavily depend on the context.

Furthermore, Black inmates experienced longer parole hearings, as measured by the page count of the corresponding PDF transcripts in our corpus, though this difference was also not statistically significant. We did not find evidence for an inmate's racial background influencing either the share of questions posed or their likelihood of being released.

To investigate whether the commissioners' gender influences their questioning behavior towards inmates of different races, we investigated the types of questions posed by male and female commissioners during the hearings. We conducted a series of hierarchical linear regressions (see Table 4 in appendix C) to analyze the relationship between the gender of the commissioners and the inmate's race. After assessing the validity of our models by conducting regression diagnostics for all included models in this article, we fixed issues of heteroscedasticity and of non-normality of residuals by using robust standard errors as a base for our calculations (Cribari-Neto and da Glória A. Lima, 2014; Pek et al., 2018). We found that, on average, male presiding commissioners asked significantly fewer closed questions than their female colleagues (see Table 4, Model (1)). Moreover, we found that female presiding commissioners posed fewer closed questions to White inmates compared to male presiding commissioners (see Table 4, Model (1)). We observe a similar pattern for Black inmates, although this finding is only significant at the 0.1 level (see Table 4, Model (1)). With regard

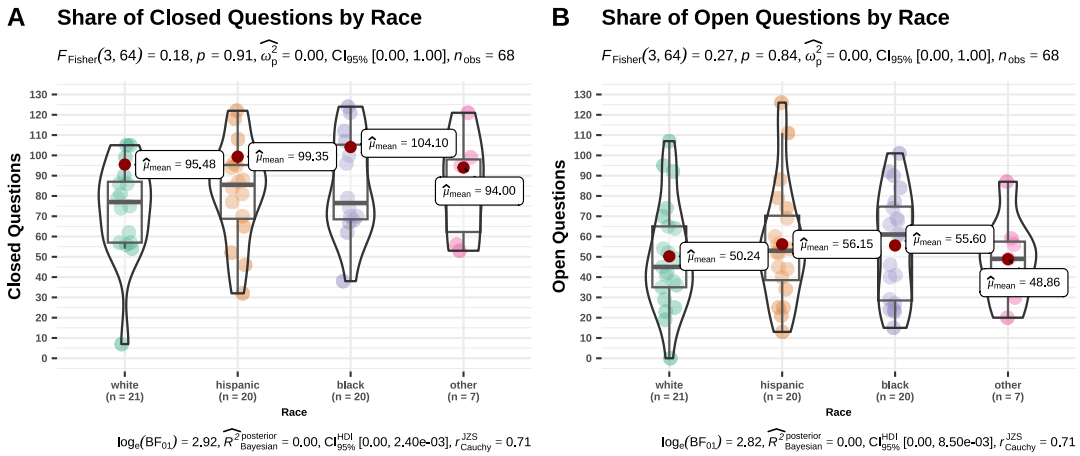

Figure 1: Distribution of Closed (A) and Open (B) Questions by Inmate's Race

to open questions, we found no evidence of a relationship between presiding commissioners' gender and inmate's race (see Table 4, Model (3)).

Assessing the questioning behavior of male and female deputy commissioners on the inmate's race, we found no statistically significant relationship between the share of closed/open questions deputy commissioners (male or female) posed and the race of the inmates. However, we did observe a negative relationship between closed and open questions from deputy commissioners (regardless of gender) and the parole decision. This suggests that, on average, when parole is granted, the share of questions (open or closed) asked by deputy commissioners is lower (see Table 4, Model (2) and (4)).

We also conducted further analysis of the data to identify other factors influencing the share of open and closed questions in the parole suitability hearings within our dataset. Table 2 shows the results of the full model of the calculated step-wise linear regressions. According to our findings, inmates with "third-striker" status were asked fewer closed questions (see 2, Model (1)). We also found that "third-strikers" were asked fewer open questions (see 2, Model (1)). Looking into the total distribution of questions among "third-strikers", we found that they were asked fewer questions in general ($\hat{\mu}_{\text{third-striker}} = 139,50$) when compared to "non-third-strikers" ($\hat{\mu}_{\text{non-third-striker}} = 159$) ($p = 0.05$). With regard to open questions, we found significant effects related to the inmates' age, their age at the time of committing the life offense, and the number of years they have been in prison (see Table 2). Our analysis revealed that older inmates were asked significantly more open questions compared to younger inmates. Furthermore, the older inmates were at the time they committed their crime, the fewer open questions they were asked. Additionally, our analysis showed that the share of open questions declines with increasing time served in prison.

	Closed Questions | Open Questions	
	(1)	(2)
Constant	93.897**	65.688***
	(37.665)	(21.105)
Age	1.901	2.136**
	(1.698)	(0.951)
Age at Crime	−2.171	−2.329**
	(1.696)	(0.950)
Years in Prison	−1.681	−2.298**
	(1.669)	(0.935)
Education	3.016	−0.797
	(5.770)	(3.233)
Gang	1.122	3.965
	(10.338)	(5.793)
Third Striker	−20.209*	−18.332***
	(10.145)	(5.684)
Violent	6.854	1.990
	(18.310)	(10.260)
Non-violent	2.705	7.858
	(21.184)	(11.870)
Sex Offender		
Observations	71	71
R^2	0.119	0.238
Adjusted R^2	0.006	0.139
Residual Std. Error (df = 62)	40.349	22.608
F Statistic (df = 8; 62)	1.051	2.417**
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2: Effect of Age, Age at Crime, Years in Prison, Education, Gang and Crime on Types of Questions.

7 Discussion and Limitations

This study aims to contribute to the research gap regarding the linguistic study of parole hearings, with a special focus on whether the types of questions asked lead to biased outcomes in parole suitability.

Based on our sample, we do not observe a relationship between the type of questions asked and the outcome of the hearings. Contrary to our expectations, open questions do not seem to be positively correlated to a positive parole outcome, implying that board members might prioritize other criteria over the inmate's ability to provide persuasive answers. External factors to the conversation held in the hearing, such as an inmate's rehabilitation progress, low-risk scores in psychological evaluations, and potential for successful reintegration into society might have a bigger impact on parole eligibility than an inmate's articulating power.

In addition, we do not find statistically significant evidence for disparities in the share of questions posed to inmates of different racial backgrounds, suggesting that the evidence regarding racial biases from the social sciences, might not be correlated with commissioners questioning behavior. The fact that female presiding commissioners ask fewer closed questions to White inmates, suggests that White inmates are given a higher chance to articulate themselves in these hearings. However, our findings do not indicate a significant effect on the parole hearing outcome. The low shares of all question types from deputy commissioners in granted parole hearings might reflect the supportive role they play, backing up the presiding commissioner. The presiding commissioners may have already gathered enough information to make a final decision, rendering further questions unnecessary. Our findings regarding the inmate's age and their age at the time of the crime align with the findings from social sciences. Older inmates were not only more likely to be found suitable for parole release, but were also given more often the chance to articulate themselves by being asked more open questions. Similarly, inmates who were very young at the time they committed their life crime were also asked more open questions, allowing them to decide with how much detail they wanted to answer. These findings align with the Board's engagement with the enactment of the Elderly Parole Program and the Youth Parole Program.

Limits to the generalizability of the findings lie in the small sample size, the incompleteness of the manually extracted metadata, and the short timeframe of data selection. Due to the skewed sample (only two female inmates) we were not able to test for the gender-responsiveness of questioning patterns used by the board members. Another limitation of our study is the lack of detailed content analysis of questions posed by the commissioners and the corresponding inmate responses, which might affect the outcome of the parole hearing. To address these limitations and to obtain more generalizable findings on potential question type bias, we plan to officially request metadata for a larger corpus of parole hearings. Regarding the annotation via generative AI, we intend to implement a human-in-the-loop approach, where human oversight will complement ChatGPT's output, ensuring greater reliability, through a combination of the AI's efficiency with the precision of human expertise.

8 Conclusion & Outlook

This study is the first to conduct an in-depth analysis of question patterns in spoken and transcribed parole hearing data, combining insights from social sciences and language technology. While our annotation approach using ChatGPT, yielded very good results, our analysis, based on a sample of 102 parole hearings, did not reveal a significant correlation between the types of questions posed and parole outcomes. Our findings also suggest, that racial disparities in parole hearings might not be correlated to a commissioner's questioning behavior or gender, but might be due to other factors discussed in Section 7.

In order to assess the complexity of the dialogical dynamics and further investigate possible relationships between question types, demographic variables and parole hearing outcome, we plan to expand our corpus. To generate more metadata for in-depth analysis, we consider developing information extraction techniques, such as those used by Hong et al. (2021).

As we are interested in the linguistic strategies employed in parole hearings, the next step is to analyze the content of questions and inmate responses to identify patterns of evasive and non-evasive responses and their potential impact on decisions made by the commissioners. This ongoing research will further bridge the gap between social sciences and computational linguistics, offering a more robust understanding of procedural justice in parole hearings.

Acknowledgments

We are deeply grateful to the California Department of Corrections and Rehabilitation (CDCR) for supplying the parole hearing transcriptions, which were essential to our research.

We also wish to acknowledge the significant contributions of our student assistant, Klymentii Myslyvyi, whose diligent efforts in collecting and organizing the metadata greatly enhanced the quality of our work.

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

Jim Bullington, Ira Endres, and M Rahman. 2007. Open Ended Question Classification using Support Vector Machines. *MAICS 2007*.

Effie Papatzikou Cochran and Cheryl Comeau-Kirschner. 2016. The language of parole: sex offenders' discourse strategy use in Indeterminate Sentence Review Board hearings. *WORD*, 62(4):244–267.

Francisco Cribari-Neto and Maria da Glória A. Lima. 2014. New heteroskedasticity-robust standard errors for the linear regression model. *Brazilian Journal of Probability and Statistics*, 28(1):83 – 95.

Steffen Eckhard and Laurin Friedrich. 2022. Linguistic Features of Public Service Encounters: How Spoken Administrative Language Affects Citizen Satisfaction. *Journal of Public Administration Research and Theory*, 34(1):122–135.

Ingrid Espinoza, Steffen Frenzel, Laurin Friedrich, Wassiliki Siskou, Steffen Eckhard, and Annette Hautli-Janisz. 2024. PSE v1.0: The first open access corpus of public service encounters. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13315–13320, Torino, Italia. ELRA and ICCL.

Rudolf Fisch and Margies Burkhard, editors. 2014. *Bessere Verwaltungssprache.: Grundlagen, Empirie, Handlungsmöglichkeiten.* Duncker Humblot GmbH.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Valentin Gold, Mennatallah El-Assady, Annette Hautli-Janisz, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 32(1):141–158. _eprint: https://academic.oup.com/dsh/article-pdf/32/1/141/11046544/fqv033.pdf.

Hyukjun Gweon and Matthias Schonlau. 2023. Automated Classification for Open-Ended Questions with BERT. *Journal of Survey Statistics and Methodology*, 12(2):493–504. _eprint: https://academic.oup.com/jssam/article-pdf/12/2/493/57170483/smad015.pdf.

Katie Hail-Jares. 2019. Weighing Words: The Impact of Non-victim Correspondence on Parole Board Decisions. *Justice Quarterly*, 38(4):678–700.

Annette Hautli-Janisz, Katarzyna Budzynska, Conor McKillop, Brian Plüss, Valentin Gold, and Chris Reed. 2022a. Questions in argumentative dialogue. *Journal of Pragmatics*, 188:56–79.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022b. QT30 : A Corpus of Argument and Conflict in Broadcast Debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Paris. European Language Resources Association (ELRA).

Marc Hertogh. 2018. *Front-Line Officials and Public Law*, pages 131–146. Palgrave Macmillan UK, London.

Clara Holzinger. 2020. 'We don't worry that much about language': street-level bureaucracy in the context of linguistic diversity. *Journal of Ethnic and Migration Studies*, 46(9):1792–1808. PMID: 32405261.

Jenny Hong, Catalin Voss, and Christopher Manning. 2021. Challenges for Information Extraction from Dialogue in Criminal Law. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 71–81, Online. Association for Computational Linguistics.

Beth M. Huebner and Timothy S. Bynum. 2008. The Role of Race and Ethnicity in Parole Decisions. *Criminology*, 46(4):907–938.

Aikaterini-Lida Kalouli, Katharina Kaiser, Annette Hautli-Janisz, Georg A. Kaiser, and Miriam Butt. 2018. A multingual approach to question classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2715–2720, Paris. ELRA. ISBN: 979-10-95546-00-9.

Zlata Kikteva, Kamila Gorska, Wassiliki Siskou, Annette Hautli-Janisz, and Chris Reed. 2022. The Keystone Role Played by Questions in Debate. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 54–63, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, page 1–7, USA. Association for Computational Linguistics.

Gabriela Lotta and Roberto Pires. 2019. Street-level bureaucracy research and social inequality. In *Research handbook on street-level bureaucracy*, pages 86–101. Edward Elgar Publishing.

Steven Maynard-Moody and Michael Musheno. 2012. Social Equities and Inequities in Practice: Street-Level Workers as Agents and Pragmatists. *Public Administration Review*, 72(s1):S16–S23.

Gaël Le Mens, Balázs Kovács, Michael T. Hannan, and Guillem Pros. 2023. Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences*, 120(49):e2309350120.

Kathryn D. Morgan and Brent Smith. 2008. The Impact of Race on Parole Decision-Making. *Justice Quarterly*, 25(2):411–435.

William H. Moseley and Margaret H. Gerould. 1975. Sex and Parole: A comparison of male and female parolees. *Journal of Criminal Justice*, 3(1):47–57.

Jolynn Pek, Octavia Wong, and Augustine C. M. Wong. 2018. How to Address Non-normality: A Taxonomy of Approaches, Reviewed, and Illustrated. *Frontiers in Psychology*, 9.

Nadine Raaphorst. 2022. Administrative Justice in Street-Level Decision-Making: Equal Treatment and Responsiveness. In *The Oxford Handbook of Administrative Justice*. Oxford University Press.

Elisabeth Scheibelhofer, Clara Holzinger, and Anna-Katharina Draxl. 2021. Linguistic diversity as a challenge for street-level bureaucrats in a monolingually-oriented organisation. *Social Inclusion*, 9(1):24–34.

Martin Silverstein. 2006. Justice in Genderland: Through a Parole Looking Glass. *Symbolic Interaction*, 29(3):393–410.

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*.

Tanya Stivers and N.J. Enfield. 2010. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626. Question-Response Sequences in Conversation across Ten Languages.

Graham Todd, Catalin Voss, and Jenny Hong. 2020. Unsupervised Anomaly Detection in Parole Hearings using Language Models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 66–71, Online. Association for Computational Linguistics.

Petter Törnberg. 2023. How to use LLMs for Text Analysis. *Preprint*, arXiv:2307.13106.

Kathryne M Young, Debbie A Mukamal, and Thomas Favre-Bulle. 2015. Predicting Parole Grants: An Analysis of Suitability Hearings for California's Lifer Inmates. *Fed. Sent'g Rep.*, 28:268.

Kathryne M. Young and Jessica Pearlman. 2022. Racial Disparities in Lifer Parole Outcomes: The Hidden Role of Professional Evaluations. *Law 38; Social Inquiry*, 47(3):783–820.

A Prompt engineering

A.1 Few-shot prompt

(2) Classify the following question as "open" (wh-questions), "closed" (yes/no or alternative questions), or "other." Also, return the probability of it being that specific question type. The output should only contain three words: "open", "closed" or "other", and the probability with two decimal points.

Examples:
1. Question: "Why were you drinking?" Output: open 0.95
2. Question: "You were not doing anything illegal?" Output: closed 0.95
3. Question: "Either on your own or through the institution?" Output: closed 0.90
4. Question: "Huh?" Output: other 0.95

Now, classify the following question: question: 'question'.

Metric	Precision	Recall	F1
open	0.92	0.94	0.93
closed	0.96	0.86	0.90
other	0.39	0.71	0.50
macro avg	0.76	0.84	0.78
weighted avg	0.91	0.88	0.89
accuracy			**0.88**

Table 3: Precision, Recall, and F1-Score of few shot prompt.

B Share of "Other" Questions by Race

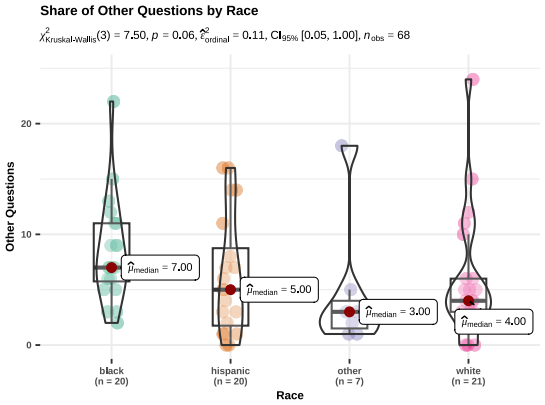

Figure 2: Distribution of "Other" Questions by Inmate's Race.

C Hierarchical Linear Regressions

	Closed Questions (1) (2) | Open Questions (3) (4)			
	Closed Presiding	Closed Deputy	Open Presiding	Open Deputy
	(1)	(2)	(3)	(4)
Constant	49.718***	83.770***	17.454**	39.330***
	(21.939)	(35.055)	(11.206)	(27.542)
Parole	5.845	−17.770**	2.638	−12.330**
	(8.374)	(8.767)	(4.774)	(6.048)
Presiding Commissioner: Male	−33.313*		−6.592	
	(25.909)		(13.817)	
Deputy Commissioner: Male		−16.708		3.868
		(34.491)		(27.284)
White	−24.287	−16.170	−2.782	−6.130
	(22.369)	(39.851)	(12.269)	(29.247)
Black	−20.324	−12.913	−1.659	−11.044
	(23.200)	(37.525)	(12.265)	(28.889)
Hispanic	−18.256	−18.077	−5.109	−6.247
	(23.085)	(37.188)	(12.261)	(27.597)
Other				
Male Presiding Commissioner * White	42.442**		12.498	
	(29.256)		(15.872)	
Male Presiding Commissioner * Black	33.234*		6.947	
	(26.714)		(14.262)	
Male Presiding Commissioner * Hispanic	26.616		16.519	
	(27.691)		(15.069)	
Male Presiding Commissioner * Other				
Male Deputy Commissioner * White		15.877		−3.006
		(40.685)		(29.699)
Male Deputy Commissioner * Black		28.206		17.998
		(38.456)		(29.649)
Male Deputy Commissioner * Hispanic		29.458		3.257
		(38.573)		(28.388)
Male Deputy Commissioner * Other				
Observations	68	68	68	68
R^2	0.084	0.081	0.084	0.121
Adjusted R^2	−0.041	−0.044	−0.040	0.002
Residual Std. Error (df = 59)	22.588	33.090	12.170	20.981
F Statistic (df = 8; 59)	0.673	0.646	0.680	1.016

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Effect of Race and Gender on Open/Closed Questions. Full Models of conducted Hierarchical Linear Regressions.