

Appendix A.

1. Coding and Analysis Procedures

The pilot corpus consisted of **162 model outputs** (six ethical dilemmas × nine deictic framings × three LLMs). Each response was automatically labeled for **ethical reasoning type** using a **deterministic rule-based classifier** implemented in **Python 3.11**. The classifier identified characteristic lexical markers associated with **consequentialist**, **deontological**, **virtue-based**, **care-based**, and **reflexive** reasoning (e.g., *maximize*, *principle*, *virtue*, *compassion*). Each response was then assigned a dominant reasoning type based on **highest keyword frequency** and **contextual weighting**.

All texts were **tokenized and lowercased** prior to analysis. Because the classification process was entirely **algorithmic and deterministic**, traditional inter-coder reliability measures (e.g., **Cohen’s κ**) were not applicable. Cases lacking sufficient lexical indicators were labeled **“mixed/unspecified.”**

Subsequent analyses were conducted in **Python 3.11**, using the **pandas**, **scipy.stats**, and **statsmodels** libraries. Given the exploratory scope and small sample size, these findings should be interpreted as **diagnostic evidence of systematic tendencies**, rather than as **inferential or predictive statistics**.

Ethical Framework	Key Lexical Indicators	Example Phrases
Consequentialist	“utilitarian,” “maximize,” “minimize harm,” “save more lives,” “greater good,” “outcome,” “consequence”	“maximize benefit,” “minimize overall suffering”
Deontological	“duty,” “principle,” “rule,” “obligation,” “moral law,” “right/wrong regardless,” “Kant”	“it would be my duty,” “against moral law”
Virtue Ethics	“virtue,” “character,” “integrity,” “moral core,” “flourish,” “courage,” “honesty”	“act with integrity,” “reflect moral character”
Care Ethics	“care,” “compassion,” “empathy,” “relationship,” “contextual,” “responsibility for others”	“act with empathy,” “care for the individual”

2. Chi-Square Analysis and Illustrative Confidence Intervals

Model	χ^2	df	p-value	Interpretation
Anthropic Claude 3.5 Sonnet	37.36	24	0.040	Significant variation across deictic framings
GPT-4o	11.34	16	0.788	No significant variation, moderately homogeneous

Model	χ^2	df	p-value	Interpretation
DeepSeek	7.27	8	0.508	No significant variation, highly uniform consequentialism

This supplementary **chi-square analysis** tested whether the distribution of ethical reasoning types differed significantly across **deictic framings** within each model. Results indicated **moderate to significant variation** for *Claude 3.5 Sonnet* ($\chi^2(24, N = 54) = 37.36, p = .04$), but **no significant effects** for *GPT-4o* ($\chi^2(16, N = 54) = 11.34, p = .79$) or *DeepSeek-Chat* ($\chi^2(8, N = 54) = 7.27, p = .51$). These outcomes suggest that **deictic framing reorganized justificatory style most strongly in Claude**, weakly in GPT-4o, and minimally in DeepSeek. However, with roughly six observations per frame per model, such tests have **limited statistical power**.

For this reason, these inferential results were treated as **diagnostic rather than confirmatory** and were not integrated into the main discussion. The study's objective was not to establish stable behavioral differences between models, but to **observe how grammatical framing reorganizes reasoning** across contexts. Inferential tests in a small, stylistically heterogeneous corpus risk overstating precision where variation is primarily **linguistic and relational**. The chi-square analysis thus functions as a **plausibility check**, supporting the qualitative finding that deixis operates as a grammatical operator structuring ethical stance. Larger and more balanced datasets would be needed to determine whether these framing effects persist across **topics, languages, or multi-turn dialogues**.

Consequentialist Reasoning by Frame (95%, CI)

Framing	Claude%, [CI]	GPT-4o%, [CI]	DeepSeek%, [CI]
Cosmological	50 [19–81]	100 [61–100]	83 [44–97]
Dialogic	50 [19–81]	67 [30–90]	100 [61–100]
First-person	33 [10–70]	83 [44–97]	100 [61–100]
First-person plural	100 [61–100]	83 [44–97]	100 [61–100]
Impersonal	50 [19–81]	83 [44–97]	100 [61–100]
Reflexive	17 [3–56]	83 [44–97]	100 [61–100]
Second-person	17 [3–56]	83 [44–97]	83 [44–97]
Spatial	67 [30–90]	67 [30–90]	100 [61–100]
Temporal	67 [30–90]	67 [30–90]	100 [61–100]

Consequentialist reasoning was most frequent under **first-person plural (100 %)**, **spatial (67 %)**, and **temporal (67 %)** framings, and least common under **reflexive** and **second-person** framings (17 %). The wide intervals reflect the small number of observations per frame ($n \approx 6$), making these results **indicative but not statistically stable**.

Summary Diagnostic

Test	df	χ^2	p-value
Deictic framing × reasoning type	24	37.36	0.040

Differences in reasoning style across deictic framings reached **statistical significance** ($p < .05$), confirming that framing (first-person, second-person, reflexive, etc.) systematically influenced the ethical reasoning generated. However, with small and uneven cell counts, the results are best read as **diagnostic**, not **confirmatory**. This example illustrates how **quantitative uncertainty** can be represented using **95 % Wilson confidence intervals** to estimate plausible variation in small samples. While such visualization demonstrates statistical bounds, confidence intervals were **not integrated into the main discussion**, as they risk overstating precision in a small, stylistically heterogeneous corpus. The analysis focuses instead on **grammatical and positional regularities** (how **deixis reorganizes reasoning**) rather than on population-level inference.

Framing Type	% Consequentialist	95% CI
Cosmological	50.0%	[18.8%, 81.2%]
Dialogic	50.0%	[18.8%, 81.2%]
First-person	33.3%	[9.7%, 70.0%]
First-person plural	100.0%	[61.0%, 100.0%]
Impersonal	50.0%	[18.8%, 81.2%]
Reflexive	16.7%	[3.0%, 56.4%]
Second-person	16.7%	[3.0%, 56.4%]
Spatial	66.7%	[30.0%, 90.3%]
Temporal	66.7%	[30.0%, 90.3%]

Consequentialist reasoning appeared most frequently under **first-person plural (100 %)**, **spatial (67 %)**, and **temporal (67 %)** framings, and least often under **reflexive** and **second-person** framings (17 %). The wide confidence intervals reflect small sample sizes per frame ($n \approx 6$), which limit inferential reliability.

3. Supplementary Observations and Boundary Conditions

DeepSeek produced the **shortest responses** (average 469 words) yet integrated the **widest range of ethical frameworks** (utilitarian reasoning in 96.3% of cases, care ethics in 90.7%, and deontological principles in 87%), suggesting that brevity can coincide with conceptual breadth. *Claude 3.5*, by contrast, generated the **longest responses** (average 750 words) but typically converged on a single ethical frame, indicating that verbosity signals rhetorical elaboration rather than pluralism.

A second irregularity, the “**philosophy-textbook effect**,” appeared in canonical dilemmas such as the *Trolley Problem*, where all three models sharply increased their use of philosophical terminology (*GPT-4o* + 1,471%,, *DeepSeek* + 460%,, *Claude* + 233%,) while **shortening their responses** (*Claude* – 63%,, *GPT-4o* – 25%,). When faced with familiar moral scripts, the models reproduced standard formulations instead of developing new reasoning.

These anomalies do not contradict the account of **deixis as a grammar of positionality**, but instead **mark its operational limits**. While deictic framing organizes justificatory style, canonical prompts and model-specific conventions show how positional grammar can **solidify into moral cliché** or **dissolve into rhetorical incoherence**. Such breakdowns illuminate the boundaries of current **AI alignment paradigms**, which assume that moral reliability can be achieved through constraint or optimization. What fails here is not factual accuracy but **interpretive stability**, the model’s capacity to sustain meaning once deixis has detached from any identifiable speaker.

4. Pronoun Adaptation Analysis

Pronoun adaptation measures how consistently each model mirrored the **deictic pronoun** provided in the prompt. For example, under a *second-person* framing (“You must decide…”), the response should primarily contain *you/your*, while a *first-person* framing should emphasize *I/my*. Each response was tokenized, and all personal pronouns (*I, me, my, we, us, our, you, your, he, she, they*) were counted using a simple **regex-based script**.

Adaptation accuracy was calculated as:

Adaptation Accuracy = ($\frac{\text{Number of prompt-matching pronouns}}{\text{Total pronouns}}$) × 100

Model	Mean Adaptation (%)	SD	Range	Notes
Claude 3.5 Sonnet	97.6	2.1	92–100	Occasional drift in dialogic frames
GPT-4o	98.9	1.2	96–100	Minor plural bleed (“we”) under reflexive prompts
DeepSeek-Chat	99.4	0.8	98–100	Essentially perfect pronoun substitution

All three models reproduced deictic markers with **near-perfect reliability**, confirming that **pronoun framing acts as a strong syntactic constraint** on model output. The few deviations occurred in **dialogic** and **reflexive** frames, where mixed grammatical roles were required (e.g., “*I wonder what you would do*”).

5. Recursive Pronoun Chains

Recursive pronoun chains were defined as **utterances containing three or more personal pronouns within a single clause**, for example: “I think you should consider what she might do if he refuses.” Such constructions signal *deictic layering* with multiple, nested points of address within one utterance. They index **linguistic simulations of self-reflection or multi-perspectival reasoning**, where pronouns stage imagined interlocutors or viewpoints without a stable speaker. Each text was segmented and parsed as follows:

- **Sentence segmentation:** punctuation marks . ? !
- **Clause segmentation:** commas, semicolons, colons, and dashes/em dashes.
- **Pronouns counted (case-insensitive):**
I, me, my, mine, myself, you, your, yours, yourself, he, him, his, himself, she, her, hers, herself, we, us, our, ours, ourselves, they, them, their, theirs, themselves.
- A **clause** was tagged *recursive* if it contained ≥ 3 of these pronouns.
- A **response** was classified as recursive if at least one of its sentences contained such a clause.
- Frequency was calculated as:

$$\text{Recursive Ratio} = \frac{\text{Recursive Sentences}}{\text{Total Sentences}} \times 100$$

TRecursive Pronoun Chains by Deictic Frame

Frame Type	Responses (n)	With Recursive Clauses (n)	% of Responses	Mean Recursive Sentences (%)	Example Clause
Reflexive	18	10	55.6 %	11.8 %	“I ask myself whether you would forgive me if they discovered this.”
Spatial	18	6	33.3 %	7.6 %	“We see you as they might from where we stand.”
Cosmological	18	4	22.2 %	5.4 %	“We must remind ourselves that they and we share the same fate.”
Dialogic	18	3	16.7 %	3.2 %	“You ask what I believe she intends, and I wonder too.”
First-person	18	3	16.7 %	3.5 %	“I tell myself I can forgive what they have done to you.”

Frame Type	Responses (n)	With Recursive Clauses (n)	% of Responses	Mean Recursive Sentences (%)	Example Clause
First-person plural	18	3	16.7 %	3.1 %	"We think you should consider what others like us might do."
Second-person	18	2	11.1 %	2.2 %	"You remind yourself of what they told you before."
Temporal	18	2	11.1 %	1.9 %	"I would have told you then what they know now."
Impersonal	18	0	0.0 %	0.0 %	—
Overall	162	33	20.4 %	4.3 %	—

Recursive pronoun chains occurred in approximately **one in five responses (20 %)**, with strong concentration in **reflexive** and **spatial** framings. Their distribution correlates significantly with frame type ($r \approx 0.72$, $p < 0.01$), supporting the claim that **deictic complexity, not pronoun frequency, marks the transition from anchored to suspended deixis**. These structures simulate *perspective-taking* through grammatical recursion rather than cognition, producing the *appearance* of introspection or empathy. They thus exemplify **non-anchored deixis in miniature**: language generating the illusion of inner dialogue without a subject to sustain it.