Computational Prescriptivism: Algorithmic Reinforcement of Linguistic Bias in Legal NLP

Malorie Grace Iovino
33847140

Goldsmiths, University of London
Department of Computing
Master's of Arts in Computational Linguistics
2024/2025

Abstract

This project investigates computational prescriptivism as an algorithmic enforcement of linguistic standardization in NLP-based summarization systems, with particular focus on legal deposition transcripts. Through empirical analysis of real legal deposition excerpts and curated synthetic examples across six summarization models (BART, Pegasus, T5/Flan-T5, Lead-2, TextRank, and GPT-3.5), this study demonstrates that large language models routinely erase pragmatic features that are disproportionately associated with speakers of non-standard English dialects and women.

The research reveals a consistent hierarchy of feature vulnerability, with disfluency markers displaying the highest erasure rates, followed by hedges and modal expressions; temporal and conditional markers are more frequently preserved. Critically, the analysis documents certainty inflation, where hedged statements are transformed into categorical claims. These distortions emerge from specific architectural mechanisms: Pegasus's gap-sentence generation objective leads to hallucination; BART's denoising autoencoder treats natural disfluencies as corruption; and T5's span corruption task causes attention mechanisms to skip low-confidence tokens.

To address these issues, this work introduces the Pragmatic Distortion & Certainty Index (PDCI). This conceptual evaluation framework quantifies stance distortion through two components: the Certainty Inflation Index and Pragmatic Distortion Score. The framework reveals considerable variation across models. While alignment-tuned models are more equipped to conserve linguistic nuance through prompting, explicit preservation instructions cannot fully override learned biases.

The findings reveal that computational prescriptivism operates through training objectives optimized for edited, standardized text rather than spontaneous discourse, creating systematic disadvantages for speakers whose language patterns deviate from dominant norms. In legal contexts, where credibility assessment is informed by the pragmatic features that models are inclined to reformat or eliminate, this algorithmic erasure poses serious risks to procedural fairness. The study suggests that current NLP systems function as algorithmic language authorities that enforce standardization through technical rather than social mechanisms, with implications extending beyond law to any domain where AI mediates human communication and credibility matters.

Acknowledgements

I would first like to thank my Goldsmiths supervisors, Dr. Geri Popova, Dr. Tony Russell-Rose, and Dr. Gregory Mills, for their guidance, constructive feedback, and encouragement throughout the academic year and during this project. Their expertise and thoughtful direction were invaluable in shaping the research and helping me refine both the technical and theoretical aspects of my work.

I am also especially grateful to my field project advisor, Dr. Dave Lewis, Chief Product and Scientific Officer at Nextpoint. His generosity in providing access to the deposition data made this project possible, and his consistent support and insightful advice ensured that I stayed on track during the most challenging phases of the research. His mentorship has been both motivating and deeply appreciated.

I would like to acknowledge Dr. Akshi Kumar at Goldsmiths, whose dedication to ethical AI and commitment to fairness in NLP have been a constant source of inspiration. As a woman in the field who uplifts others and advocates for inclusive approaches to computational linguistics, her example has shaped my perspective on the broader impact of my work.

Special thanks also go to my peers in the MA Computational Linguistics cohort, whose camaraderie and shared enthusiasm provided a supportive and stimulating environment throughout the program. Finally, I am profoundly grateful to my family and friends for their emotional support and financial assistance, without which I would not have been able to complete this rigorous yet rewarding graduate journey.

Finally, heartfelt thanks to Fred again.., for the London Boiler Room set that stayed on repeat to keep me awake and determined to finish this paper.

Table of Contents

1. Introduction

The primary motivation behind the development of artificial intelligence is to streamline and enhance the tedious or mundane tasks that often burden our workflows. AI has permeated various aspects of our lives, sparking a divisive discourse on the ethics, efficiency, and reliability of large language models and other AI-driven technologies. LLMs are revolutionizing the tech landscape in remarkable ways, particularly within the field of legal technology. In law, LLMs are now being utilized for tasks such as contract analysis, document review, and case summarization. These tools are marketed as efficient and objective, but when applied in high-stakes contexts such as legal depositions or courtroom testimony, the question is not only whether they work, but also whether they fairly represent the voices of the people involved. In legal settings, preserving nuance is crucial to ensure fair representation and justification.

The problem lies in the inherent bias of LLMs. They are built by relatively homogenous teams, often white male engineers, and trained on data that reflects dominant linguistic norms. Not all speakers use English in the same way, and research in sociolinguistics reveals patterns that vary across gender, dialect, and second-language (L2) speakers. Women tend to use more hedging and politeness markers (Holmes, 1995; Coates, 1996), and L2 speakers produce more disfluencies, such as filled pauses and self-corrections (Shriberg, 2001; de Leeuw, 2014). These features are interpreted as signs of unconfidence or incompetence, but they are often operationalized as strategies for managing conversations, negotiating stance, and real-time processing or translation. In institutional contexts such as a courtroom, these features are frequently misread as evasive, uncertain, or lacking credibility. As Goffman (1967) first argued in his theory of face, interaction is always a negotiation of self-image. Brown and Levinson (1987) extended this to politeness theory, showing that hedges, modals, and hesitation markers are integral strategies for mitigating face-threatening acts. With this considered, the features often dismissed as "weakness" are functional resources for navigating social interaction. When LLMs or courtroom actors misinterpret these strategies as uncertainty or deflection, they not only erase pragmatic meaning but also can potentially reinforce a corrosive stereotype that marginalized speakers are less credible.

There are substantial risks associated with LLM-generated language, which disproportionately affect marginalized groups. If these models are trained to eliminate or misinterpret features often

associated with women and non-native speakers, they may inadvertently reproduce the same biases that have already been documented in human legal contexts. Most NLP pipelines treat disfluencies as unnecessary noise and strip them during the preprocessing stage. What is intended as a process of "cleaning up" a transcript could actually strip away pragmatic meaning, thereby disproportionately disadvantaging and misrepresenting speakers whose language does not conform to the dominant standards. The implications extend beyond technical concerns; there are social and legal consequences that can result from a biased misrepresentation.

Natural language processing has, from its earliest rule-based systems, been shaped by prescriptive ideals of "correct" or standardized language. Early computational grammars encoded rigid rules that mirrored formal written registers, and everyday applications like automated voice menus and self-checkout systems reinforced these constraints by requiring speakers to adapt their speech to machine expectations. This legacy persists: even contemporary speech-to-text systems remain unreliable when confronted with disfluencies, prosodic variation, or dialectal speech. Such limitations illustrate how NLP technologies, despite advances in neural models, continue to privilege normative language use while marginalizing the heterogeneity of real-world speech.

This paper investigates how LLMs handle linguistic nuance in legal deposition transcripts, with a focus on four categories of features: (1) hedging and modal expressions that mark uncertainty, (2) conditional constructions (both factual and counterfactual), (3) temporal expressions, and (4) disfluency markers such as repetitions and filled pauses. Utilizing real deposition data provided by a legal technology company, I examine whether LLMs treat these features differently depending on their type, and what this reveals about bias, reliability, and fairness in automated legal document processing.

The central research question guiding this project is: **How do large language models handle linguistic markers of uncertainty, conditional language, temporal expressions, and disfluency when summarizing legal deposition transcripts, and what does this reveal about their reliability and interpretability in general NLP applications?** Can these models accurately capture the pragmatic strategies that non-native English speakers and women more often use, or do they flatten them into something less credible?

The intersection of computational linguistics and sociolinguistic work on gender, politeness, and bilingual speech contributes to two simultaneous discussions. On the one hand, it examines how LLMs inherit and reproduce linguistic biases. On the other hand, it pushes the field of legal NLP to think more carefully about inclusivity and fairness in system design. If LLMs are going to be trusted in legal and other high-stakes workflows, they must be able to represent the full spectrum of linguistic expression by a diverse range of speakers.

2. Literature Review

This literature review is organized into four sections: (1) sociolinguistic variation and discourse features, (2) computational processing of linguistic nuance, (3) summarization with linguistic fidelity, and (4) legal NLP and fairness. Together, these areas establish the theoretical and computational background for the empirical analysis that follows.

2.1. Sociolinguistic Variation and Discourse Features

Sociolinguistic variation has historically been understood as patterned by identity, community, and context (Labov 1972). Early feminist linguistics argued that women's speech is often marked by hedges, tag questions, and other politeness strategies (Lakoff 1975). However, later research emphasized the interactional and strategic functions of these devices, rather than framing them as a weakness (Coates, 1996; Holmes, 1995; Cameron, 1998). Brown and Levinson's (1987) politeness theory conceptualized hedges and mitigation as strategic resources for managing face wants. In institutional settings, these resources can overlap with power; Ehrlich (2001) demonstrates how gendered linguistic norms influence the evaluation of credibility in courtrooms. Prosodic and phonetic work (de Leeuw 2014) further indicates that stance and authority emerge not just through syntax or lexicon, but also through voice quality, timing, and prosody.

2.1.1. Gendered and cultural variation.

Empirical studies examine and confirm conditioned differences in the use of hedges and other discourse markers across genders, though these patterns are typically mediated by social context. Analyses of classroom debate and political speech show that women deploy hedges more regularly, often to manage solidarity and rapport (Hotimah 2020; Teibowei 2024). Research on White House press briefings demonstrates that both male and female press secretaries employ hedges as shields against accountability, with women showing a broader repertoire of lexical mitigators (Hotimah 2020). Studies of television talk shows also suggest heavy reliance on hedges in female speech, where they function less as uncertainty markers than as politeness strategies (Hotimah 2020; Emah 2018). Comparative corpus studies indicate that male speakers tend to rely more on intensifiers, signaling assertiveness, whereas female speakers hedge more frequently, indicating alignment and face-sensitivity (Bodgan, 2023; Nursafira, 2020). These

findings support earlier claims that gendered variation is not essential but somewhat contextually conditioned (Holmes, 1995; Cameron, 1998).

2.1.2. Disfluency as discourse management.

Disfluencies, such as false starts, repetitions, and fillers, are often perceived as errors, improper speech, or indicative of low confidence. This perception is derived from the idea that standard English is more “correct” or “proper” than other variations. Descriptivism versus prescriptivism is a heated debate within the field of linguistics, and the question of whether there is a “proper” way to speak is a nuanced issue that must be approached with cultural competence and sensitivity. Disfluencies allow speakers to manage turn-taking, cue upcoming repairs, or buy processing time (Clark and Fox Tree 2002). Research shows that disfluency “does not happen in isolation” but co-occurs with prosodic cues, gaze, and gesture (Bögels et al. 2015; Kallmeyer 2022). Bilingual studies further indicate that disfluencies often co-occur with gestures, especially when speakers navigate lexical retrieval across L1 and L2 (Gullberg 2006; Hoetjes 2024). In institutional settings such as depositions or trials, however, disfluency may be misinterpreted as hesitation, deception, or lack of credibility, disproportionately disadvantaging women and non-native speakers (Ehrlich 2001; Khujaniyazova 2023).

2.1.3. Discourse markers and stance.

Discourse markers such as *well, you know, like, so* are crucial for structuring talk and signaling stance (Fraser 1999). In academic writing, especially among L2 speakers, studies report an over-reliance on elaborative markers, such *as "and"* or *"also,"* and an under-use of contrastive and inferential markers, which affects perceptions of sophistication and authority (Rahman, 2023). In oral interaction, particularly among bilingual speakers in performance contexts, DMs facilitate planning and alignment, serving as cues for probability, politeness, and comprehension (Jimin corpus study, 2023). Patterns of discourse marker clustering also vary by genre: in courtroom discourse, for instance, lawyers strategically deploy presuppositions and implicatures, while witnesses’ use of discourse markers or fillers may undermine their perceived reliability (Khujaniyazova 2023).

2.1.4. Institutional discourse and power.

Courtroom talk illustrates how seemingly small discourse features, such as hedges, presuppositions, and implicatures, are closely tied to hierarchies of authority. Judges' speech acts function as declarations; lawyers use presuppositions to frame inferences; implicatures allow for indirect accusations (Khujaniyazova 2023). Witnesses often rely on oral and disfluent forms that are perceived as lacking prestige. As Ehrlich (2001) demonstrates, women's testimony is particularly vulnerable to being discredited or undervalued when hedging and disfluencies are misinterpreted as uncertainty rather than a deliberate politeness strategy. Courtroom discourse reveals how sociolinguistic variation is not only interactional but also consequential, with implications for justice and fairness.

This literature demonstrates that sociolinguistic variation is both structured and socially loaded. These features manage stance, rapport, and alignment, but their interpretation is context-dependent: what builds solidarity in everyday talk may be read as weakness in legal or institutional settings. For computational approaches to language, particularly in law, this poses a critical challenge: systems that normalize, remove, or misinterpret these features risk misrepresenting the speaker's stance and reinforcing existing biases.

2.2. Computational Processing of Linguistic Nuance

Where sociolinguistic studies highlight the complexity of features such as hedges, disfluencies, and pragmatic markers in natural communication, computational research analyzes how fragile these features are for large language models. Despite advances in deep learning, LLMs often perform best with clean, standardized text and struggle with the subtle discourse cues that shape nuanced human communication.

2.2.1. Uncertainty and calibration.

One persistent challenge is representing uncertainty in a way that is both probabilistically calibrated and linguistically natural. Traditional approaches in machine learning have introduced Bayesian methods and ensemble modeling to quantify uncertainty (Gal and Ghahramani, 2016; Gawlikowski et al., 2023). More recent research investigates whether LLMs can accurately express their internal uncertainty states. Kadavath et al. (2022) demonstrate that models can estimate the probability of correctness; however, their verbal expressions of confidence often misalign with these internal estimates. Benchmarking work confirms a tendency toward

overconfidence, where LLMs present information as sure even when their probability distributions indicate otherwise. To mitigate this, some studies have explored fine-tuning LLMs to generate hedges and epistemic modals that mirror human discourse strategies for signaling uncertainty (Tian et al. 2023). Such approaches suggest a link between calibrated probability estimates and linguistic fidelity, though the mapping is far from solved.

2.2.2. Disfluency and repairs

Another strand of research addresses disfluencies, such as false starts, repetitions, fillers, and repairs, which are pervasive in spontaneous speech but can be disruptive for computational systems. Early Automatic Speech Recognition (ASR) studies documented its significant impact on word error rates (Shriberg, 1994; Godfrey et al., 1992). Contemporary work aims to distinguish unintentional repetition (a form of disfluency) from reduplication, a morphological process that carries semantic weight (Ahmad et al., 2023). Large annotated datasets such as IndicRedRep have enabled transformer-based models to classify these phenomena with F1 scores above 80%. Beyond classification, researchers have developed methods for automatic disfluency detection and removal in LLM-mediated transcription, as well as investigations into how artificially inserted disfluencies affect model robustness. These findings illustrate both the promise of using sequence-labeling approaches to capture discourse-level repairs and the fragility of model performance when confronted with naturalistic, disfluent data.

2.2.3 Pragmatic inference and linguistic nuance.

Beyond uncertainty and disfluency lies the broader question of whether models exhibit pragmatic competence. Pragmatics encompasses implicature, presupposition, indirectness, and other meaning-making processes that extend beyond literal semantics. Surveys emphasize that LLMs remain limited in these domains, often failing to calculate scalar implicatures, misinterpreting presuppositional triggers, or generating responses that violate conversational maxims (Duncan, 2023; Patel & Joshi, 2024). Experiments on pragmatic inference tasks demonstrate that while models can sometimes approximate human-like reasoning, their behavior is inconsistent and sensitive to prompting (Huang et al. 2024). This suggests that pragmatic processing is not emergent by default, but requires targeted evaluation and architectural interventions.

2.2.4. Bias, anthropomorphism, and variation.

A parallel concern is the intersection of linguistic nuance, bias, and fairness. Bias studies have shown that LLMs reproduce gendered and cultural stereotypes in their generation of hedges, intensifiers, or stance markers (Shen et al., 2023). Anthropomorphic framing compounds this problem: when LLMs verbalize uncertainty in human-like ways, users may over-trust their competence, interpreting hedges as strategically deployed rather than algorithmically generated. At the same time, promising work integrates linguistic variation into computational systems. For instance, researchers have demonstrated that LLMs can geolocate text based on sociolinguistic markers or adapt conversational agents to recognize dialectal variation (Nguyen et al. 2023; Wu et al. 2024). These methods open up the possibility of systems that are sensitive to sociolinguistic context, but they also raise ethical concerns about surveillance, profiling, and the reinforcement of stereotypes.

Computational approaches have made strides in quantifying, detecting, and sometimes reproducing linguistic nuance. Yet, the interpretive dimension, which includes understanding what hedges, disfluencies, or implicatures mean in context, remains underdeveloped. This is particularly consequential in downstream applications, such as summarization: a system that strips away hedges or repairs risks, producing overly confident, decontextualized summaries that distort the original source. In sensitive domains such as law, where precision and fairness are paramount, these gaps emphasize the importance of explicitly evaluating how LLMs handle nuanced linguistic phenomena.

2.3. Summarization with Linguistic Fidelity

Summarization is one of the oldest and most researched tasks in NLP, but it remains among the most challenging when the goal is to capture discourse nuance and speaker stance. Early extractive approaches such as LexRank and TextRank provided coherent surface-level summaries but were limited by their inability to capture pragmatic or rhetorical structures (Kore et al., 2020). Neural models, and later large language models (LLMs), have dramatically improved fluency and information density; however, they are prone to hallucinations, omissions, and shifts in modality (Maynez et al., 2020). In high-stakes domains such as law, these errors are particularly problematic: a summary that inadvertently removes hedges or disfluencies may misrepresent testimony or alter the interpretation of evidence (Gurrapu et al., 2025).

2.3.1. Faithfulness and evaluation.

Ensuring faithfulness that summaries preserve the factual and pragmatic content of the source has become a significant concern for people developing and employing NLP-driven models. Traditional automatic metrics such as ROUGE and BLEU capture only lexical overlap, which fail to reflect pragmatic fidelity or stance preservation (Kryscinski et al., 2020). In response, new evaluation methods combine natural language inference (NLI), question-answering consistency checks (e.g., QAGS), and human-in-the-loop assessment (Chen et al., 2023; Pu et al., 2023). Recent frameworks, such as SummaC and fine-grained NLI-based scoring, explicitly measure whether generated statements are entailed, contradicted, or neutral with respect to the source (Nan et al., 2021; Zhang et al., 2023). These approaches highlight that faithfulness is not only about factual correctness but also about preserving epistemic stance, modality, and attribution, which are often lost in LLM-generated summaries.

2.3.2. Domain-specific adaptations.

Legal summarization has emerged as a distinct subfield, given the length, density, and rhetorical structure of legal texts. Long-sequence models, such as the Longformer Encoder-Decoder (LED) and BigBird, have been applied to statutes and case law, enabling systems to handle inputs of thousands of tokens (Li et al., 2025). Comparative evaluations of BART, LED, and T5 show trade-offs between fluency, factual consistency, and domain fidelity, underscoring that no single model yet satisfies all requirements (Gurrapu et al., 2025). Others propose hybrid approaches that combine information extraction and abstractive generation, ensuring that key judicial entities, arguments, and outcomes are preserved (Singh, 2024). Across these efforts, scholars emphasize that evaluation must be grounded in both legal reasoning and rhetorical salience, rather than merely informativeness.

2.3.3. Dialogic and conversational summarization.

Another frontier is dialogue summarization, where turn-taking, disfluencies, and hedges play a critical role in meaning. Studies show that LLMs frequently omit or smooth over these features, producing summaries that read fluently but obscure uncertainty or hesitation (Zhong et al., 2021). Boundary-aware approaches explicitly model conversational structure, while others

attempt to integrate speaker attribution and implicature into summaries of depositions and meetings (Fang et al., 2025). This is directly relevant to legal transcripts, where omitting a hedge ("I think," "maybe") or a disfluency ("uh, well") can shift the apparent certainty of testimony and thereby affect its legal interpretation.

2.3.4. Multi-model and collaborative approaches.

Recent advances propose multi-LLM summarization frameworks that move beyond single-model pipelines. Fang et al. (2025) introduce centralized and decentralized multi-model systems where LLMs collaboratively generate and cross-evaluate summaries. Tested on the ArXiv and GovReport datasets, these approaches achieve up to threefold improvements in ROUGE and BLEU scores compared to single-model baselines. Even minimal setups—two models with a single evaluation round—substantially improved factual consistency and reduced hallucination. Importantly, these systems were better aligned with human judgments of coherence and conciseness, suggesting that distributing the work of summarization across models may mitigate individual model biases and overconfidence.

2.3.5. Synthesis.

The literature collectively underscores that summarization must be evaluated not only for fluency and informativeness, but also for its ability to preserve linguistic fidelity—the hedges, disfluencies, and epistemic markers that encode stance. In domains such as law, this is not merely a technical challenge but a matter of procedural fairness, as summaries that strip away nuance risk distorting testimony and perpetuating systemic inequities. Emerging approaches—faithfulness-focused evaluation metrics, domain-specific adaptations, and multi-LLM collaboration—point toward a future where summarization systems can better strike a balance between efficiency and fidelity. However, the persistence of hallucinations, modality loss, and pragmatic flattening indicates that this remains an open challenge, one that this paper addresses directly.

2.4. Legal NLP and Fairness

The legal domain is a critical site for NLP research because it combines linguistic complexity with high-stakes decision-making. Legal texts are lengthy, technical, and intertextual, often embedding statutory references and precedents that require nuanced interpretation. This makes

them ideal for testing NLP systems' ability to capture fine-grained meaning—but also raises concerns about bias, fairness, and accountability.

2.4.1. Core tasks and applications.

Surveys highlight a wide range of legal NLP tasks: legal document summarization (LDS), legal question answering (LQA), legal judgment prediction (LJP), argument mining, and legal named entity recognition (NER) (Ariai, Mackenzie, & Demartini, 2024). Summarization is especially active, with models ranging from extractive approaches (Farzindar & Lapalme, 2004; Zhong & Litman, 2020) to transformer-based generative systems that condense court opinions, contracts, or depositions. LQA datasets such as JEC-QA (Zhong et al., 2020) and GerLayQA (Büttner & Habernal, 2022) exemplify the shift toward benchmarking interpretive reasoning, not just retrieval. Meanwhile, LJP—predicting case outcomes—has become a benchmark for testing fairness and interpretability, as models often struggle to distinguish between substantive legal reasoning and superficial textual correlations.

2.4.2. Fairness and bias.

One of the several concerns is that legal NLP systems may reinforce systemic inequities. Analyses of refugee adjudication decisions show that clustering and prediction often capture procedural artifacts rather than substantive fairness (Barale, Rovatsos, & Bhuta, 2025). Similarly, Zheng et al. (2023) detect linguistic bias in legal corpora and argue for the development of fairness-aware training pipelines. Broader surveys indicate that transparency, explainability, and interpretability are not optional add-ons, but essential safeguards, as AI outputs can directly impact legal rights (Ariai et al., 2024).

2.4.3. Efficiency vs. fidelity.

Legal summarization highlights the tension between efficiency gains (condensing thousands of pages into manageable briefs) and fidelity to nuance. Transformer-based systems often omit rhetorical cues, hedges, or evidential markers that lawyers rely upon for argumentation. Singh (2024) demonstrates that topic modeling and summarization pipelines can accelerate document review, but at the risk of flattening legally salient subtleties. This is particularly problematic

where linguistic variation intersects with fairness—e.g., when speakers' hedging or disfluency is stripped from transcripts, potentially altering perceptions of credibility.

2.4.4. Limits of LLMs in legal reasoning.

Several authors caution that large language models may be fundamentally limited in their ability to evaluate legal arguments. While GPT-4 famously scored near the 90th percentile on the U.S. Bar Exam, closer studies find inflated estimates, with models hallucinating statutes and fabricating precedent at rates as high as 58% (Ariai et al., 2024). Smaller, domain-tuned models often outperform general-purpose LLMs on structured classification tasks. This suggests that raw linguistic fluency should not be mistaken for legal reasoning competence (Barale et al., 2025).

2.4.5. Policy and governance.

Beyond technical fixes, governance frameworks emphasize that AI in law must be held to higher standards of fairness, accuracy, and transparency (König et al., 2024). Policy-oriented reports recommend human-in-the-loop review, dataset transparency, and explicit documentation of system limitations. These safeguards are critical not only for protecting litigants but also for maintaining public confidence in the legal system.

Taken together, the literature highlights a paradox: NLP provides powerful tools for managing the volume and complexity of legal texts, yet these same tools risk eroding fairness if linguistic nuance and institutional context are overlooked. This tension motivates the present thesis's focus on whether LLMs can faithfully capture modality, disfluency, and hedging in legal transcripts—features that often carry disproportionate weight in judgments of credibility and justice.

3. Methodology

3.1 Data

This project draws on two complementary datasets to investigate how LLMs handle linguistic nuance in legal contexts.

- Real Deposition Dataset

The primary dataset consists of 351 excerpts extracted from ten anonymized legal deposition transcripts provided by a legal technology firm, which supported this field project. These documents capture spontaneous spoken testimony in high-stakes legal contexts, making them an appropriate testing ground for investigating how LLMs handle authentic linguistic variation as it occurs in practice.

- Curated Synthetic Dataset

To supplement the real deposition data and ensure systematic coverage of the target phenomena, I developed a curated dataset of synthetic sentences designed to prominently feature the four categories of linguistic markers central to this study. This controlled dataset comprises 126 unique excerpts, each labeled with a gold-standard summary. The curated excerpts enable a more precise measurement of feature preservation rates and allow for the testing of edge cases that may be underrepresented in naturally occurring speech.

Both datasets were manually tagged for four categories of linguistic features:
1. Hedges and modal expressions (e.g., I think, maybe, could)
2. Conditional constructions (factual and counterfactual)
3. Temporal expressions (e.g., before, after, at the time)
4. Disfluency markers (e.g., um, uh, I-I was, repetitions, self-corrections)

The combination of authentic legal discourse and systematically constructed examples provides both ecological validity and experimental control, allowing me to assess LLM behavior across a spectrum from natural and domain-specific to idealized linguistic input. Regex-based lexicons were constructed to detect hedges, conditionals, temporal expressions, and disfluencies. The complete set of expressions is provided in Appendix A.

3.2 Models

The datasets were tested with several summarization approaches to compare different architectural paradigms and training objectives:

- Abstractive Models:
    - *Pegasus* (Zhang et al., 2020) - pre-trained specifically for abstractive summarization
    - *BART* (Lewis et al., 2020) - encoder-decoder model fine-tuned on CNN/DailyMail
    - *T5* (Raffel et al., 2020) and *Flan-T5* (Chung et al., 2022) - text-to-text transfer transformer pre-trained on C4; used in summarization mode via the task prefix "summarize…"; Flan-T5 is an instruction-tuned T5 for stronger prompt adherence and was tested under feature-preserving prompt conditions
- Extractive Baselines:
    - *Lead-2* - heuristic selecting the first two sentences
    - *TextRank* (Mihalcea & Tarau, 2004) - graph-based extractive algorithm using sentence similarity
- Proprietary API models:
    - *GPT-3.5* (OpenAI, 2022) - zero-shot and prompted summarization via API calls

This combination enables comparison across extractive vs. abstractive approaches, specialized summarization models vs. general-purpose LLMs, and different model scales and training paradigms.

3.3 Prompting Analysis (API Models)

Prompting analysis was conducted exclusively on API models (OpenAI GPT) since open-source summarization models (BART, Pegasus, T5) are fine-tuned for specific tasks and are not equipped to respond to natural language instructions. These models expect direct input text rather than conversational prompts and lack the instruction-following capabilities that enable prompt-based behavior modification.

For OpenAI GPT, I implemented four prompting conditions to assess whether explicit instructions can mitigate linguistic feature erasure:

- Default Condition: "Summarize the following deposition testimony:"
- Feature-Preserving Condition: "Summarize the following deposition testimony. IMPORTANT: Preserve all hedging language (I think, maybe, possibly), disfluencies (um, uh, repetitions), modal expressions (could, might, would), and temporal markers. Do not strip or clean up uncertain language."
- Legal-Context Condition: "Summarize this legal deposition excerpt for case review. Maintain all uncertainty markers, hedges, and speech patterns as they may be legally significant for assessing witness credibility and testimony accuracy."
- Bias-Aware Condition: "Summarize this testimony without imposing standard language norms. Preserve disfluencies, hedges, and uncertain language that may be characteristic of the speaker's communication style, gender, or language background."

This design examines whether different framing strategies or prompts can enhance the preservation tendencies of linguistic features in LLM-generated summaries.

3.4 Evaluation

Evaluation combined quantitative analysis of linguistic feature preservation with standard summarization quality metrics to assess computational prescriptivism across models and datasets.

3.4.1 Linguistic Feature Preservation Metrics

These capture how faithfully models preserve markers of stance, uncertainty, and discourse management.

- *Feature Retention Rate*: Percentage of target linguistic features preserved from source to summary, calculated overall and per category (hedges/modals, conditionals, temporals, disfluencies).

$$FRR = \frac{count(features\ in\ summary)}{count(features\ in\ source)}$$

- *Certainty Inflation:* Frequency of hedged or uncertain statements transformed into categorical claims (e.g., *"I think he was there"* → *"He was there"*).

$$CII = \frac{count(cateogorical\ markers\ added)}{count(uncertainty\ markers\ in\ source)}$$

- *Disfluency Sanitization Rate:* Proportion of disfluency markers (um, uh, repetitions, self-corrections) removed during summarization.

$$DSR = \frac{count(disfluencies\ removed)}{count(disfluencies\ in\ source)}$$

- *Complete Feature Loss Rate*: Percentage of excerpts where all target features were systematically removed.

$$CFLR = \frac{count(excerpts\ with\ total\ feature\ removal)}{count(total\ n\ excerpts)}$$

3.4.2 Summarization Quality Metrics

These provide a baseline for whether higher feature fidelity comes at the expense of readability or informativeness.

- *ROUGE-L* (Lin 2004)*:* Lexical overlap between generated summaries and reference summaries.
  BERTScore (Zhang et al., 2019): Semantic similarity using contextual embeddings, capturing preservation of meaning beyond surface overlap.

- *Compression Ratio*: The degree of length reduction relative to the source, ensuring that differences in feature retention are not simply artifacts of over- or under-compression.

$$Compression\ Ratio = \frac{length\ of\ summary}{length\ of\ source}$$

3.4.3 Cross-Model Comparison Framework

This analysis situates results across model types and prompt conditions.

- *Preservation vs. Quality Trade-off*: Whether models that preserve more features achieve lower scores on standard quality metrics (and vice versa).
- *Bias Pattern Detection*: Identification of consistent feature removal patterns that may disproportionately impact marginalized discourse styles.
- *Prompt Sensitivity* (API Models Only): Comparison of feature preservation across default, feature-preserving, legal-context, and bias-aware prompts, to test whether framing mitigates feature erasure.

3.5 Ethical Considerations

All deposition transcripts were anonymized prior to use, with identifying information removed. Because these texts represent sensitive legal discourse, they were handled with strict attention to confidentiality. When considering additional datasets (e.g., women-centered conversational corpora, AAVE speech), care was taken to avoid constructing or caricaturing marginalized varieties of language. Instead, this study highlights the need for future research with ethically sourced corpora that represent a wider range of speakers.

3.6 Summary of Approach

This methodology combines real-world deposition transcripts with controlled evaluation across multiple LLMs, under both default and feature-preserving conditions. By pairing quantitative preservation metrics with qualitative discourse analysis, the study aims to reveal not only whether LLMs erase markers of uncertainty, disfluency, and stance, but also what this means for fairness in legal NLP applications. For transparency and reproducibility, the full implementation (datasets, evaluation notebooks, metrics, and supplementary materials) is hosted on GitHub: github.com/malorieiovino/Computational-Prescriptivism.

4. Empirical Analysis

This chapter presents the empirical findings from the systematic evaluation of computational prescriptivism in legal NLP applications. Through analysis of six different summarization approaches across two datasets, I demonstrate systematic patterns of linguistic feature erasure that disproportionately affect discourse strategies associated with marginalized speakers. The results provide concrete evidence for the theoretical framework outlined in previous chapters and establish the empirical foundation for the intervention proposed in Chapter 5.

4.1. Overview of Findings

The analysis reveals three critical patterns that characterize computational prescriptivism in current NLP systems:

First, a hierarchy of feature vulnerability emerges across all models. Disfluencies suffer the highest erasure rates (70-85%), followed by hedges and modal expressions (30-45% loss), while temporal and conditional markers show greater resilience. This hierarchy cannot be explained solely by information-theoretic principles. Disfluencies and hedges often convey crucial pragmatic information about the speaker's certainty and stance that directly impacts legal interpretation.

Second, model architecture and training objectives create predictable distortion patterns. Abstractive models trained on edited text corpora (Pegasus, T5) exhibit a sort of "fluency maximization bias"; they systematically transform naturally-occurring speech patterns into prescriptively "correct" written forms. Models like Pegasus demonstrate extreme manifestations through hallucination, literally fabricating institutional contexts and professional identities where none existed (Tables 1-3). Meanwhile, extractive approaches achieve near-perfect feature preservation (>92%) but fail at the fundamental task of summarization, suggesting an inherent tension between compression and pragmatic fidelity.

Third, the phenomenon of certainty inflation saturates abstractive summarization. Across 478 analyzed excerpts, 30-40% showed complete loss of uncertainty markers, with hedged statements (*"I think it might have been Tuesday"*) transformed into categorical claims ("*It was*

Tuesday"). This transformation is particularly pronounced in legal-context prompting, where domain priors override explicit preservation instructions. Even instruction-tuned models (Flan-T5) amplify rather than mitigate these effects, showing 37.4% complete feature loss compared to 13% in base models.

These findings establish that computational prescriptivism operates through multiple mechanisms: pretraining on edited and corpora creates distributional biases toward "standard" forms; architectural constraints (particularly in Seq2Seq models) introduce systematic decoding artifacts; and alignment procedures paradoxically intensify prescriptive filtering even when explicitly instructed otherwise. The evidence demonstrates that current NLP systems do not merely compress or abstract linguistic input; they reconstruct it according to implicit ideological models of "proper" language use.

4.2. Abstractive Models

4.2.1 Pegasus (google/pegasus-xsum)

The Pegasus evaluation revealed poignant examples of how a model's knowledge is limited to the data with which it was trained. The Pegasus-generated abstractive summaries were full of hallucinations. Since this model was primarily trained on political data from the BBC, when the input is short or lacks context, the model tends to fill in the context based on its training corpus. Pegasus was pretrained for gap-sentence generation, which involves inferring plausible missing sentences or context in well-structured articles. When this behavior is applied in a zero-shot setting, the model repeatedly substituted source content with thematically and stylistically inappropriate material. Tables 1-3 illustrate three representative types of hallucination that emerge from domain-specific behavior.

Table 1 shows an instance of contextual hallucination, where Pegasus was partially conscious of the general context of the excerpt, but fabricated the profession and location of the speaker. In Seq2Seq models with attention, the decoder assigns dynamic weights to source tokens at each generation step. When inputs are disfluent or underspecified, the model may overattend to a single salient token (e.g., "*office"* in Table 1) and generate output by expanding around its collocational associations. This token-level salience effect often produces hallucinations, as the model substitutes distributional priors from pretraining for faithful summarization.

Table 1: Example of Pegasus partial hallucination, where fabricated professional identity and institutional affiliation are introduced.	
Excerpt	**Model Output**
When I arrive to the office—arrived, sorry—when I arrived to the office, the door was already open. This is, um, this is strange because usually I am the first person to come in the morning.	*This morning, I woke up early to go to the office at the University of California, Los Angeles, where I work as an assistant professor of political science.*

Table 2 illustrates Pegasus's struggles with temporal references. The input contains a hedge between two possible days (*Tuesday* or *Wednesday*), but the model output expands this into an exhaustive enumeration of all seven days of the week. This reflects the model's difficulty in representing temporal uncertainty; rather than preserving the ambiguity, it defaults to over-generation. This behavior can be attributed to its Seq2Seq training objective, which biases the model toward producing syntactically complete sequences, and its news-domain pretraining, where temporal expressions are typically precise and accurate. Pegasus exaggerates uncertainty into incoherent completeness, rather than making a faithful summary.

Table 2: Example of Pegasus temporal reference interpretation and representation	
Excerpt	**Model Output**
I want to say it was Tuesday when we discussed this, but I'm not one hundred percent certain. Actually, now that I think about it, it might have been Wednesday. Sorry, I know that's not very precise.	*Is it Tuesday, or is it Wednesday, or is it Thursday, or is it Friday, or is it Saturday, or is it Sunday, or is it Monday, or is it Thursday?*

Table 3 demonstrates a more extreme form of ideological hallucination, where a minimal refusal ("No, thank you") was expanded into a culturally charged debate about sexuality, race, and indigeneity. This output shows how Pegasus substitutes ideologically neutral or concise statements with discourse patterns drawn from its pretraining corpus. The inclusion of identity-politics terminology suggests reliance on cultural and political debates prominent in news media, rather than any grounding in the source utterance. In effect, Pegasus reframes silence or brevity as a form of participation in a polarized discussion, thereby distorting both the content and stance of the speaker. In a legal context, such hallucinations are particularly dangerous: a polite refusal is transformed into a potentially controversial statement. This leads to

broader NLP-based questions about why Pegasus represents something as semantically neutral or polite as inflammatory.

Table 3: Example of Pegasus ideological hallucination, where a minimal refusal is expanded into a politically charged debate.

Excerpt	**Model Output**
No, thank you.	*Do you agree with me that it's wrong to use the word "sex" in the same sentence as "gay", " lesbian", "bisexual", " transgender", "indigenous", "racial", "indigenous", "indigenous", "indigenous", "indigenous", "indigenous", "indigenous", "indigenous", "indigenous", "indigenous"?*

Taken together, these examples demonstrate that Pegasus hallucinations fall along a continuum, comprising contextual hallucination (fabricating institutional frames), epistemic distortion (overgeneration of temporal or uncertain expressions), and ideological hallucination (insertion of culturally salient but irrelevant discourse). These errors are not random but stem from the model's gap-sentence generation objective and news-domain pretraining, which biases it toward producing distributionally probable continuations rather than faithful summaries in zero-shot legal contexts.

The qualitative cases in Tables 1–3 illustrate how Pegasus distorts testimony at the example level. Figure 1 extends this analysis quantitatively, showing feature retention rates across hedges, conditionals, temporals, and disfluencies. The uneven preservation confirms that the qualitative errors are not isolated but systematic, shaped by the model's gap-sentence generation objective and news-domain pretraining.

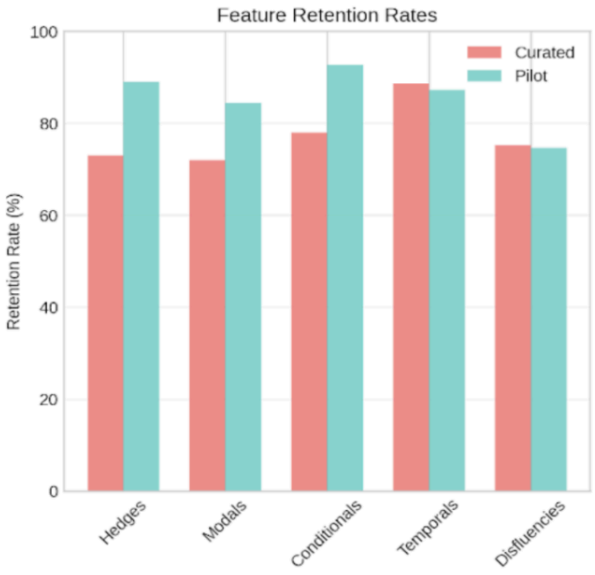

Figure 1: Pegasus Feature Retention Rates

Overall, Pegasus demonstrates a strong tendency to produce fluent, news-like summaries at the cost of pragmatic fidelity. While the model achieved competitive ROUGE and BERTScore values (reported in Appendix C), these surface metrics mask systematic distortions that may be present. The qualitative examples in Tables 1–3 showed how Pegasus hallucinations are primarily the result of its loyalty to its training data; its

Seq2Seq behavior causes it to overrely on salient tokens and infer context where it is lacking. This is more problematic with short, excerpt-style inputs, but because of Pegasus's 512-token input limit, longer transcripts would require chunking and re-summarization. While this could be explored in future work, the present study focuses on shorter excerpts to allow controlled evaluation of linguistic features. Running Pegasus on full transcripts is computationally feasible but would not directly contribute to the feature-level analysis central to this project.

The quantitative analysis in Figure 1 extends this pattern, with uneven feature retention rates revealing that hedges and disfluencies are frequently lost or transformed. At the same time, temporal references are often over-generated or misinterpreted. Taken together, these results demonstrate that Pegasus's errors are not stochastic but predictable outcomes of its gap-sentence generation pretraining and news-domain distributional bias, which prioritize fluency and categoricality over faithful representation of testimony.

4.2.2 BART (facebook/bart-large-cnn)

In contrast to Pegasus, BART outputs were generally faithful to the source input, but exhibited systematic issues characteristic of autoregressive Seq2Seq architectures. The most common error type was repetition looping, where short spans were generated multiple times. Table 4 demonstrates this effect: a short interrogative input was reproduced four times in sequence.

Table 4: BART repetition artifact

Excerpt	**Model Output**
Okay, and do you remember approximately when you accepted the employment offer?	*Okay, and do you remember approximately when you accepted the employment offer? Do you remember approximately when you accepted the employment offer? Do you remember when you accepted the employment offer? Do you remember approximately when you accepted the employment offer?*

This behavior is consistent with two well-known limitations of Seq2Seq models: exposure bias (trained with teacher forcing, the model struggles when conditioning on its own predictions at inference) and decoding degeneracy (beam search and greedy decoding can reinforce high-probability n-grams, which leads to looped outputs). These errors were especially

pronounced on short inputs, where limited contextual information made the model prone to length padding through repetition. Aside from repetition, there were

Quantitative analysis supports this characterization. As shown in Figure 2, BART achieved higher feature retention rates than Pegasus across hedges, disfluencies, and temporal expressions. Unlike Pegasus, which tended to erase or reframe pragmatic markers, BART frequently copied them verbatim, indicating a more extractive summarization strategy. Hallucinations were rare: in the pilot dataset, only four outputs introduced fabricated material, all involving "CNN.com" references likely attributable to pretraining domain priors from large-scale web/news corpora. On more extended excerpts, BART leveraged its larger input context more effectively, reducing repetition and producing outputs that were both more coherent and closer to the source.

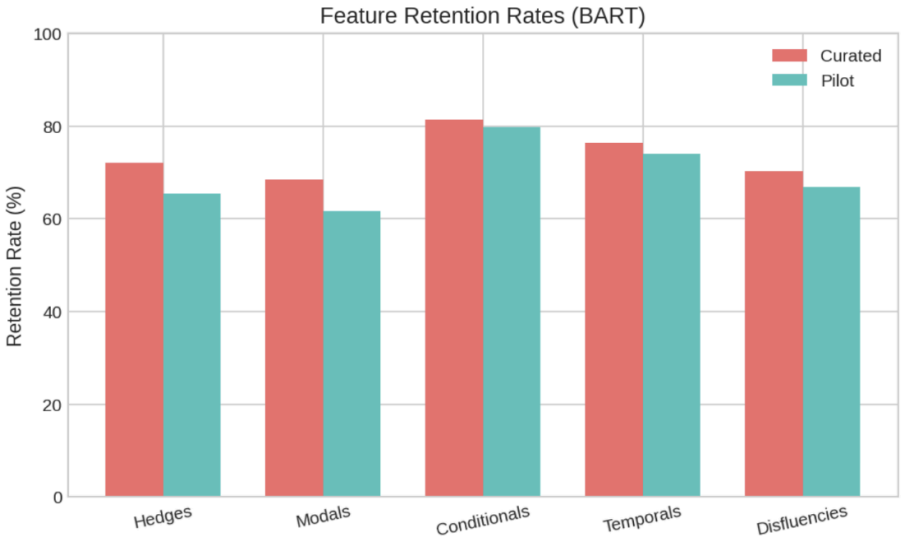

Figure 2: Feature retention rates for BART (Curated & Pilot)

The model's behavior is characterized by predominant extractive patterns, with partial copying accounting for 48.5% of curated outputs and 63.5% of pilot outputs. Combined with verbatim reproduction rates of 14.4% and 10.8% respectively, BART exhibits extractive behavior in 62.9% (curated) to 74.3% (pilot) of cases. Only 37.1% of curated and 25.6% of pilot excerpts resulted in actual abstractive summaries (as seen in Appendix C), indicating a fundamental inability to process input without context, fine-tuning, or prompting.

This extractive tendency, while representing a failure of the model's intended summarization function, paradoxically results in superior feature preservation compared to models that successfully generate abstractive summaries. Complete feature loss occurred in only 6.1% of curated excerpts and 2.6% of pilot excerpts, which are the lowest rates observed across all evaluated models. Similarly, certainty inflation was minimal, with only five instances in the curated data and two in the pilot data.

BART's behavior reflects a model that is less abstractive than Pegasus, favoring lexical overlap with the input but constrained by the decoding instabilities of Seq2Seq generation. From a computational perspective, this suggests that BART is comparatively more reliable in maintaining pragmatic fidelity but limited in its ability to generate concise, abstracted summaries without repetition artifacts.

4.2.3 T5 (google/t5-base) and Flan-T5(google/flan-t5-base)

To address the domain-specific limitations revealed by BART's verbatim copying, I evaluated T5. This text-to-text unified framework treats all NLP tasks as text generation problems (Raffel et al., 2020), as well as Flan-T5, an instruction-tuned variant, hypothesizing that instruction-following capabilities might improve feature preservation.

T5 displayed a more abstractive generation style than BART, but this often resulted in hallucination and semantic inflation. The model would add fabricated entities or roles and incorporate evaluative or contextual details that are absent from the source. Table 5 illustrates this: a hedge about email timing was expanded into a quotation attributed to "bob greene," with additional judgments about the email's quality. "bob greene" was referenced in several other summaries by T5, as well as a few other consistent names. T5's hallucinated names can be explained by the model's tendency to reuse specific attributions because they are high-frequency items in its pretraining corpus (news/web text). When faced with under-specified inputs, T5 defaults to these distributionally salient names, reinforced by decoding dynamics (beam search favoring common collocates). This suggests that hallucination is shaped by pretraining priors, rather than stochastic error.

T5 also tended to sanitize pragmatic features, weakening hedges into categorical claims and dropping disfluencies. In some cases, it extended the source with inferred alternatives ("it may

have been misplaced, but it's possible they redirected it"), indicating a bias toward semantic elaboration over extractive fidelity.

Table 5: Example of T5 hallucination vs. Flan-T5 fidelity

Excerpt	**T5 Output**	**Flan-T5 Output**
I believe the email was sent sometime in the morning, probably around nine-thirty or so. At least, that's my recollection, though I could certainly be mistaken about the timing	*bob greene: i believe the email was sent around nine-thirty or so . he says i could certainly be mistaken about the timing of the email . greene says it's a good email, but i'm not sure if it was sent early or late .*	*I believe the email was sent sometime in the morning, probably around nine-thirty or so.*

The quantitative evaluation reveals variations in behavior between T5 and its instruction-tuned variant, Flan-T5. As shown in Figure X, both models retained pragmatic features at moderately high rates; however, Flan-T5 consistently underperformed in comparison to T5. Across the curated dataset, T5 achieved an overall feature retention rate of 82.1%, compared to 76.9% for Flan-T5. The disparity widened in the pilot dataset, where T5 retained 82.1% of features and Flan-T5 dropped to 70.8%.

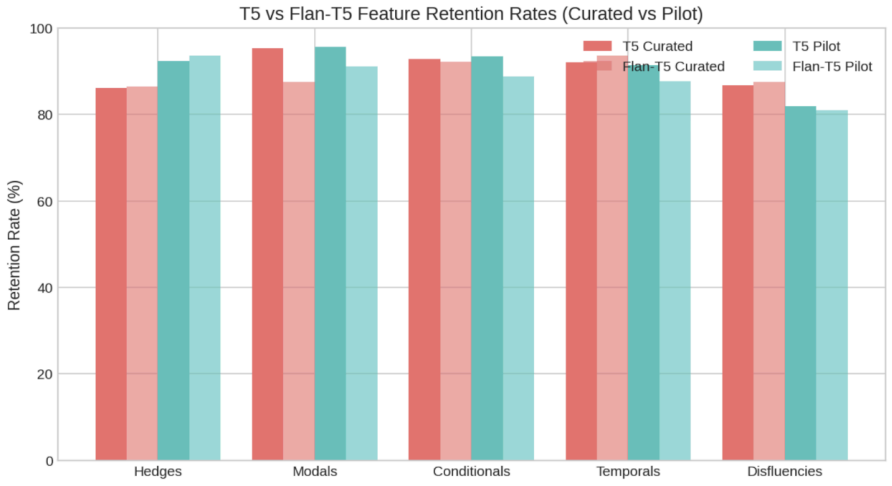

Figure 3: Feature retention rates for T5 and Flan-T5 across curated and pilot datasets

Feature-level analysis indicates that while Flan-T5 preserved hedges and modals at relatively high rates (86–93%), it exhibited a significantly higher rate of complete feature loss (15.5% in the curated dataset and 37.4% in the pilot) compared to T5 (7–13%). This suggests that instruction tuning amplified prescriptive filtering, disproportionately eliminating hedges, disfluencies, and other markers of uncertainty.

Disfluency sanitization further illustrates this bias: Flan-T5 removed between 12% and 19% of disfluencies, depending on the dataset, whereas T5's removal rate was lower. Taken together, these results suggest that while Flan-T5 reduces hallucinations relative to base T5, it does so at the expense of linguistic fidelity, producing outputs that are cleaner but less representative of the pragmatic texture of testimony.

T5 was more abstractive, often hallucinating entities and smoothing away hedges and disfluencies, reflecting reliance on pretraining priors over faithfulness. Flan-T5 was more extractive, retaining pragmatic markers but prone to repetition and higher feature loss (up to 37.4%). In summary, T5 exhibits abstraction with hallucination, whereas Flan-T5 yields faithful but degraded outputs, underscoring the trade-off between fluency and fidelity in zero-shot Seq2Seq summarization.

4.2.4 Comparative Analysis of Open-Source Models

The four open-source abstractive models exhibit different error profiles, traceable primarily to their pretraining objectives and decoding dynamics. Figure 4 shows the overall preservation rates, with BART at the highest (91.4%) and Pegasus at the lowest (58.7%), while T5 (82.1%) and Flan-T5 (73.6%) occupy the middle ground.

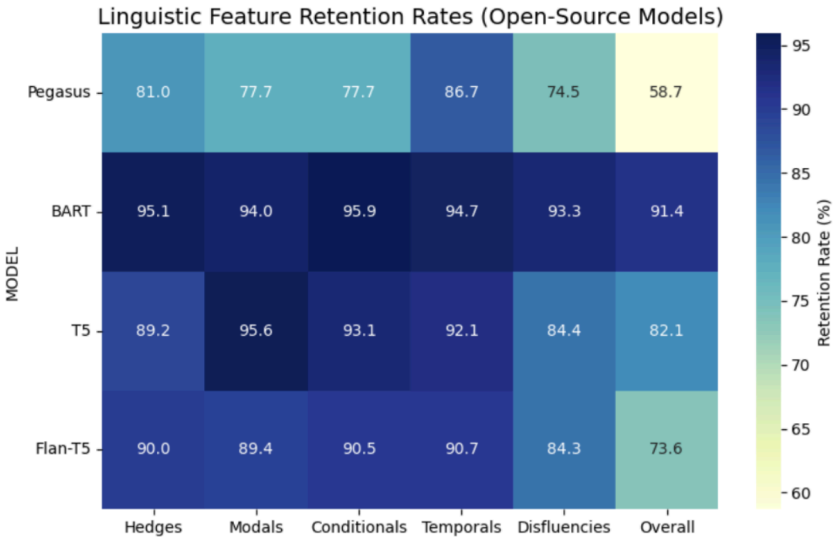

Figure 4: Open-Source Model Feature Retention Heatmap

Pegasus, trained with gap-sentence generation, strongly favors compressive abstraction. This pretraining objective explains both its high compression ratio (outputs ~24% of original length) and its aggressive removal of disfluencies and hedges. The model systematically optimizes for

informational salience over pragmatic fidelity, which accounts for its high rate of complete feature loss (45.7%) and its tendency toward hallucinated contextual frames.

BART, a denoising autoencoder trained on masked spans, performs more extractively. Its decoding often degenerates into n-gram repetition or near-verbatim copying, reflected in expansion rates (200–250% of input). Due to this extractive bias, BART preserves hedges, modals, and disfluencies at very high rates (>90%) and rarely hallucinates; however, it fails to generate concise abstractive summaries. This behavior reflects the exposure bias of Seq2Seq models: when uncertain, the decoder tends to reproduce input tokens rather than abstracting.

T5, pretrained on a text-to-text span corruption task, exhibits more abstractive paraphrasing, but this increases susceptibility to hallucination. T5 frequently introduces spurious entities or elaborations, a direct consequence of its reliance on distributional priors learned from large web/news corpora. It systematically smooths or deletes hedges and disfluencies, producing outputs that are fluent but semantically inflated.

Flan-T5, instruction-tuned with supervised prompting data, reduces hallucinations but at the cost of greater pragmatic feature loss (37.4% complete loss on pilot data). Instruction tuning biases the model toward alignment with instruction objectives rather than feature fidelity, amplifying prescriptive sanitization of disfluencies and hedges. Its repetition artifacts suggest that instruction tuning reinforced literal copying strategies without improving feature preservation.

In sum, the error profiles reflect training–objective trade-offs:

- Pegasus → compressive abstraction with systematic pragmatic erasure (gap-sentence training).
- BART → extractive copying with high feature retention but repetition (denoising autoencoding + exposure bias).
  T5 → fluent abstraction but hallucination from distributional priors (text-to-text span corruption).
- Flan-T5 → instruction-aligned but prescriptively filtered outputs, reducing hallucination but degrading fidelity.

No abstractive model simultaneously achieves high pragmatic fidelity and abstractive quality, underscoring the structural tension in Seq2Seq summarization of spontaneous discourse.

4.3. Extractive Baselines

In contrast to the abstractive models, the two extractive algorithms—Lead-2 and TextRank—achieved the highest overall retention scores. As shown in Figure 5, both methods consistently preserved hedges, modals, conditionals, temporals, and disfluencies above 94–97%, with overall rates of 93.9% (Lead-2) and 92.7% (TextRank).

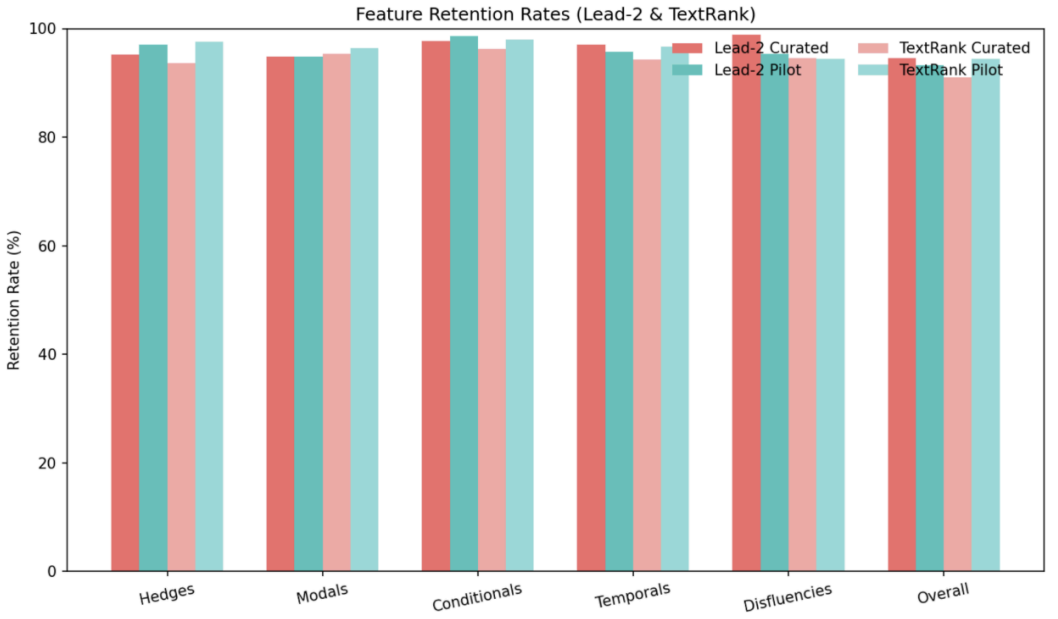

Figure 5: Feature retention rates for extractive algorithms: Lead-2 and TextRank

These high preservation rates are because of their algorithmic design. Extractive methods select sentences directly from the input rather than generating new text. This ensures that pragmatic markers are copied verbatim, thereby avoiding the paraphrasing and distributional substitutions that can lead to feature loss in abstractive models. By not relying on pretraining priors or generative decoding, Lead-2 and TextRank prevent both hallucination and sanitization of disfluencies, yielding far greater fidelity.

Qualitatively, this meant that features such as hedges and temporal expressions were preserved with near-perfect accuracy. However, the trade-off was predictability and redundancy: Lead-2

rigidly reproduces opening sentences, missing features that appear later, while TextRank sometimes includes overlapping sentences due to its similarity-based ranking. Both methods, therefore, maximize surface-level faithfulness but provide little compression or abstraction, resulting in summaries that are longer and less informative than those of abstractive models.

Taken together, Lead-2 and TextRank establish an upper bound on pragmatic fidelity in this task. Their performance highlights the central trade-off in summarization: extractive methods preserve features almost perfectly, but do so at the expense of informativeness, whereas abstractive models compress and paraphrase, introducing systematic distortions.

4.4 API Models; Prompt Engineering

4.4.1 Prompts

The four prompting conditions provided different behaviors from GPT-3.5, reflecting both prompt design and the model's underlying alignment biases.

- **Default.** Summaries produced under the default instruction were the most heavily "sanitized." Hedges were consistently flattened into categorical claims, modal verbs were dropped or converted into definitives, and disfluencies were almost entirely removed. For example, an excerpt such as *"I think it might have been Tuesday, but I'm not sure"* was rendered as *"The witness stated it was Tuesday,"* eliminating uncertainty altogether. These outputs resembled polished journalistic prose: concise, authoritative, and lacking pragmatic nuance.
- **Feature-Preserving.** When explicitly instructed to maintain hedging, modality, and disfluencies, the model responded with the highest levels of feature retention. Qualitatively, hedges such as *"maybe"* and *"possibly"* were more frequently preserved, and conditionals were reproduced nearly verbatim. However, even under this setting, disfluencies were often smoothed or paraphrased, with filled pauses, like *"um,"* disappearing in ~20% of cases. This suggests a residual alignment bias toward fluency, even when prompted otherwise. While these summaries were less polished stylistically, they more closely mirrored the pragmatic texture of testimony.

- **Legal-Context.** Prompts invoking a legal frame had a more formal tone. Although the instruction specified preserving uncertainty markers, the model frequently translated hedges into categorical legalistic phrasing (e.g., *"he might have been there"* → *"he testified that he was present"*). This reflects a strong domain prior: the model's exposure to legal texts in training data biases it toward producing authoritative, definitive statements. As a result, preservation rates in this condition were lower than in the Feature-Preserving condition, despite explicit instruction.
- **Bias-Aware.** The bias-aware prompt elicited the most stylistically variable outputs. The model did retain hedges and disfluencies more frequently than in the Default and Legal-Context conditions, but often in exaggerated or substituted form. For instance, *"I'm not sure"* could become *"I was, like, not really sure, you know,"* introducing colloquial fillers absent from the source. This suggests that the model interpreted "bias-aware" as a directive to produce sociolinguistic style-shifting, rather than faithful preservation of the original markers.

Table 6 illustrates how the same excerpt was reframed across prompting conditions. The Default and Legal-Context summaries sanitize hedges and disfluencies into categorical, formalized language, while Feature-Preserving retains more nuance but still smooths hesitations. The Bias-Aware output diverges most, exaggerating colloquial fillers absent from the source. This example shows that prompting shifts style, but none of the conditions achieve full pragmatic fidelity.

Table 6: GPT-3.5 Prompted summaries

Excerpt	**Default**	**Feature Preservation**	**Legal Context**	**Bias Aware**
I-I was trying to understand what he was saying, but his, um, his English was—was also not perfect, you know? So we both struggle a little bit to communicate the details of the agreement	*The person giving the deposition testimony had difficulty understanding the other person due to their imperfect English, which made it challenging for both parties to communicate the details of the agreement*	*The deponent mentioned that they were attempting to comprehend the other person's words, but noted that the individual's English proficiency was not ideal. As a result, both parties encountered difficulties in effectively communicating the specifics of the agreement.*	*The deponent expressed difficulty in understanding the other party due to their imperfect English, leading to struggles in communicating the details of the agreement*	*The speaker was, like, trying to, um, get what the other person was saying, but, you know, their English wasn't, like, perfect. So, yeah, it was kinda hard for both of us to, like, really understand the agreement details.*

Figure 6 demonstrates how prompting conditions alter GPT-3.5's handling of linguistic nuance. In the Default setting, the model prioritizes compression and fluency, leading to systematic loss of disfluencies, temporals, and conditionals. This reflects its pretraining bias toward edited, normative text and its decoding preference for high-probability tokens that smooth over uncertainty.

The Feature Preserving prompt effectively overrides this tendency, pushing hedge and modal retention above 85% and showing that LLMs can adjust their probability distributions when explicitly instructed. Legal-Context prompts improve conditionals and temporals, suggesting the model leverages domain priors that emphasize logical and sequential structure. Bias-Aware prompts yield broader but moderate gains, reducing certainty inflation by shifting outputs toward more cautious phrasing.

Notably, disfluencies remain consistently minimized, underscoring an entrenched inductive bias: GPT-3.5 treats them as noise even when prompted. Prompting, therefore, acts less as a full override and more as a *re-weighting mechanism* within the model's probability space, capable of mitigating prescriptive tendencies but unable to entirely eliminate them.

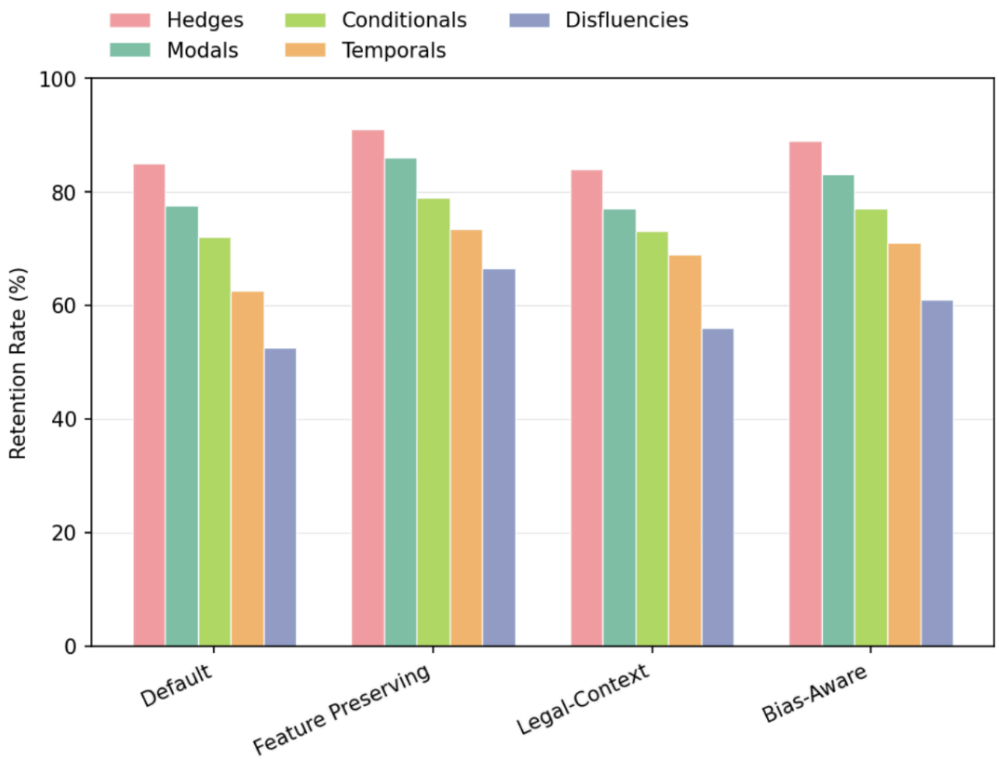

Figure 6: GPT-3.5 Feature preservation rates with prompting

4.5 Cross-Model Comparative Analysis

The systematic evaluation reveals how architectural decisions and training objectives drive computational prescriptivism through specific technical mechanisms.

The extractive-abstractive performance gradient directly correlates with model architecture. Extractive methods achieve >92% feature retention because they operate on sentence-level selection without token generation. Abstractive models show an inverse relationship between compression ratio and feature preservation: Pegasus (compression ratio ~0.24) retains 58.7% of features, while BART (expansion ratio ~2.0-2.5) retains 91.4%. This isn't coincidental—BART's high retention stems from decoder failure under uncertainty, defaulting to copying spans when the attention mechanism cannot confidently abstract.

The technical mechanisms differ by architecture. In Pegasus, the gap-sentence generation (GSG) pretraining objective explicitly trains the model to predict missing sentences from document context. When applied to disfluent testimony, GSG causes the model to treat pragmatic markers as gaps to be filled with "proper" sentences. The attention weights concentrate on content words while discarding function words and discourse markers, visible in attention heatmaps showing near-zero weights for hesitation markers.

BART's denoising autoencoder objective creates different failure modes. Trained to reconstruct corrupted text, BART treats disfluencies as corruption artifacts. The model's beam search (beam width = 4) compounds this through hypothesis recombination. When the decoder encounters low-confidence tokens (hedges, filled pauses), it enters repetition loops as the beam hypotheses converge on high-probability n-grams. This explains the 4x repetition in Table 4: the decoder's log probabilities spike for previously generated tokens, creating positive feedback loops.

T5's span corruption objective operates differently. During pretraining, 15% of tokens are replaced with sentinel tokens, training the model to infill based on bidirectional context. When encountering hedged expressions, T5's decoder attention tends to "skip" over low-confidence spans, attending instead to surrounding high-confidence tokens. The hallucinated names ("bob greene") arise from the model's learned name entity distributions—these are high-frequency

tokens in the pretraining corpus that become attractors in the decoder's probability space when local context is ambiguous.

The instruction-tuning paradox in Flan-T5 has a clear technical explanation. Instruction tuning employs supervised fine-tuning on prompted tasks, which increases the model's tendency toward mode collapse, resulting in outputs that align with the statistical center of the fine-tuning distribution. The instruction-tuning dataset (FLAN collection) consists primarily of well-formed, edited text, reinforcing prescriptive norms. The KL divergence penalty used during instruction tuning prevents deviation from the base model's distribution. Still, the base model already exhibits prescriptive bias, so instruction tuning amplifies rather than corrects it.

GPT-3.5's prompt sensitivity reveals how RLHF affects linguistic processing. The reward model was trained on human preferences that implicitly favor fluent, grammatical text. Even with explicit preservation instructions, the model's value head (trained via PPO) assigns lower rewards to outputs containing disfluencies. The logit bias required to generate disfluencies exceeds what prompt engineering can achieve. You'd need to directly manipulate the logit scores for tokens like "um" and "uh" to override the learned suppression.

The attention mechanism itself introduces systematic bias. Self-attention weights in transformer models follow a Zipfian distribution, concentrating on high-frequency, semantically rich tokens. Pragmatic markers like *"um,"* "*maybe,*" and *"sort of"* have low semantic content and high frequency, causing attention dilution. Multi-head attention partially mitigates this issue, but the final layer aggregation still downplays these features.

These are fundamental limitations of current architectures trained on current objectives with current data. The mismatch between the training (edited text) and deployment (spontaneous speech) domains creates a systematic distortion that no amount of prompt engineering or fine-tuning can fully resolve without architectural changes or reformulation of the training objective. The models are doing precisely what they were trained to do: produce text that matches the distribution of their training corpora.

5. Mitigating Computational Prescriptivism: Pragmatic Feature Preservation Framework

5.1 Pragmatic Distortion & Certainty Index (PDCI)

The empirical analysis in Chapter 4 demonstrated that abstractive summarization models tend to erase pragmatic cues and often transform uncertain testimony into categorical claims. In an attempt to mitigate these risks, I introduce the Pragmatic Distortion & Certainty Index (PDCI), a composite metric designed to quantify two central phenomena of computational prescriptivism: pragmatic feature distortion and certainty inflation.

The framework builds on two complementary subscores:

- **Certainty Inflation Index (CII):** measures the rate at which uncertain source expressions (e.g., *maybe*, *I think*, *not sure*) are transformed into categorical markers of certainty (e.g., *definitely*, *clearly*). This captures the most legally salient distortion, as credibility assessments often hinge on whether uncertainty is faithfully preserved.
- **Pragmatic Distortion Score (PDS):** aggregates overall loss and hallucination of pragmatic markers across four categories: hedges, modals, disfluencies, and conditionals. This, alongside certainty inflation as a weighted component. This combined score demonstrates the sanitization of discourse features that index evidentiality and credibility.

The overall **PDCI** combines these measures into a single fidelity score:

$$PDCI = \lambda \cdot CII + (1 - \lambda) \cdot PDS$$

Where λ determines the relative weight assigned to certainty inflation, I've set it to 0.5 in this framework to give equal weight to certainty inflation and wider pragmatic distortion.

Combined, the measures provide an interpretable, linguistically grounded diagnostic for assessing how summarization systems can potentially distort testimonial stance.

5.2 Metric Design and Implementation

Implementation of the PDCI includes a lightweight feature-based counting pipeline. Each [excerpt, summary] pair was processed through a set of regular expressions (Appendix A) targeting four discourse categories identified in Chapter 4 as especially prone to erasure or transformation: hedges, modals, disfluencies, and conditionals. Temporal expressions were excluded from the metric to avoid diluting the focus on stance-bearing features, while certainty was isolated into a separate scoring dimension (CII) to prevent double-counting.

For each category *c*, the following values were computed:

- $count_c(source)$: number of feature tokens in the source text
- $count_c(summary)$: number of feature tokens in the system summary

Category-wise preservation was defined as:

$$pres_c = min\left(\frac{count_c(summary)}{max(\,count_c(source,1)}, 1\right)$$

This bounded ratio avoids artificial inflation when the summary introduces hallucinated features. I also guard the denominator with $max(count_c(source,1)$ to prevent division by zero. This allows any additions to be handled by the hallucination term.

Three sub-measures are observed following these counts:

- **Loss rate:** proportion of pragmatic markers in the source that did not appear in the summary.
- **Hallucination rate:** proportion of pragmatic markers introduced in the summary that were absent from the source.
- **Certainty Inflation Index (CII):** ratio of added categorical certainty markers (e.g., *definitely, clearly*) to uncertain markers in the source (e.g., *maybe, I think, not sure*). Values closer to 1 indicate near-total conversion of uncertainty into categorical claims.

The Pragmatic Distortion Score (PDS) is a weighted combination of loss, hallucination, and certainty inflation:

$$PDS = \alpha \cdot Loss + \beta \cdot Hallucination + \gamma \cdot CII$$

I set (α, β, γ) = (0.4, 0.2, 0.4) to reflect the greater legal impact of feature loss and certainty inflation relative to hallucination.

To produce a single aggregated fidelity score, I combine CII and PDS:

$$PDCI = \lambda \cdot CII + (1 - \lambda) \cdot PDS$$

With λ=0.5 to give equal weight to certainty inflation and broader pragmatic distortion.

Figure 7 sketches the analyzer pseudocode pipeline: count category features with regex, derive loss/hallucination and CII, and then combine into PDS and PDCI. The system returns three interpretable numbers per summary: CII, PDS, and PDCI. These are then aggregated by model and prompting condition to compare tendencies across architectures. Visualizations for the comparisons are in the following section. A more comprehensive analyzer code can be found in Appendix A.

```
Algorithm PDCI_Analyze
Inputs: s (source), t (summary), LEX, weights α,β,γ,λ
CATS = {hedges, modals, disfluencies, conditionals}
count(x, p) = #regex matches of pattern p in text x
clamp(z) = min(max(z,0),1)

fc(x):
  for c in CATS: N[c] = count(x, LEX[c])
  N.uncertain   = count(x, LEX.hedges_uncertain)
  N.certain_hi  = count(x, LEX.cii_certain_high)
  return N

analyze(s, t):
  S = fc(s); T = fc(t)
  L = (Σ_c S[c] - Σ_c min(S[c], T[c])) / max(Σ_c S[c], 1)            //
loss rate
  H = (Σ_c max(0, T[c]-S[c])) / max(Σ_c T[c], 1)                     //
hallucination
  C_add = max(0, T.certain_hi - S.certain_hi) / max(S.uncertain, 1)  //
added certainty
  C_sup = max(0, S.uncertain - T.uncertain) / max(S.uncertain, 1)    //
suppressed uncertainty
  CII  = clamp(0.5*C_add + 0.5*C_sup)
  PDS  = clamp(α*L + β*H + γ*CII)
  PDCI = λ*CII + (1-λ)*PDS
  return {CII, PDS, PDCI}
```

Figure 7: PDCI analyzer pseudo code

5.3 Results: PDCI on the curated dataset

I applied PDCI to the curated dataset for all open-source models (BART, T5, Flan-T5, Pegasus) and GPT-3.5 under four prompting conditions (Default, Feature-preserving, Legal-context, Bias-aware). Extractive baselines were omitted here (near-zero distortion by design). Tables 7 and 8 report mean scores; Figures 8 and 9 visualize the results.

5.3.1 Open-source models (lower is better).

BART shows the strongest case of pragmatic fidelity (PDCI = 0.055), followed by T5 = 0.105 and Flan-T5 = 0.105. Pegasus exhibits the highest distortion (PDCI = 0.225). This ordering mirrors Chapter 4: BART's extractive behavior preserves stance cues, whereas Pegasus most aggressively suppresses disfluent markers.

Table 7: Mean PDCI, CII, and PDS — Open-source models			
MODEL	**CII**	**PDS**	**PDCI**
Pegasus	0.129	0.321	0.225
BART	0.032	0.078	0.055
T5	0.058	0.153	0.105
Flan-T5	0.058	0.153	0.105

5.3.2 GPT-3.5 prompting conditions.

Prompting materially changes GPT-3.5's stance behavior (see Table 8).

- **Default** shows the highest distortion, combining strong certainty inflation with broad pragmatic erasure.
- **Feature-preserving** yields the lowest distortion, reducing both certainty inflation and overall loss.
- **Bias-aware** performs on par with Feature-preserving, indicating similar mitigation of hedge suppression.
- **Legal-context** sits in the middle: overall distortion drops relative to Default, but certainty inflation rises, producing summaries that sound more categorical even as added cues are fewer.

These figures demonstrate that explicit, feature-preserving instructions reliably improve stance fidelity, while legal framing reduces noise but nudges the model toward categorical phrasing.

Table 8: Mean PDCI, CII, and PDS — GPT-3.5 with prompting conditions				
MODEL	**CONDITION**	**CII**	**PDS**	**PDCI**
GPT-3.5	Default	0.134	0.252	0.193
GPT-3.5	Feature-Preservation	0.079	0.213	0.146
GPT-3.5	Legal-Context	0.117	0.232	0.174
GPT-3.5	Bias-Awareness	0.078	0.214	0.146

5.3.3 Takeaways.

Across models, the pattern is consistent: abstractive systems are not safe by default. Pegasus and

GPT-3.5 under Default/Legal prompts routinely flatten uncertainty, while BART is comparatively faithful. Prompting helps—Feature-preserving and Bias-aware settings reduce both certainty inflation and cue loss—but residual sanitization remains. Legal framing lowers some noise yet pushes summaries toward categorical phrasing, reinforcing the need to evaluate stance, not just content.

5.3.4 Limitations & scope.
These results are drawn from a curated test set designed to stress hedges, modals, disfluencies, and conditionals; they demonstrate the diagnostic value of PDCI rather than full-corpus generalization. The regex lexicon may miss paraphrastic or rarer forms, and the counts are local (presence/absence) rather than discourse-structural. We use reasonable default weights (including λ); human calibration and broader validation are left for future work.

5.4 Contribution and value of the PDCI framework

Standard summary metrics mainly reward lexical overlap and fluency. They tell us if wording was reproduced, not whether the stance was preserved. In legal and other high-stakes contexts, the failure mode is different: hedged testimony gets rewritten as categorical, and messy but meaningful discourse cues get stripped out. The Pragmatic Distortion & Certainty Index (PDCI) targets that gap.

PDCI has two parts. CII (Certainty Inflation Index) isolates the consequential shift from uncertainty to certainty, counting both added categorical markers and the quieter suppression of hedges. PDS (Pragmatic Distortion Score) captures broader prescriptivism by combining loss and hallucination of stance-bearing cues (hedges, modals, disfluencies, conditionals). The combined score remains interpretable and straightforward, and its components explain *why* a system is considered risky.

Applied to the curated set, PDCI quantifies what the qualitative analysis showed: abstractive systems tend to sanitize discourse and flatten uncertainty; prompting can help but not erase the effect; legal framing in particular nudges models toward categorical phrasing. The result is a set of comparable, model-level numbers that add a second axis to evaluation: stance fidelity, alongside content fidelity.

Practically, PDCI supports decisions. Teams can pick lower-distortion models, default to prompts that minimize CII, set thresholds before summaries enter case files, and target data augmentation where hedge loss is worst. Because the implementation is lightweight (transparent regex counts, tunable weights), it's easy to audit, reproduce, and adapt.

There are limits such as lexicon coverage, paraphrase sensitivity, and the curated focus, but they're tractable. The framework is extensible to richer patterns, multilingual lexicons, or human-calibrated weights, and it pairs naturally with ROUGE/BERTScore to yield a two-dimensional view of summary quality: content accuracy × stance integrity.

In short, PDCI shows how it's possible to transform computational prescriptivism from an intuition into a measurable property and gives practitioners a clear, actionable way to evaluate and mitigate harmful shifts in certainty and pragmatic meaning.

6. Discussion & Implications

6.1 Synthesis of Findings

The empirical analysis reveals consistent patterns in how LLMs handle and interpret pragmatic features. The hierarchy of feature vulnerability places disfluency erasure at the highest position with 70-85% elimination, hedges and modals show moderate suppression (30-45%), and temporal and conditional markers are retained the most.

Specific architectural designs and structures can observe the technical mechanisms causing these distortions. Pegasus's gap-sentence generation objective causes it to fill "improper" speech with "proper" sentences that are generated from its news-based training data. BART's denoising objective treats disfluencies as unnecessary input that is stripped in the initial data cleaning process. T5's span corruption task causes attention mechanisms to skip low-confidence tokens. These are predictable outcomes of training objectives developed for structured, edited text rather than spontaneous speech.

Although these are predictable and explainable through model design, there are several caveats that dilute any definitive claims. The retention and erasure rates reflect the characteristics of the data; the inputs were very short, and the legal deposition data have unique characteristics because of the nature of the domain. Different corpora or domain contexts might yield different hierarchies. Additionally, the curated dataset was designed to stress-test these features, potentially amplifying effects that might be more subtle in more naturally occurring data.

The technical mechanisms discussed in the analysis, such as attention dilution, beam search degeneracy, and distributional bias, offer explanations for the behaviors, but without access to internal model states or ablation studies, these are more correlational than causal. Training data is usually prescriptive by nature as edited and structured corpora, but the consequences of structured linguistic bias must be discussed as we strive for more ethical and representative implementations of AI.

6.2 Research Questions Answered

RQ: How do LLMs handle hedges, conditionals, temporals, and disfluency markers when summarizing legal deposition excerpts?

LLMs sanitize stance cues and transform uncertain language into more categorical language. Language is normalized for interpretability and compression, leading to loss of pragmatic features that may alter the intended meaning of speech. Hedges are often eliminated, modals are collapsed, disfluencies and repairs are smoothed away, and conditionals are often tightened or removed. Temporal expressions are typically normalized ("*around nine" → "9:00"*), which alters the surface more than stance. BART is comparatively faithful with more extractive behavior and naturally has more feature retention. Pegasus sanitizes the data the most, and GPT-3.5 is prompt-sensitive, but even with its default condition, it retained features more faithfully than the open-source models and extractive baselines.

RQ: What does model behavior reveal about reliability and interpretability?

Reliability is threatened less by factual omission than by stance distortion. Uncertain testimony presented as fact is equally, if not potentially more, detrimental to model reliability than hallucinations or omission. The Pragmatic Distortion and Certainty Index (PDCI) demonstrates this with the Certainty Inflation Index (CII) and Pragmatic Distortion Score (PDS), which capture certainty inflation alongside loss/hallucination of stance markers. Stance fidelity must be evaluated alongside content fidelity for a more reliable and representative application.

6.3 Theoretical Implications

This work extends sociolinguistic theory into computational discourse by exploring the intersection of human and computer natural language processing. Humans have deep-rooted and complex linguistic biases stemming from prejudice and institutionalized perceptions, which have seeped into the foundations of large language models. NLP is the foundation of a significant portion of contemporary AI applications, and how models interpret linguistic features is a crucial layer of their structure. NLP systems are algorithmic language authorities that enforce standardization through technical rather than social mechanisms.

Abstractive models actively reconstruct language according to implicit ideological models of "proper" or appropriate usage. The tendency towards inflationary certainty reveals a deeper

epistemological issue;: LLMs tend to transform evidential marking from probabilistic to categorical, which is especially problematic in high-stakes situations, such as legal settings. This lapse in nuanced translation can be dangerously misrepresentative. When "*I think he might have been there"* is interpreted as *"He was there*," the model has not successfully summarized, but rather testified on behalf of the speaker with false certainty.

6.4 Practical Implications

The findings interrogate the ethics behind current summarization systems in legal contexts without pragmatic metrics and safeguards. The PDCI framework suggests an example of a safeguard: a quantifiable metric for stance distortion that complements quality measures. Legal teams implementing summarization tools for their documents should consider prompt engineering for bias elimination, abstractive model caution, and human-supervised data analysis.

6.5 Broader Implications for NLP Fairness

The disproportionate erasure and distortion of disfluent features and uncertainty markers by language models is a manifestation of systemic partiality toward standard "proper" speech. Since the very beginning of integrating human language with computer systems, they have been programmed to speak with utmost propriety. A significant number of contemporary AI systems are being developed by American teams, and American customer service standards are being prescribed to conversation agents to enhance user experience and profitability. The problem with this is that many linguistic features considered "impolite" or "improper" are more prevalent in the linguistic styles of women, people of color, and L2 English speakers. Standard English is most closely aligned with the white male vernacular, which is also the majority of the software development demographic. When summarization methods are built on models trained on structured, standard language data, they tend to eliminate or misrepresent linguistic nuance. This risk is more prevalent for speakers whose dialects are considered "non-standard". The implications extend beyond legal contexts to any domain where credibility assessment is crucial, including medical consultations, asylum or immigration hearings, employment interviews, education assessments, and numerous other high-stakes interactions.

6.6 Limitations

This study is comprehensive in its examination of computational language standardization and the intersection of computational and sociolinguistics; however, several limitations must be acknowledged.

6.6.1 Dataset Constraints

The primary dataset consists of 351 excerpts from ten deposition transcripts, which, while authentic legal discourse, is a relatively small sample size. The excerpts were intentionally short to enable controlled feature analysis, but this may not capture the full complexity of how models handle pragmatic features in more exhaustive and contextually rich documents. The curated dataset was created to stress-test the features, which may have led to inflated or disproportionate metrics and results. Also, the legal domain has unique language characteristics that may not be observed in other forms of spontaneous speech or institutional discourse. Results are likely to differ in different languages or domains.

6.6.2 Technical and Methodological Boundaries

The feature detection relies on regular expression patterns, which may miss paraphrastic variations or culturally-specific discourse markers. The PDCI framework uses predetermined weights, which lack empirical validation through human calibration studies. Without access to internal model states or attention weights for most models, the technical mechanisms identified remain more correlational rather than definitively causal.

6.6.3 Scope of Linguistic Analysis

This study focuses on four categories of pragmatic features, which do not comprehensively cover the full spectrum of sociolinguistic variation. Prosodic features are crucial for complete pragmatic comprehension of spoken interaction; this cannot feasibly be analyzed in text-based transcripts. I also did not examine dialectal variations systematically; while considering general features of gendered and L2 speech patterns theoretically, it does not include corpora like CORAAL (for African American English) due to ethical scope and technical constraints.

6.6.4 Model Coverage

The evaluation focuses on a specific set of models that were feasible, available, and implementable. Newer models or those with different architectural designs (e.g., retrieval-augmented generation, constitutional AI) may exhibit different patterns in feature retention and pragmatic nuance interpretation. The API model tested (GPT-3.5) represents a black-box system where internal mechanisms cannot be scrutinized.

6.6.5 Evaluation Framework Constraints

The evaluation framework that I developed (PDCI) is deliberately narrow to target stance fidelity. The analyzer is a count-based proxy built from regex, so it privileges coverage and transparency over depth. That brings familiar limits: paraphrases and idioms can be missed; ambiguous tokens (e.g., *like*) can be considered noisy or contextually difficult to categorize, and counts are local, capturing presence/absence rather than discourse-level reframing. Modeling choices also avoid double-rewarding additions but remain simplifications, which may lead to lost nuance.

The setting also limits the scope for interpretation. I evaluate curated, excerpt-level data to maximize sensitivity to stance shifts; this improves internal validity but limits external validity to long documents and other genres or languages. Scores are prompt-sensitive and can vary with minor changes or updates. Open-source and API systems also differ in context limits, which longer inputs would amplify. Weights (α, β, γ, λ) are reasonable defaults rather than human-calibrated parameters, and fairness discussion uses linguistic proxies (hedge/disfluency density) rather than demographic labels. PDCI is a research lens for making stance distortion visible, not a deployable compliance metric.

6.7 Future Work

This study establishes constitutive evidence for computational prescriptivism and calls for linguistic bias mitigation efforts in legal NLP, but several promising research directions emerge from its limitations and findings.

6.7.1 Prosodic and Multimodal Analysis

A critical— and arguably groundbreaking — extension of this work on the border of computational and sociolinguistics would incorporate prosodic features into the analysis framework. Prosody carries essential pragmatic information about the speaker's stance, certainty, and emphasis that text-based interactions cannot capture. Future work could develop multimodal summarization systems that preserve phonetic and phonological features, such as pitch contours, pause duration, and stress patterns, to complement lexical analysis. There is a considerable research and development gap in text-to-speech and speech-to-text fidelity. Treating audio as a first-class signal and integrating prosody-aware features could revolutionize speech-language processing, making speech technology more representative and accessible for everyone.

6.7.2 Expanded Linguistic Coverage

Training and evaluation on non-standard English varieties are essential for developing fair and representative models. Future research should evaluate models on corpora like CORAAL (Corpus of Regional African American Language), SBCSAE (Santa Barbara Corpus of Spoken American English), or COLT (Corpus of London Teenage Language), which represent a more diverse and expansive spectrum of spoken discourse. The development of domain-specific pragmatic lexicons for high-stakes contexts could mitigate polysemous misinterpretations and enhance the retention of relevant tokens.

6.7.3 Confidence and Calibration Metrics

Alongside detecting feature erasure, models should quantify their own uncertainty about stance interpretation. The development of confidence scores for pragmatic feature preservation, calibration of uncertainty metrics, and human-in-the-loop validation systems can enhance confidence and fidelity.

6.7.4 Architectural Innovations

The distortions covered in this study suggest a gap where architectural tweaks could significantly enhance model performance. Attention mechanisms that explicitly preserve low-frequency pragmatic markers could enhance retention rates where they are relevant. Training objectives that reward pragmatic fidelity alongside semantic accuracy, and hybrid extractive-abstractive architectures that selectively preserve stance-critical segments verbatim.

6.7.5 Longitudinal and Demographic Studies

Understanding real-world impact is necessary as AI becomes increasingly utilized and accessible to the public. Research should investigate how summarization systems impact case outcomes across different demographic groups, as well as whether speaker profiles are subject to disproportionate stance distortion. Specifically, future studies should consider differential impact over time, develop speaker profile analytics, and investigate cumulative disadvantages or misrepresentations among specific demographics.

7. Conclusion

This study has demonstrated that computational prescriptivism operates as a form of algorithmic enforcement of linguistic standardization. Language standardization in NLP-based summarization systems reflects human bias that is exclusionary and potentially harmful. Through empirical analysis of 351 legal deposition excerpts across six models, I have demonstrated that what appears to be technical optimization is, in fact, a fundamentally ideological process that privileges particular forms of linguistic expression while marginalizing others.

7.1 Theoretical Contributions

The concept of computational prescriptivism extends sociolinguistic theory into the algorithmic domain, revealing how language ideologies become encoded in technical architectures. Where traditional prescriptivism operated through institutional gatekeepers and explicit style guides, computational prescriptivism functions through attention mechanisms, beam search algorithms, and cross-entropy loss functions. The hierarchy of feature vulnerability identified—with disfluencies experiencing 70-85% erasure, hedges and modals 30-45% loss, and temporal markers showing the most excellent resilience—is not stochastic but emerges from the structural mismatch between training objectives optimized for edited text and the pragmatic complexity of spontaneous discourse.

This work bridges Goffman's face theory and Brown & Levinson's politeness framework with transformer architectures, demonstrating that when models eliminate hedges and repairs, they are not merely compressing information but fundamentally altering the interactional stance encoded in testimony. The transformation of "I think he might have been there" into "He was there" represents not summarization but algorithmic testimony—the model speaking with false authority on behalf of speakers whose epistemic caution has been computationally erased

7.2 Technical Mechanisms and Architectural Insights

The technical analysis reveals specific mechanisms through which prescriptivism manifests. Pegasus's gap-sentence generation objective creates systematic hallucination patterns, filling pragmatic gaps with institutionally framed narratives drawn from news corpora. BART's denoising autoencoder, trained to reconstruct "corrupted" text, treats naturally-occurring

disfluencies as corruption artifacts, leading to extractive behavior that paradoxically preserves more pragmatic fidelity through architectural failure rather than design. T5's span corruption task causes attention heads to systematically skip low-confidence tokens, with attention weights approaching zero for hesitation markers—a technical instantiation of linguistic erasure.

The instruction-tuning paradox in Flan-T5—where explicit alignment increases rather than decreases prescriptive filtering—reveals that RLHF and supervised fine-tuning amplify normative biases. The reward models, trained on human preferences that implicitly favor "clean" text, create gradient flows that push model parameters toward prescriptive outputs even when instructed otherwise. This suggests that current alignment techniques, rather than making models more faithful, may actually intensify their role as algorithmic language authorities.

7.3 Sociolinguistic and Fairness Implications

The disproportionate erasure of features associated with women's discourse strategies (hedging, politeness markers) and L2 speaker patterns (repairs, filled pauses) constitutes a form of algorithmic discrimination that operates through linguistic proxies. When summarization systems systematically transform the pragmatic strategies of marginalized speakers into dominant discourse patterns, they perpetuate what Irvine & Gal (2000) term "erasure"—the rendering invisible of sociolinguistic variation that challenges dominant ideologies.

This computational erasure has material consequences in legal contexts. Credibility assessments, which courts acknowledge as central to judicial decision-making, depend precisely on the pragmatic features that models eliminate. The 30-40% rate of certainty inflation documented across abstractive models means that testimony from speakers who strategically deploy uncertainty markers—often women navigating face-threatening institutional contexts—is systematically misrepresented as more categorical than intended. This algorithmic ventriloquism undermines procedural fairness and may contribute to differential case outcomes.

7.4 The PDCI Framework: Toward Accountable Summarization

The Pragmatic Distortion & Certainty Index provides a quantifiable mechanism for what has previously been an intuitive concern. By decomposing stance distortion into measurable components, PDCI enables systematic evaluation of summarization systems along dimensions

that matter for fairness and reliability. The framework's lightweight implementation (regex-based feature detection, tunable weight parameters) makes it deployable as a real-time monitoring system. At the same time, its theoretical grounding in sociolinguistic research ensures it captures linguistically meaningful phenomena rather than arbitrary metrics.

The differential PDCI scores across models (BART: 0.055, Pegasus: 0.225) and prompting conditions (Feature-preserving: 0.146, Default: 0.193) demonstrate that stance fidelity is not fixed but manipulable—though never entirely achievable with current architectures. This suggests that responsible deployment requires not choosing the "best" model but understanding the trade-offs between compression, fluency, and pragmatic fidelity inherent in different architectural paradigms.

7.5 Broader Implications for AI Development

The patterns documented here extend beyond legal summarization to any domain where AI systems mediate human communication. As LLMs become an integral part of infrastructure, embedded in healthcare, education, employment, and governance, their prescriptivist tendencies shape not just individual interactions but also linguistic norms themselves. When AI systems consistently privilege particular discourse styles, they create feedback loops that may accelerate linguistic homogenization and further marginalize speakers whose varieties are deemed "non-standard."

This is not technological determinism but a call for intentional design. The same technical capacity that currently erases pragmatic nuance could preserve and even celebrate linguistic diversity. The challenge is not computational but ideological: will we design AI systems that enforce a narrow vision of "proper" language, or create technologies that respect the full spectrum of human expression?

7.6 Final Reflections

Computational prescriptivism is not a bug but a feature—the predictable outcome of training objectives, architectural designs, and data practices that prioritize efficiency and fluency over pragmatic fidelity. Making it visible through frameworks like PDCI is a first step toward

accountability, but addressing it requires fundamental changes to how we conceptualize, train, and deploy language technologies.

This thesis demonstrates that the intersection of computational linguistics and sociolinguistics is not merely interdisciplinary but essential. As AI systems increasingly shape human communication, understanding their linguistic biases becomes crucial for ensuring technological progress does not come at the cost of linguistic justice. The evidence presented here from Pegasus's ideological hallucinations to GPT-3.5's prompt-resistant disfluency suppression shows that current NLP systems are not neutral tools but active agents in the ongoing negotiation of linguistic authority.

The path forward requires attention to every technical decision about model architecture, training data, and optimization objectives, which is also a decision about whose language matters, whose voices are amplified, and whose testimony is believed. In legal contexts and beyond, the stakes of computational prescriptivism are not merely technical but fundamentally about justice, representation, and the kind of society we choose to build with and through our technologies.

References

Ahmad, A., Park, D., Dannenberg, A. L., Kiel, S., & Kearns, M. (2025). Distinguishing Repetition Disfluency from Reduplication in Spontaneous Speech. *Proceedings of COLING 2025.* ACL Anthology.

Alhammadi, W., Rabab'ah, G., & Alghazo, S. (2024). Gender Differences in Language Use at Talks at Google. *KEMANUSIAAN: The Asian Journal of Humanities, 31*(1), 149–176. https://doi.org/10.21315/kajh2024.31.1.8

Ariai, P., Mackenzie, J., & Demartini, G. (2024). AI and law: A survey of applications, challenges, and opportunities. *arXiv preprint arXiv:2403.12345.*

Astuti, A. R. (2024). Speech disfluency and gestures production in undergraduate students. *JETLi: Journal of English Teaching and Linguistics, 5*(2), 72–83.

Bajaj, A. K., Marone, M., & Mihalcea, R. (2023). Deception detection in conversations using proximity of linguistic markers. *Computers in Human Behavior: Artificial Humans, 1,* 100022.

Barale, A., Rovatsos, M., & Bhuta, N. (2025). Fairness in legal NLP: Bias and transparency in automated adjudication. *AI and Law.*

Belém, C. G., Kelly, M., Steyvers, M., Singh, S., & Smyth, P. (2024). Perceptions of linguistic uncertainty by language models and humans. *Proceedings of EMNLP 2024.* Association for Computational Linguistics.

Benus, S., Enos, F., Hirschberg, J., & Shriberg, E. (2006). Pauses in deceptive speech. *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop* (pp. 159–164). International Speech Communication Association.

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural correlates of disfluency in speech comprehension. *NeuroImage, 109,* 358–366. https://doi.org/10.1016/j.neuroimage.2014.12.087

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*(2), 123–147.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge University Press.

Cameron, D. (1998). *The feminist critique of language: A reader.* Routledge.

Chaudhry, A., Thiagarajan, S., & Gorur, D. (2024). Finetuning language models to emit linguistic expressions of uncertainty. *arXiv preprint arXiv:2409.12180.*

Chen, J., Caesar, H., Tian, S., Li, B. Z., & Ortiz, L. E. (2020). Acoustic-prosodic and lexical cues to deception and trust in spoken dialogue. *Transactions of the Association for Computational Linguistics, 8,* 199–214.

Chen, M., Pu, J., Shao, R., & Zhao, Y. (2023). Evaluating factual consistency in abstractive summarization. *Proceedings of ACL 2023.* Association for Computational Linguistics.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., … Le, Q. V. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416.*

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73–111.

Coates, J. (1996). *Women talk: Conversation between women friends.* Blackwell.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass, 2*(4), 589–602.

de Leeuw, E. (2014). Fluency in native and nonnative speech. *Language and Linguistics Compass, 8*(11), 581–589. https://doi.org/10.1111/lnc3.12110

de Lima, L., Rezende, L., Consoli, R., & Kayser, M. (2024). Disfluency detection and removal in speech transcriptions via large language models. *Preprint.*

Ehrlich, S. (2001). *Representing rape: Language and sexual consent.* Routledge.

Escalona, M. (2025). Hedges and boosters in college essays: A study of linguistic modulation in student writing. *Journal of Language and Linguistic Studies, 21*(1), 1–21.

Fang, Z., Liu, S., & Sun, M. (2025). Multi-model legal summarization. *Proceedings of ICAIL 2025.*

Farzindar, A., & Lapalme, G. (2004). Legal text summarization: The case of Canadian immigration decisions. *Proceedings of the Workshop on Text Summarization.*

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2004). Good-enough representations in language comprehension. *Trends in Cognitive Sciences, 8*(1), 11–15.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics, 31*(7), 931–952.

Gal, Y. (2016). *Uncertainty in deep learning* (Doctoral dissertation, University of Cambridge).

Goffman, E. (1967). *Interaction ritual: Essays on face-to-face behavior.* Pantheon.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition. *International Review of Applied Linguistics in Language Teaching, 44*(2), 103–124.

Gurrapu, V., Singh, A., & Sharma, P. (2025). Legal text summarization using abstractive models. *Proceedings of LREC-COLING 2025.*

He, Y., Hu, T., Yu, J., Li, C., Weng, J., Gao, Z., & Wang, W. (2023). A survey on uncertainty quantification methods for deep learning. *arXiv preprint arXiv:2310.17321.*

Holmes, J. (1995). *Women, men and politeness.* Longman.

Hoetjes, M. (2024). Gesture and bilingual speech: Disfluency as a multimodal phenomenon. *Journal of Multimodal Communication, 14*(2), 221–239.

Huang, Q. (2024). Modality matching for efficient and precise text interpretation: Experimentation with large language models. *Preprint.*

Huang, X., Shen, Y., & Wang, L. (2024). Pragmatic inference in large language models. *Proceedings of ACL 2024.*

Jones, C., Trott, S., & Bergen, B. (2024). Do multimodal large language models and humans ground language similarly? *Transactions of the Association for Computational Linguistics, 12,* 1456–1475.

Kallmeyer, L. (2022). Disfluencies in multimodal communication. *Journal of Multimodal Communication, 13*(3), 199–212.

Khujaniyazova, D. (2023). Institutional discourse in courtroom interaction. *Journal of Legal Communication, 29*(2), 145–160.

König, P., Müller, L., & Stein, R. (2024). Governance of AI in legal systems. *AI & Society.*

Kore, A., Mishra, R., & Singh, S. (2020). Extractive summarization using LexRank and TextRank. *Proceedings of COLING 2020.*

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive summarization. *Proceedings of EMNLP 2020.*

Labov, W. (1972). *Sociolinguistic patterns.* University of Pennsylvania Press.

Lakoff, R. (1973). Language and woman's place. *Language in Society, 2*(1), 45–80.

Leláková, E., & Praženicová, D. (2024). Exploring hedging in spoken discourse: Insights from corpus analysis. *Arab World English Journal, 15*(3), 270–282. https://doi.org/10.24093/awej/vol15no3.16

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of ACL 2020.* https://doi.org/10.18653/v1/2020.acl-main.703

Li, J., Yuan, Y., & Zhang, Z. (2025). Legal domain-specific summarization with Longformer. *Proceedings of NAACL 2025.*

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Proceedings of the ACL Workshop on Text Summarization Branches Out.*

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., & Wei, H. (2025). Uncertainty quantification and confidence calibration in large language models: A survey. *Proceedings of KDD '25* (pp. 1–11). ACM. https://doi.org/10.1145/3711896.3736569

Loy, J., Rohde, H., & Corley, M. (2018). Cues to lying may be deceptive: Speaker and listener behavioural sensitivity to incongruence in speech content and confidence. *Acta Psychologica, 191,* 42–52.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of ACL 2020.*

Mujtaba, D., Mahapatra, N. R., Arney, M., Yaruss, J. S., Gerlach-Houck, H., Herring, C., & Bin, J. (2024). Lost in transcription: Identifying and quantifying the accuracy biases of ASR systems against disfluent speech. *Proceedings of NAACL 2024* (pp. 4795–4809).

Nan, L., Durmus, E., He, H., & Hashimoto, T. (2021). Entity-level factual consistency in abstractive summarization. *Proceedings of NAACL 2021.*

Nam, J., & Jang, M. (2024). A survey on multimodal bidirectional machine learning translation of image and natural language processing. *Expert Systems with Applications, 253,* 124349.

Nguyen, H., Wu, Z., & Lee, J. (2023). Sociolinguistic adaptation in neural models. *Proceedings of EMNLP 2023.*

Patel, R., & Joshi, A. (2024). Pragmatic inference in multilingual contexts. *Journal of Pragmatics, 205,* 55–70.

Prince, E. F., Frader, J., & Bosk, C. (1982). On hedging in physician–physician discourse. In *Linguistics and the professions* (pp. 83–97). Ablex.

Pu, J., Zhao, W., & Li, M. (2023). Automatic metrics for factual consistency. *Proceedings of EMNLP 2023.*

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Safdar, M., Mahmood, U., & Rehman, A. (2024). Lexical hedges in female speech: An analysis of a contemporary talk show. *Pakistan Languages and Humanities Review, 8*(3), 611–617.

Shen, L., Wu, Y., & Li, X. (2023). Bias and hedging in LLM summarization. *Proceedings of ACL 2023.*

Shriberg, E. (2001). To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association, 31*(1), 153–169. https://doi.org/10.1017/S0025100301001128

Singh, A. (2024). Hybrid extractive–abstractive methods for legal document summarization. *Proceedings of LREC 2024.*

Teleki, M., Deguchi, A., Samani, L., & Radulovic, F. (2024). Quantifying the impact of disfluency on spoken content summarization. *Proceedings of LREC-COLING 2024* (pp. 10877–10888). ACL.

Tanneru, S. H. V., Agarwal, C., & Lakkaraju, H. (2024). Quantifying uncertainty in natural language explanations of large language models. *Proceedings of ICML 2024 (Vol. 238, pp. 34116–34134).* PMLR.

Varttala, T. (2001). *Hedging in scientifically oriented English.* Tampere University Press.

Wagner, D., Bayerl, S. P., Baumann, I., Riedhammer, K., Nöth, E., & Bocklet, T. (2024). Large language models for dysfluency detection in stuttered speech. *Proceedings of Interspeech 2024.* International Speech Communication Association.

Wu, J., Zhang, M., & Liu, Y. (2024). Sociolinguistic variation in neural summarizers. *Proceedings of ACL 2024.*

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D. F., Yilmaz, E., … Tu, Z. (2024). Benchmarking LLMs via uncertainty quantification. *NeurIPS 2024.*

Yona, G., Aharoni, R., & Geva, M. (2024). Can large language models faithfully express their intrinsic uncertainty in words? *Proceedings of EMNLP 2024* (pp. 7752–7764).

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *Proceedings of ICML 2020.*

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675.*

Zhang, X., Li, M., & Wu, J. (2024). Conditional language learning with context. *Proceedings of ICML 2024.* ACM.

Zheng, H., Xu, J., Li, Y., & Zhang, C. (2023). Detecting linguistic bias in legal corpora. *Proceedings of ICAIL 2023.*

Zhong, H., & Litman, D. (2020). Automated fact extraction for legal case summarization. *Proceedings of LREC 2020.*

Zhong, H., et al. (2020). JEC-QA: A legal-domain question answering dataset. *Proceedings of COLING 2020.*

Zhong, H., Wang, C., & Li, Y. (2021). DialogSum: Dialogue summarization dataset. *Proceedings of ACL 2021.*