# Hydrangea: A Dataset of Context Loss Statements for Contextomy Detection and Quote Recontextualization

**Anonymous ACL submission**


## Abstract

Contextomy—the selective cropping of quotations to alter a speaker's intended meaning—is a longstanding rhetorical tactic to shift public opinion, yet its recontextualization remains a manual task performed by journalists and has not been operationalized as a computational problem. Prior contextomy detection work uses semantic similarity as a heuristic for meaning retention, but contextomy in the wild is multidimensional and inherently subjective. We propose Context Loss Statement Classification, a new task in which a model predicts pluralistic human judgments (Agree, Disagree, Unclear) of natural language statements describing how a cropped quote relates to its source. To support this task, we introduce Hydrangea, a human-annotated (n=804) dataset of 13,770 context loss statements built from 918 LLM-selected crops of 306 real-world excerpts from speeches, interviews, and press conferences sourced via Snopes. We train models to predict human agreement with context loss statements at the annotator level and the statement level and report scenarios where zero-shot LLMs perform competitively with finetuned BERT models. Overall, our findings demonstrate that contextomy is best modeled as a multidimensional, socially situated inference problem rather than a single faithfulness judgment.


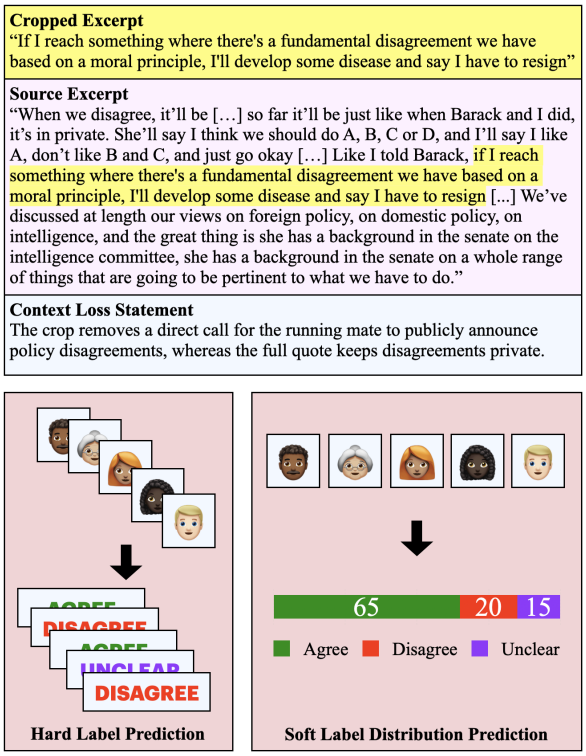


Figure 1: Context Loss Statement Classification task. Given a cropped excerpt, its source passage, and a context-loss statement describing how meaning may shift between the two, the model predicts either (1) a soft-label distribution over {Agree, Disagree, Unclear} corresponding to the response a target annotator population would give, or (2) a hard label produced by a particular annotator. Optionally, the model is conditioned on demographic attributes of the annotator(s).

## 1 Introduction

The phenomenon of *contextomy*, in which a quotation is selectively cropped to alter the speaker's intended meaning (McGlone, 2005), exploits the authority and authenticity that comes from directly quoting a speaker's words verbatim (Zelizer, 1989; Nylund, 2003) for political purposes. In *The Art of Quoting Out of Context*, McGlone illustrates in detail how a single sentence from Martin Luther King Jr.'s "I Have a Dream" speech has been deployed for decades to argue against affirmative action, despite King's explicit advocacy for race-based reparations elsewhere in his writings (McGlone, 2005). Contextomy has been documented as a common rhetorical tactic in political discourse, and can demonstrably shift public opinion (McGlone, 2005; Han and Lee, 2013). Yet, most recontextualization of contextomies today is done by journalists and news outlets.

A substantial challenge from an NLP perspective is that determining whether a quote is actually a contextomy is inherently subjective and

multi-faceted. In our dataset we show empirically that a quote could be out of context for a variety of reasons (e.g. it misstates or omits the speaker's goals, removes linguistic qualifiers, or simply shifts the speaker's emotional tone towards a topic). This marks a stark contrast to summarization where faithfulness to speaker intended meaning is typically operationalized as a single judgment of whether a shorter text preserves the meaning of a longer one and only explicit interpretations are considered (Maynez et al., 2020; Choi et al., 2021; Es et al., 2024)

We propose a new task, **Context Loss Inference Classification**: given a cropped excerpt and the full excerpt it was drawn from, a model must predict human judgments (AGREE, DISAGREE, or UNCLEAR) of a natural language statement describing how the crop relates to its source. This enables the NLP community to go beyond simple metrics like semantic similarity (Song et al., 2023) to generate and evaluate fine-grained, human-readable, assertions of context loss.

To support this task, we introduce HYDRANGEA, a dataset of 13,770 context loss hypotheses constructed from 918 LLM-selected crops of 306 real-world excerpts drawn from primary source materials such as speeches, interviews, and press conference transcripts used by Snopes' fact-checking and news editors. The context loss statements capture a diversity of viewpoints spanning 15 contextual dimensions (e.g., Emotional Tone, Supporting Rationale, Linguistic Qualifiers) plus a top-level binary judgment. Each context loss statement is labeled AGREE, DISAGREE, or UNCLEAR by 4.7 annotators on average, with various demographic features of the annotators included such as ethnicity, age, gender, and political affiliation. We also capture concepts such as annotators' agreement with the statement, perception of the statement as harmful, and more. The variation in annotator demographics together with the taxonomy applied to real-world excerpts produces a resource that supports both majority-label modeling and pluralistic analysis of how annotator perspective shapes judgments of contextual shift (Plank, 2022; Cabitza et al., 2023).

We evaluate fine-tuned encoders and zero-shot LLMs on two task formulations and find that while fine-tuned models better predict annotator label distributions, both models perform about on-par when classifying specific annotator labels. Supplementing with annotator demographic information yields mixed results.

Our contributions are:

- The first dataset of contextomy hypotheses with human judgments, structured along 15 dimensions of contextual shift and a top-level binary judgment.
- A taxonomy and template for generating contextomy statements, empirically derived from human-written contextomy descriptions.
- Baseline LLM and fine-tuned model results on two operationalizations of the Context Loss Inference Classification task.

## 2 Related Work

### 2.1 Contextomy and Decontextualization

The term *contextomy* was coined by historian Milton Mayer to describe the deliberate truncation of quotations to reverse or distort their meaning, and was formally analyzed as a rhetorical phenomenon by McGlone (2005), who documented its "deployment across contemporary mass media to defame public figures and misappropriate rhetoric". A key insight of McGlone's analysis is that verbatim quoting does not guarantee faithful representation: a quote can reproduce a speaker's exact words while still fundamentally misrepresenting their intent, simply by excluding nearby context that clarifies what was meant; there is no false statement to rebut, only an omission to expose. This can be compared fundamentally to decontextualization (Choi et al., 2021) which explicitly aims to take a sentence together with its context and rewrite it to be "interpretable out of context if feasible, while preserving its meaning". Choi et al. (2021) restrict meaning preservation to truth-conditional content and explicature, while acknowledging the broader pragmatics tradition they set aside. Hydrangea bridges these two threads: it shares decontextualization's goal of surfacing what a standalone excerpt fails to convey, while expanding the scope of "meaning preservation" beyond truth-conditional content to include the implicit, speaker-oriented dimensions McGlone treats as the harm of contextomy.

### 2.2 Semantic Similarity as a proxy for Contextomy

QuoteCSE (Song et al., 2023), detects contextomy in Korean-language journalism by classifying whether a headline quote has a semantically similar embedding as a quote in the article body. Fundamentally, and apart from our contribution, there is

no human evaluation of speaker intent preservation; in QuoteCSE, the semantic similarity of embeddings is used as a ground truth measure of contextomy. Critically, the CtxSimFit metric (Yerukola et al., 2023) has shown that standard automatic metrics such as ROUGE and SBERT embedding comparisons correlate poorly with human judgments of contextual faithfulness, motivating our use of human annotations on context loss statements as the primary evaluation signal rather than surface-level similarity scores.

### 2.3 Pragmatic and Implicit Meaning in NLP

Contextomy detection requires both semantic and pragmatic understanding; harm arises both from literal statements in the cropped text and the implicatures about the speaker that can be made due to what was omitted. Jeretic et al. (2020) find that NLI models trained on MultiNLI acquire some pragmatic reasoning, but coverage is partial and inference-type-dependent. The Implied NLI dataset (Havaldar et al., 2025) similarly shows that fine-tuning on implied entailment improves model understanding of non-explicit meaning, though models struggle to generalize across domains. In a more applied setting, Hoyle et al. (2023) demonstrate that explicitly enumerating implications from a politician's tweet can improve downstream performance in predicting the legislator's voting behavior, motivating our approach of enumerating several dimensions of context loss capturing both implicit and explicit meaning shifts.

Our approach also builds on the use of structured linguistic frames and natural language explanations to characterize socially situated language phenomena, including COBRAFRAMES for offensive language (Zhou et al., 2023) and Misinfo Reaction Frames for misinformation (Gabriel et al., 2022). These frameworks demonstrate that free-form explanations grounded in a structured taxonomy of dimensions can capture nuanced sociopragmatic phenomena that resist single-label classification — a property we argue is essential for modeling contextomy. We bring these threads together to tackle contextomy as a multidimensional, socially situated inference problem. Doing so requires ecologically valid source material, hypotheses grounded in a validated taxonomy of contextual shift, and pluralistic annotation that treats disagreement as signal rather than noise.

## 3 Dataset Creation

We construct Hydrangea through a four-stage pipeline: (1) curating real-world primary source excerpts from Snopes, (2) generating LLM crops simulating real world contextomies, (3) deriving structured templates of contextomy dimensions from pilot annotation studies, (4) generating crop-specific statements from templates, and (5) collecting human annotations on the crops and statements.

### 3.1 Snopes Excerpt Collection

We constructed a set of 306 source excerpts by extracting highlighted primary source excerpts that are used as evidence in articles from Snopes, a professional fact-checking website that routinely embeds extended primary source material — including speech transcripts, interview recordings, and press conference statements — directly within its articles. Because Snopes articles are motivated by viral media coverage or social media resharing, every extended excerpt in our dataset represents a public figure that was recontextualized (either faithfully or unfaithfully) in a way that demonstrably reached and influenced a public audience.

### 3.2 LLM Crop Generation

For each of the 306 source excerpts, we prompted Claude Sonnet 4.6 to generate three distinct cropped substrings of the full excerpts: two contextomies and one faithful crop, resulting in 918 crop–full excerpt pairs. To generate the contextomies, the prompt instructs the model to prioritize crops that would plausibly appear as a headline or social media post. For the faithful crop, the model is simply instructed to choose a crop that best retains the speaker's intention. We then check the generations with (1) a validator prompt to make sure the substrings are valid crops structurally (e.g. clear unambiguous noun subject, no fragmented phrases that can't standalone) and (2) a quality prompt that ensures the contextomy generations are at least moderate distortions of speaker intent. Any crops that fail these two checks are removed and regenerated. The full crop generation prompt and crop reviewer prompts are provided in Appendix A.

### 3.3 Contextomy Dimension Template

To generate the contextomy dimension template, we conducted two Prolific pilot studies to understand how annotators perceive meaning shift and context loss. In the first pilot study a pool of 35

annotators labeled 31 pairs of cropped and source excerpts derived from Snopes. We asked them to share in free text (1) how the explicit meaning of an excerpt changed from the crop to the full and (2) how the implicit meaning of the crop differend from the full. Each excerpt was annotated 5 times, resulting in 155 free text annotations for implicit change and explicit change. From these explanations we extracted 15 qualitative themes on how the meaning of an an excerpt might differ from the speaker's intention. We ran a second pilot study on 100 examples from Snopes to validate that the 15 dimensions were relevant and comprehensive aspects of contextomy. We found that for each dimension, crops were significantly more likely to be annotated as a contextomy if the annotator also marked a meaning shift along that dimension. We also found that when selecting types of meaning shift, the option of "Other" was rarely selected when the crop was a contextomy and frequently selected when the crop was not a contextomy, suggesting that these dimensions cover the pragmatic ways meaning shifts in dimensions that are important to determining contextomy.

### 3.4 Context Loss Statement Generation

In the dimension template pilot studes we saw meaningful annotator disagreement about the relevant dimensions contextomy, even when annotators agreed about whether the meaning had changed. To capture this diversity while avoiding the cognitively intensive task of generating strictured inferences, we used an LLM to generate specific claims of contextomy based on the 15 template dimensions. For each of the 918 (crop, full) pairs, we prompted Claude Sonnet 4.6 to instantiate all 15 context loss template statements, creating contextomy assertions from each contextomy dimension. To verify that generated statements faithfully instantiate their assigned templates, we conducted a manual audit detailed in Appendix B.

Examples of dimension templates and their instantiations are shown in Table 1.

### 3.5 Human Annotation

We recruited human annotators through Prolific's AI Task Builder. Because the corpus consists predominantly of political quotes, we redacted the speaker name in both source and cropped excerpts to reduce speaker bias in annotators' determinations of whether context was lost.

Each task assigns an annotator to one Primary

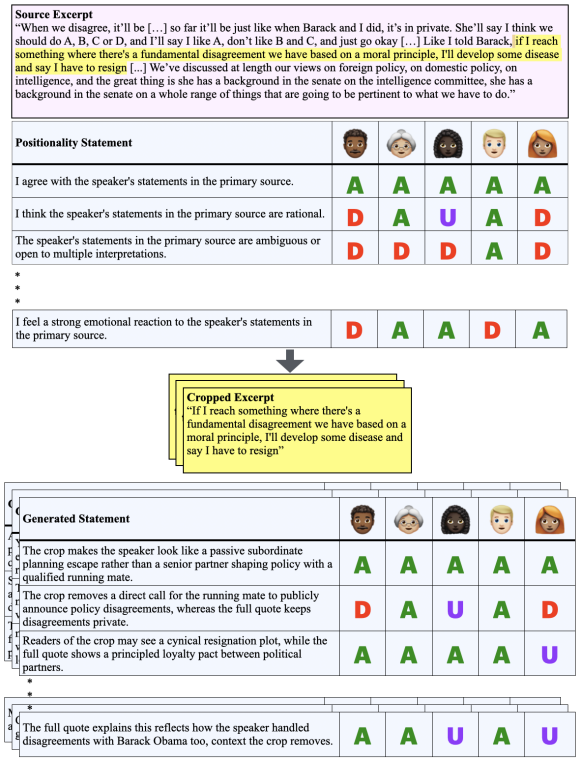


Figure 2: The structure of the annotation task given to Prolific participants. Annotators first answer 8 questions addressing potential covariates with contextomy judgements before assessing whether they agree with each of the context loss statements for each of the 3 crops associated with the the source excerpt.

Source Excerpt and three Cropped Excerpts drawn from it, producing 62 judgments per annotator on a uniform AGREE/DISAGREE/UNCLEAR scale: 8 positionality items on annotator priors, speaker recognition, and affective response are administered at the start, followed by three crop-specific blocks of 18 items each. Every crop-specific block contains a standardized top-line contextomy question, then 15 generated inference statements, then an attention check, and finally a repetition of the top-line contextomy question to capture how contextomy perceptions shift after exposure to the 15 dimensional statments. Annotators are limited to a maximum of five tasks to constrain per-annotator influence on the dataset, and five annotators are recruited per Primary Source, so the three crops drawn from a single source and their corresponding context loss statements are rated by the same annotators.

Annotators who failed attention checks were removed, leaving 4.7 annotations per triplet on average. Further task design detials are found in Appendix C, and full task instructions in Appendix D.

| Template Statement | Generated Instantiation |
|---|---|
| The cropped excerpt omits context about the speaker's audience and/or the speaker–listener relationship, altering the meaning compared to the source excerpt. | The crop hides that the speaker was addressing an audience about criminal justice reform for young offenders. |
| Given the original source of the quote, the speaker's words in the cropped excerpt were presented in a manipulative or misleading way. | The crop presents a rhetorical crowd-engaging remark as a genuine insult toward young people. |
| The cropped excerpt significantly changes the tone or emotional attitude of the speaker. | The crop sounds harshly dismissive while the full quote is compassionate and reform-oriented toward young offenders. |
| The cropped excerpt omits language that clarifies the speaker's moral or ethical stance toward the event being discussed. | The crop omits the speaker's advocacy for treating young offenders with developmental understanding not punishment. |
| The cropped excerpt removes the reasoning behind the statement which was present in the source excerpt. | The crop removes the reasoning that 18–24-year-olds need guidance like college students, not adult criminal sentences. |

Table 1: Examples of LLM-generated contextomy inference statements instantiated from template statements for a single (crop, full) pair.

## 4 The Hydrangea Dataset

The annotation study recruited 804 unique annotators, each annotator could complete between 1 and 5 task batches of 3 crops each, for a total of 3 to 15 crops per annotator. Most annotators (n=500) completed the minimum set of 3 crops, with the full breakdown shown in Table 2.

| # crops | 3 | 6 | 9 | 12 | 15 |
|---|---|---|---|---|---|
| # annotators | 500 | 132 | 67 | 51 | 54 |

Table 2: Annotator workload summary

In addition, after the initial recruitment and filling out of the annotation task, we filtered out all responses from annotators who failed two or more attention checks, allowing Prolific to backfill the task with more annotators to complete it. In the majority of the cases (681) we get all five of the annotators that we need, but there are some cases that receive fewer anotations: 204 cases receive four annotators, 30 receive three annotators, and 3 cases receive only two annotators.

### 4.1 Contextomy Judgment Post-Reasoning Shifts

**Task design.** Each annotator made a holistic contextomy judgment on the cropped excerpt as a whole, separately from labeling the individual LLM-generated contextomy assertions. Annotators responded *Agree*, *Disagree*, or *Unclear* to the statement: *"The Cropped Excerpt omits context from the Primary Source in a way that significantly distorts the speaker's intended meaning, misrepresenting the speaker's views."* This judgment was elicited twice: once *before* the annotator labeled any of the 15 LLM-generated contextomy assertions, and once *after* labeling all 15. This design lets us measure how exposure to model-generated reasoning shifts annotator judgments.

**Results.** The marginal distribution of *Agree* responses rose from 50% before reasoning to 62% after, while *Disagree* fell from 42% to 31%. Exposure to the 15 model-generated contextomy assertions appears to surface miscontextualizations that annotators would not otherwise have flagged, shifting roughly 12 percentage points of mass from *Disagree* to *Agree*. Table 3 provides the full joint distribution. The largest off-diagonal cell, *Disagree* $\rightarrow$ *Agree* at 0.166, is the mechanical driver of the marginal shift; the reverse movement (*Agree* $\rightarrow$ *Disagree*) accounts for only 0.064. The asymmetry suggests the reasoning step is not merely adding noise to judgments but disproportionately persuading annotators *toward* the contextomy label, consistent with the interpretation that LLM-generated assertions surface distortions the annotator had not initially noticed.

## 5 Dataset Item Difficulty

Before turning to model performance, we characterize how human annotators distribute across agreement buckets, how those distributions shift between contextomy sections, and how LLM zero-shot predictions compare to the human distribution. Throughout this section, we summarize per-item agreement into four ordered buckets: $N$ (no major-

| | After reasoning | | | |
|---|---|---|---|---|
| Before reasoning | Agree | Disagree | Unclear | Total |
| Agree | 0.414 | 0.064 | 0.022 | 0.500 |
| Disagree | 0.166 | 0.226 | 0.025 | 0.417 |
| Unclear | 0.040 | 0.019 | 0.023 | 0.082 |
| Total | **0.621** | **0.309** | **0.071** | **1.000** |

Table 3: Joint distribution of annotator contextomy judgments before (rows) and after (columns) labeling the 15 model-generated contextomy assertions. Cells give the share of annotator–bundle pairs in each before/after combination; the interior sums to 1.00.

| Inference | $N$ | $A_0$ | $A_+$ | $A_{++}$ |
|---|---|---|---|---|
| Contextomy (Initial) | 21% | 37% | 31% | 11% |
| Audience Relationship | 15% | 26% | 32% | 27% |
| Manipulative Framing | 13% | 26% | 36% | 25% |
| Emotional Tone | 12% | 28% | 38% | 23% |
| Ethical Stance | 12% | 26% | 33% | 28% |
| Supporting Rationale | 9% | 19% | 36% | 36% |
| Inferential Conclusions | 11% | 26% | 35% | 28% |
| Recommended Actions | 10% | 22% | 37% | 30% |
| Speaker Goals | 11% | 24% | 37% | 28% |
| Situational Context | 10% | 20% | 38% | 33% |
| Statement Salience | 12% | 34% | 34% | 20% |
| Identity Based Alienation | 19% | 32% | 33% | 16% |
| Sociocultural Context | 9% | 21% | 36% | 34% |
| Linguistic Qualifiers | 21% | 34% | 30% | 15% |
| Speaker Role | 14% | 30% | 34% | 23% |
| Speaker Portrayal | 14% | 31% | 34% | 21% |
| Contextomy (Posterior) | 14% | 33% | 34% | 18% |

Table 4: Human agreement bucket distribution by inference type ($N = 918$ items per row). $N$ = no majority, $A_0$ = weak majority ($\geq 50\%$), $A_+$ = strong majority ($\geq 70\%$), $A_{++}$ = unanimous ($\geq 90\%$). Rows for the holistic contextomy judgment (Initial and Posterior) bracket the 15 template dimensions.

ity), $A_0$ (weak majority, 50%+), $A_+$ (strong majority, 70%+), and $A_{++}$ (near-unanimous, 90%+).

### 5.1 Human Agreement Buckets by Section

We first ask how annotator agreement varies across the three section types of the instrument: the bare-crop baseline (C0), the 15 dimension prompts (C1–C15), and the open-ended final prompt (C17). If the template dimensions C1–C15 are well-calibrated to the way humans perceive contextual shift, we should expect higher proportions of strong-majority and unanimous agreement on those items than on the unstructured C0 and C17 prompts. Table 4 reports the per-section distribution.

The pattern supports our hypothesis. The template prompts (C1–C15) produce the highest proportion of unanimous agreement (25.8%) and the lowest proportion of no-majority items (12.7%), while the unstructured prompts yield substantially flatter distributions: C0 has the highest no-agreement rate (20.6%) and lowest unanimous rate (11.0%), with C17 falling between them. This is consistent with the interpretation that the template dimensions surface concrete, discriminable aspects of contextual shift that annotators converge on, whereas open-ended prompts leave more room for divergent readings. Further examination of the agreement structure is provided in Appendix F.

## 6 Modeling Contextomy Inference

In this section, we present baseline LLM and fine-tuned contextomy inference models as a way of further exploring the Hydrangea dataset. We briefly formalize the two versions of the task introduced in Figure 1 before describing our modeling methodology and results.

### 6.1 Task Formulation

We experiment with two versions of the contextomy inference task: a statement-level *soft label* experiment that aggregates targets across annotators, and an annotator-level *hard label* task with targets derived from specific annotator responses.

In the following, we refer to a data set item as a tuple $(s_i, c_{i,j}, h_{i,j,k}, d_{i,j,a}, l_{i,j,k,a})$, where $s_i$ is the $i$th source excerpt, $c_{i,j}$ is one of the LLM-generated crops of $s_i$ ($1 \leq j \leq 3$), $h_{i,j,k}$ is the contextomy inference statement generated from template $k$ (C1–16) or one of the static statements (C0 or C17), $d_a$ is the sequence of demographic features of annotator $a$, and $l_{i,j,k,a}$ is the label assigned by the annotator.

**Statement-level task** We construct soft label targets as the response distribution ($y \in \{Agree, Disagree, Unclear\}$) for a particular item:

$$l^{\text{soft}}_{i,j,k}[y] \propto |\{a : l_{i,j,k,a} = y\}|$$

Given an input item, $(s_i, c_{i,j}, h_{i,j,k})$, a statement-level model produces a prediction $\hat{l}^{\text{soft}}_{i,j,k}$. Following Nie et al. (2020), we evaluate models based on the average Jensen-Shannon divergence and Wasserstein distance[1] between $l_{i,j,k}$ and $\hat{l}^{\text{soft}}_{i,j,k}$. In principle, the model could also be provided with a multiset of annotator demographics features, but we find

[1]The Wasserstein distance metric (also known as *earth mover's distance* assumes that the support of the two distributions is a metric space; we implicitly assume that semantic distance between the distance between *Agree* and *Unclear* is the same as between *Unclear* and *Disagree*.

it more informative to experiment with annotator features with the hard label task.

**Annotation-level task** The annotator-level task is a more straight-forward 3-way classification task. Given an input item $(s_i, c_{i,j}, h_{i,j,k}, d_{i,j,a}, l_{i,j,k,a})$, the model makes a prediction $\hat{l}_{i,j,k,a}$. For this task, our primary metric is macro-average F1 score, but we also consider the label-specific F1 scores.

### 6.2 Baselines

We consider three weak baselines for each task. The *zero-knowledge* baselines are uniform distribution (statement-level task) and a random baseline (annotation-level task). The *label distribution-aware* baselines are the mean (train set) label frequency for the statement-level task and the majority class for annotation-level. Finally, we consider *template-aware* baselines: the template-wise mean label frequency and majority class. This operates as a best-case-scenario hypothesis-only baseline — it would be achieved by a model that knows the train-set distribution of responses stratified by contextomy template and can recognize which template a test-set statement comes from.

### 6.3 Modeling Verification of Contextomy Inference Statements

To provide a strong baseline for these two tasks, we test LLMs prompted with a basic prompt describing the task (see Appendix A), and several transformer encoder models fine-tuned as cross-encoder classifiers on the training split.

### 6.4 LLM annotation

We tested two frontier LLMs, Claude Opus 4.7 and GPT-5.5, varying whether annotator demographics are available to the model. Without annotator demographics, the models are prompted to predict the target label counts that 5 annotators might provide for the given item triple, corresponding to the statement-level task. With annotator demographics, the models predict the specific label an annotator with a given demographic profile would supply to a input item. The prompts are provided in Appendix H, and the results for the two models and two conditions are presented in Table 5.

### 6.5 Fine-tuned models

We fine-tune BERT-base (Devlin et al., 2019), ModernBERT-base, and ModernBERT-large (Warner et al., 2024).[2] For both soft- and hard-label tasks, we use the models as cross-encoders, with full cross-attention on all three three primary inputs texts demlimited by `[SEP]` tokens. Following (Fleisig et al., 2023), in versions that include annotator demographics, we concatenate a demographics string, also delimited by `[SEP]` tokens. To differentiate the constant contextomy inference statements (C0 and C17), which have identical text, we append the strings "`[pre]`" and "`[post]`" to those statements. This will give models the opportunity to learn how the inferential characteristics of the constant statement change as a result of the annotator having responded to statements about specific dimensions of contextomy.

We did not perform a comprehensive hyperparameter search, but a limited development process revealed that different training procedures were most effective for the two different tasks (see Appendix G). For the **statement-level task** (soft labels), we report results on models trained with cross-entropy loss on annotation-level items (i.e. the standard classification setup, where uncertainty is manifest through different labels appearing with the same training instance).

For the **annotation-level task** (hard labels), we found that focal loss (Lin et al., 2020) improved performance on the development set, helping to deal with label imbalance. We ran an ablation experiment to test the effect of adding annotator demographics to the model. In particular, fed the model a string with text corresponding to seven demographic characteristics collected from Prolific: *age*, *ethnicity*, *ethnicity (simplified)*, *gender*, *sex*, *LGBTQ* (Yes or No), and *Political Spectrum (US)*.

### 6.6 Results & Discussion

The results for both tasks are reported in Table 5. On the statement-level task, all models achieved better label distribution prediction than the template-aware baseline, demonstrating that there is inferential information to be leveraged in the full and crop with respect to the contextomy statement. For fine-tuned models, we see that ModernBERT performs better than BERT, especially on Wasserstein Distance — ModernBERT is better at assigning weight between *Agree* and *Disagree*. ModernBERT-Large performs only very slighly better.

[2] Because BERT has a more limited context window (512 tokens), around 10% of the examples need to be truncated.

| | Statement-level | | Annotation-level | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | JSD | WD | Agr F1 | | Dis F1 | | Unc F1 | | Macro Avg. | |
| *demographics N/Y* | | | N | Y | N | Y | N | Y | N | Y |
| BERT-base | 0.256 | 0.382 | 62.1 | 68.8 | 47.9 | 46.2 | 6.0 | 5.0 | 38.7 | 40.0 |
| ModernBERT-base | 0.254 | 0.364 | **72.3** | 72.1 | 43.5 | **48.9** | 11.5 | 5.1 | **43.0** | 42.0 |
| ModernBERT-large | **0.252** | **0.363** | 64.0 | 64.6 | 45.6 | 27.8 | **13.2** | 12.2 | 41.0 | 34.9 |
| GPT 5.5 | 0.378 | 0.557 | 72.7 | 71.8 | **52.5** | 51.4 | 3.5 | **7.6** | 42.9 | 43.6 |
| Claude Opus 4.7 | 0.317 | 0.428 | **77.8** | 75.7 | 50.9 | 51.5 | 0.0 | 4.0 | 42.9 | **43.7** |
| zero-knowledge | 0.391 | 0.600 | 45.4 | | 32.2 | | 12.62 | | 30.0 | |
| label distribution | 0.278 | 0.433 | 78.6 | | 0.00 | | 0.00 | | 26.2 | |
| template-aware | 0.267 | 0.399 | 78.1 | | 16.3 | | 0.00 | | 31.5 | |

Table 5: Performance on the Hydrangea dataset annotator response distribution task (left) and label classification task (right). Classification performance is reported for models with (Y) and without (N) access to annotator demographics. Best scores for each metric and model category (LLM, encoder) are highlighted in **bold**.

On the annotation level task, we see similar patterns when comparing within encoder models — ModernBERT performs better, but the larger model offers little or even negative performance gains. Demographics-aware models have a slight advantage on some but not all metrics. ModernBERT-base exhibits a slightly better Macro Average F1 score without demographics than with (43.0/42.0), but this is due mainly to degraded performance on the *Unclear* label, which is an extreme minority class and semantically less distinct. Performance is improved on *Disagree* (43.5/48.9) and nearly unchanged on *Agree* (72.3/72.1). BERT-base exhibits different behavior, with the largest performance increase on *Agree* (62.1/68.8). The results for the best performing prompted LLM, Claude Opus 4.7, are similar to the fine-tuned models, even though no fine-tuning was performed; the model performs best on the majority class, *Agree*, and demographic information gives a slight boost on *Disagree* and *Unclear*. Model architecture and pre-training may have a substantial impact on the information models are able to take advantage of, especially when it comes to annotator demographics.

Table 6 shows results for the two top-line contextomy statements (C0 and C17) relative to the template-based statements (C1-C15). Performance on the C0 statement is substantially worse than the C17 statement, which behaves more similarly to the dimensional statements. This is consistent with the fact that annotators also displayed greater levels of agreement on the final contextomy statement in comparison to the inital one.

| | Demo | C0 | C17 | C1-15 |
|---|---|---|---|---|
| Agree | N | 35.4 | 66.3 | 71.6 |
| | Y | 36.1 | 62.5 | 72.3 |
| Disagree | N | 59.8 | 47.6 | 34.6 |
| | Y | 61.2 | 47.2 | 43.0 |
| Unclear | N | 16.7 | 0.0 | 8.8 |
| | Y | 0.0 | 4.8 | 4.1 |
| Macro avg. | N | 37.3 | 38.0 | 38.3 |
| | Y | 32.4 | 38.2 | 39.8 |

Table 6: ModernBERT-base results on the annotator-level (hard label) task, stratified by contextomy statement. C0 and C17 are the "generic" contextomy statement, asked at the beginning and end of the questionnaire. The C1-15 column shows the average across the template-based statements.

**Conclusion** Contextomy is a complex and multifaceted phenomenon with implications for media and healthy discourse. In this paper, we presented Hydrangea, a dataset of context loss inference statements designed to facilitate research on the identification and recontextualization of contextomies. To capture the complex nature of the phenomenon, the dataset includes context loss inference statements along 15 different dimensions of meaning shift. Our modeling results confirm that contextomy detection is a difficult task, but appropriately accounting for the subjective and multi-dimensional nature of the phoenomenon makes it possible to leverage inferences from fine-tuned and zero-shot LLMs.

## 7 Limitations

Several limitations of this design warrant discussion. Although our full excerpts are drawn from Snopes and represent real-world cases of viral contextomy, the cropped excerpts are LLM-generated rather than naturally occurring. While we motivated this decision by the need for systematic variation across political framings and the practical infeasibility of sourcing four naturally occurring crops per excerpt, it means that the crops themselves may not fully reflect the linguistic properties of organically produced contextomy. Asking annotators whether each crop is realistic during annotation partially addresses this concern by allowing dataset users to discard crops that annotators judge as implausible, but this filter is applied after generation rather than as a generative constraint. We find this to be an acceptable limitation because as AI systems are used for an increasing range of tasks, it is reasonable to expect they may be used to crop sources in the future, making LLM-generated crops an ecologically valid target domain in their own right.

Additionally, the political balance of LLM-generated crops remains largely unevaluated beyond annotator political affiliation and their stances towards source excerpts and their contextomy judgments as proxies. However, because we sample nearly evenly across the political ideological spectrum in our annotator pool, we expect that even if generated crops or statements carry latent bias, the human evaluations of those crops and statements will not systematically reflect that bias. Future work should examine whether crops differ systematically across political framings in ways that could introduce ideological bias into the dataset.

## 8 Ethical Considerations

The Hydraenga dataset is intended to facilitate computational identification and re-contextualization of contextomy, which is potentially benneficial for a healthy media environment. Nevertheless, there is a risk that models trained on the dataset could be used to automatically generate contextomies from media sources or public figures.

The median pay to annotators was $13/hr; the median time for the annotation task was 20 mins.

## References


Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. Tex.ids= Devlin2019a.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jeongsub Han and Chang-Ho Lee. 2013. Quoting practices in korean online news: Contextomized quotations in political reporting. *Asian Journal of Communication*, 23(4):397–414.

Shreya Havaldar, Hamidreza Alvari, John Palowitch, Mohammad Javad Hosseini, Senaka Buthpitiya, and Alex Fabrikant. 2025. Entailed between the lines: Incorporating implication into NLI. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32274–32290, Vienna, Austria. Association for Computational Linguistics.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919.

Matthew S. McGlone. 2005. Contextomy: The art of quoting out of context. *Media, Culture & Society*, 27(4):511–522.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Mats Nylund. 2003. Quoting in front-page journalism: Illustrating, evaluating and confirming the news. *Media, Culture & Society*, 25(6):844–851.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Seonyeong Song, Hyeonho Song, Kunwoo Park, Jiyoung Han, and Meeyoung Cha. 2023. Detecting contextomized quotes in news headlines by contrastive learning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 697–704, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint*. ArXiv:2412.13663 [cs.CL].

Akhila Yerukola, Xuhui Zhou, Elizabeth Clark, and Maarten Sap. 2023. Don't take this out of context!: On the need for contextual models and evaluations for stylistic rewriting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11419–11444, Singapore. Association for Computational Linguistics.

Barbie Zelizer. 1989. "saying" as collective practice: Quoting and differential address in the news. *Text & Talk*, 9(4):369–388.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. CobraFrames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*.

# A LLM Crop Generation

This appendix contains the full prompts used in the Hydrangea generation, validation, and quality scoring pipeline. All prompts are reproduced verbatim from the production code.

Both the cropped excerpts and the contextomy inference statements are generated by Claude Sonnet 4.6. While the 15 inference dimensions are derived from human pilot studies and final labels come from human annotators, the intermediate stimuli are shaped by a single model's priors about salience, distortion, and faithful representation. We mitigate this by constraining generation with human-authored templates and a fixed dimension set, separating crop generation and statement generation into independent prompts, and including faithful crops as a control condition. Nonetheless, a model with systematically different priors might surface different crops or instantiate the templates in subtly different ways. We view the dataset as a characterization of contextomy as instantiated through one capable model, not as a model-independent ground truth.

## A.1 Shared Rules

The following rules apply to all generated substrings (contextomies and faithful) and are referenced by both the generator and validator prompts.

0. VERBATIM ONLY — copy the substring character-for-character from the quote. No paraphrasing, no added words, no reordering of words.

1. DO NOT include or span any editorial markers such as `[. . . ]`, `[sic]`, or any bracketed interpolation. If such markers appear in the

quote, the substring must begin and end entirely within the clean text on one side of them. Never bridge two segments separated by an editorial marker.

2. STANDALONE PROPOSITION — the substring must contain a complete, self-contained claim with a named or explicitly stated subject and a full predicate. It must pass two tests: (a) Grammatical test: fits the frame "The speaker said "_____"" with no trailing fragments. (b) Headline test: a reader with no prior context could write a specific factual news headline from the crop alone. Fail any crop where the subject is an unresolved pronoun (*he, she, it, they, this, that, them, we*) whose referent does not appear within the crop itself, where a key noun is implied but absent, or where the main claim cannot be named without guessing.

3. The contextomy crops should clearly support a misleading narrative that misrepresents the speaker's intent, actions, stance, or experiences. Only select crops that misrepresent the speaker in a specific and nameable way when removed from context — not just become vague or incomplete.

4. If the quote only supports one type of shift, the same label may be used for both contextomies (e.g. two `consequence_shift`); however, avoid text overlap between the two contextomies and the faithful substring as much as possible.

5. The faithful substring must be representative of the speaker's actual intentions and overall message.

6. If the quote is a dialogue, crop only from the primary speaker's words (typically not the interviewer or host).

## A.2 Generator Prompt

### System prompt.

You are a media analysis expert building an academic dataset about contextomies (misleading quote cropping). A contextomy is a verbatim substring of a longer source excerpt that, when extracted, distorts the speaker's original meaning.

Produce exactly TWO contextomies (verbatim substrings that distort meaning when cropped), each labeled `intent_shift`, `situation_shift`, or `consequence_shift`. Labels may repeat.

**Contextomy types** (label each crop with the best fit):

**`intent_shift`** distorts the speaker's purpose, motive, or reasoning. E.g. stripping conditionals ("if", "unless"), removing justification, or shifting tone from calm to aggressive.

**`situation_shift`** misrepresents the literal conditions of the statement. E.g. hiding the speaker's role, the setting, the audience, or cultural context that would make the statement unremarkable.

**`consequence_shift`** implies different outcomes, conclusions, or calls to action. E.g. making consequences sound more alarming, more certain, or more benign than intended.

Choose crops a politically motivated actor would plausibly deploy as a headline or social media post — either making the speaker look worse (extreme, hypocritical, incompetent) or sanitizing them (hostile → neutral).

**Headline test.** Before selecting any crop, ask: "Could a reader unfamiliar with this story write a specific, factual one-sentence news headline using ONLY the words in this crop?" If the answer is no — because key nouns are missing, the subject is a pronoun, or the claim is too vague to name — the crop fails.

A good crop is specific enough that a reader could name exactly what belief, claim, or demand it expresses without seeing the original quote. Illustrative contrasts appear in Table 7.

[*Shared rules (§A.1) inserted here.*]

**Output.** Single JSON object, no markdown/code fence/preamble:

```
{"contextomies": [{"substring": "...",
            "label": "intent_shift|situation_shift|consequen
                  ...],
 "faithful": {"substring": "..."}}
```

Exactly 2 contextomy objects. Labels may repeat. All substrings verbatim from the quote.

### User prompt template.

```
SPEAKER: {speaker}
ROLE: {role}
SOURCE_EXCERPT:
{source_excerpt}
```

## A.3 Validation Follow-up Prompt

After validation, only the failed contextomy slots and the faithful substring (if it failed) are regenerated using the following follow-up message:

Your previous response failed validation on the items below. Reread the rules in the system prompt.

`VALIDATOR OUTPUT: {checks_json}`

`FAILED ITEMS: {failed_slots_detail}`

| Bad crop (fails) | Good crop (passes) |
|---|---|
| "we've never gotten anything" — *we* and *anything* have no referent inside the crop. | "Black Americans have never gotten the forty acres and a mule that was promised after emancipation" — reframes a historical grievance as a present-day accusation. |
| "yesterday was a really bad day for him, and this is what he did" — *him*, *this*, and *he* have no in-crop referent. | "After losing the primary, Senator Collins had a really bad day, and this is what he did to the party volunteers" — turns a sympathetic moment into an implied scandal. |
| "We solved it twice" — *it* is underspecified. | "Border control — we solved it twice" — hides a specific political claim about a contested policy. |
| "many of them were the wrong people" — no referent for who or what. | "We invited the scientists but many of them were the wrong people for the job of public safety" — reframes institutional critique as blanket dismissal. |
| "we declare that any strange winds" — trails off mid-clause. | "we declare that any strange winds blowing in this direction will not disturb our course" — reveals a specific declaration of purpose. |
| "art lasts forever" — no referent for context. | "art lasts forever, and no one can change what JFK and our shared history stands for" — sanitizes a sharp anti-Trump speech into a generic patriotic statement. |

Table 7: Illustrative good/bad crop pairs included in the generator system prompt to operationalize the headline test.

Generate severe contextomy replacements that maximize the gap between what a reader would infer from the substring alone and what the full `SOURCE_EXCERPT` shows the speaker meant — as long as every rule passes for the corrected text.

Before finalizing each replacement, apply the headline test: a reader with no prior context must be able to write a specific, factual news headline using only the words in the crop. If they would need to guess a referent, reselect.

Produce corrected replacements ONLY for the failed items listed above. All substrings must be verbatim from the `SOURCE_EXCERPT` in the first user message.

JSON only (no markdown, no code fence):

```
{"replacements": [{"slot_index": 0, "substring": "...",
                   "label": "..."}, ...],
 "faithful": {"substring": "..."}}
```

`replacements`: one object per failed contextomy slot (0 or 1); omit passing slots. `faithful`: include ONLY if faithful failed above; omit entirely otherwise.

## A.4 Validator Prompt

### System prompt.

You are a strict checker for contextomy crops. Given the `source_excerpt`, two contextomy substrings, and one faithful substring, check each against the rules below. Output JSON only.

Reject crops that lack a complete standalone clause — including:

- Predicate-only fragments (e.g. "will weaponize intelligence by leaking it" — no subject).
- Phrase fragments (e.g. "other than Day One" — no subject or verb).
- Dependent clauses alone (e.g. "when the deal closes" — nothing to attach to).
- Pronoun-only subjects whose referent does not appear inside the crop (e.g. "he said it was wrong" — who is *he* and what is *it*?).
- Demonstrative subjects with no in-crop antecedent (e.g. "this is what happened" — what is *this*?).
- Crops where the core noun phrase is implied but absent (e.g. "we've never gotten anything" — gotten what, *we* who?).

A crop passes Rule 2 only if a reader with NO prior context could write a specific, factual news headline given the speaker's identity, using only the words in the crop. If the headline would require the reader to guess a referent, the crop fails Rule 2.

[*Shared rules (§A.1) inserted here.*]

**Output** — JSON only, no markdown/preamble:

```
{"checks": [<contextomy_0>, <contextomy_1>, <faithful>]}
```

Each object: `{"rule_adheres": [7 booleans, one per rule 0-6], "reason": "R2 failed: ..." or null}`

### User prompt template.

```
SOURCE_EXCERPT:
{source_excerpt}

CONTEXTOMIES TO CHECK (index 0, 1):
0. substring: "{c0_substring}"
  label: {c0_label}
1. substring: "{c1_substring}"
  label: {c1_label}

FAITHFUL substring (index 2 —
representative of speaker intent):
"{faithful_substring}"

Output JSON with key "checks" (array
of 3 objects: two contextomies then
faithful, each with "rule_adheres" [7
booleans, R0-R6] and "reason").
```

**Rule applicability.** Rules 0–6 are not all applicable to every slot. Contextomy slots (indices 0 and 1) are evaluated on rules {0, 1, 2, 3, 4, 6}; rule 5 is faithful-only and the validator sets index 5 to `true` (N/A) for contextomies. The faithful slot (index 2) is evaluated on rules {0, 1, 2, 4, 5, 6}; rules 3 and 4 are contextomy- or cross-pair-only and the validator sets indices 3–4 to `true` (N/A) for the faithful substring.

### A.5 Quality Scoring Prompt

**System prompt.**

> You are evaluating contextomy crops for distortion strength. A contextomy is a verbatim substring that, when removed from its source context, produces a meaningfully misleading impression.
>
> For each contextomy, rate how misleading it would be if a reader encountered it without the original context, on this scale:
>
> **1** No distortion. The crop faithfully represents the speaker's meaning.
> **2** Minimal distortion. Slightly vaguer or less nuanced, but not misleading.
> **3** Moderate distortion. A reader would form a noticeably different impression than intended.
> **4** Strong distortion. A politically motivated actor could plausibly use this crop to advance a narrative.
> **5** Severe distortion. The crop inverts, fabricates, or dramatically misrepresents the speaker's position.
>
> Consider both directions: making the speaker look worse OR sanitizing them.
>
> Each `score` must be a single integer from 1 through 5 inclusive. Emit exactly one rating object per contextomy slot 0 and 1 (structured output enforces the schema).

**User prompt template.**

```
SOURCE_EXCERPT:
{source_excerpt}
CONTEXTOMY 0: "{c0_substring}" (label: {c0_label})
CONTEXTOMY 1: "{c1_substring}" (label: {c1_label})
Rate each contextomy for distortion strength (1-5 scale per system prompt).
```

**Output schema.** The quality scorer returns JSON conforming to:

```
{"scores": [
  {"slot_index": 0|1,
   "score": 1|2|3|4|5,
   "reasoning": "..."}, ...]}
```

with one entry per contextomy slot. Structured output enforces the score domain.

## B Context Loss Statement Audit

For each of the 918 (crop, full) pairs, we prompted Claude Sonnet 4.6 to instantiate all 15 context loss template statements, creating contextomy assertions from each contextomy dimension. To verify that generated statements faithfully instantiate their assigned templates, we conducted a manual audit on a stratified sample of 150 statements (10 per dimension, 1% of the 13,725 total). 2 Non-author annotators were shown each statement, its associated crop, and the template it was generated from, and label yes/somewhat/no whether the statement is a faithful instantiation of the template. We found that 89% of the statements were marked as faithful, 8.5% as somewhat faithful, and 2.5% as not faithful instantiations. When reviewing somewhat faithful and non faithful instantiations, we found the model sometimes rewrote the template or didn't make the template specific enough to the crop, but did not hallucinate off-template.

## C Annotation Task Design

**Structured vs. Free-Text Annotation.** Our final annotation task asks humans to rate agreement with LLM-authored contextomy statements rather than produce free-text descriptions of meaning shift, as in earlier pilot studies. This is a deliberate trade-off: free-text annotation at the scale of 918 excerpt pairs across 15 dimensions would be prohibitive both in annotator cost and in cognitive load per item, and would yield unstructured outputs that are difficult to aggregate or model. The structured format makes large-scale annotation tractable and produces directly comparable judgments across pairs and dimensions. The cost is that we measure agreement with a specific framing of meaning shift rather than independently elicited perception.

A single Prolific question is asked for all statements listed below:

> "Based on the material shown, do you Agree or Disagree with the given statement, or is it Unclear?"

**Sample Prolific Questionnaire**

**Crop None — Source 1 (Positionality Statements).**

**B00** I agree with the speaker's statements in the primary source. *(Agree / Disagree / Unclear)*

**B01** I think the speaker's statements in the primary source are rational. *(Agree / Disagree / Unclear)*

**B02** I am familiar with who the speaker is. *(Agree / Disagree / Unclear)*

**B03** I have a positive opinion of the speaker. (If you don't know them, mark *Unclear*.) *(Agree / Disagree / Unclear)*

**B04** The speaker's statements in the primary source are harmful or immoral. *(Agree / Disagree / Unclear)*

**B05** The primary source provides enough context to fairly evaluate the speaker's statements. *(Agree / Disagree / Unclear)*

**B06** The speaker's statements in the primary source are ambiguous or open to multiple interpretations. *(Agree / Disagree / Unclear)*

**B07** I feel a strong emotional reaction to the speaker's statements in the primary source. *(Agree / Disagree / Unclear)*

**Crop 1 — Source 1 (Inference Statements).**

**S00** This statement has been taken out of context. *(Agree / Disagree / Unclear)*

**S01–S15** LLM-generated inference statements. *(Agree / Disagree / Unclear)*

**S16** Attention check. *(Agree / Disagree / Unclear)*

**S17** This statement has been taken out of context. *(Agree / Disagree / Unclear)*

**Crop 2 — Source 1 (Inference Statements).**

**S00** This statement has been taken out of context. *(Agree / Disagree / Unclear)*

**S01–S15** LLM-generated inference statements. *(Agree / Disagree / Unclear)*

**S16** Attention check. *(Agree / Disagree / Unclear)*

**S17** This statement has been taken out of context. *(Agree / Disagree / Unclear)*

**Crop 3 — Source 1 (Inference Statements).**

**S00** This statement has been taken out of context. *(Agree / Disagree / Unclear)*

**S01–S15** LLM-generated inference statements. *(Agree / Disagree / Unclear)*

**S16** Attention check. *(Agree / Disagree / Unclear)*

**S17** This statement has been taken out of context. *(Agree / Disagree / Unclear)*

## D Annotation Task Instructions

**Task Name:** Evaluating How Cropped Quotes Change Meaning (Batch N)

### Task Introduction

When quotes are taken out of context—for example, in a headline or social media post—they can give a very different impression than what the speaker originally meant. A longer quote might reveal that the speaker was being sarcastic, responding to a specific question, or making a point that the short version leaves out entirely. This kind of distortion is called *contextomy*.

In this task, you will help us evaluate whether specific statements accurately describe how a cropped quote differs from its original source.

### Task Steps

**What You'll See.** On each page you will answer a question about either a positionality statement or an inference statement. The page will show the following:

- **Cropped Excerpt**—A short version of a quote, similar to what might appear in a headline or social media post.
- **Primary Source**—A longer passage from the original source that provides the full context around the cropped version.
- **Positionality Statement**—A statement about your relationship or attitude toward the source material. This field is populated for positionality questions and shows "None" for inference questions.
- **Inference Statement**—A claim about how the cropped excerpt changes or distorts the meaning of the source excerpt. This field is populated for inference questions and shows "None" for positionality questions.

**Instructions**

For every question in this task you will answer *Agree*, *Disagree*, or *Unclear* based on just a Primary Source or a Primary Source and one of three Cropped Excerpts. There is one Primary Source that is shared across all 62 questions in this task (8 positionality statements + 18 inference statements for each of 3 cropped excerpts = 62 total).

First, you will answer *Agree*, *Disagree*, or *Unclear* to 8 questions about your perspective on the Primary Source. Then, for each of the 3 Cropped Excerpts, you will answer the 18 questions comparing it against the Primary Source.

**Which question type am I on?** Check the Positionality Statement and Inference Statement fields. If the Positionality Statement has content and the Inference Statement says "None," answer based on your own view of the Primary Source. If the Inference Statement has content and the Positionality Statement says "None," answer based on how the Cropped Excerpt differs from the Primary Source.

**Positionality Statements.** For each Positionality Statement, read the primary source carefully, then decide whether you *Agree*, *Disagree*, or find it *Unclear* as a description of your own view, reaction, or knowledge. These statements are about *you*—your opinions, familiarity, and reactions—not about whether one excerpt distorts another. Answer honestly based on your genuine response to the source.

- *Agree*—The statement accurately reflects your view, knowledge, or reaction.
- *Disagree*—The statement does not reflect your view, knowledge, or reaction. If you partially agree but don't fully agree, select *Disagree*.
- *Unclear*—You don't have enough information to answer, or the statement itself is too ambiguous to judge against your view. Use this only when you genuinely can't form a response—not as a middle ground between *Agree* and *Disagree*.

**Inference Statements.** For each Inference Statement, read the cropped excerpt and source excerpt carefully, then decide whether you *Agree*, *Disagree*, or find it *Unclear* that the statement accurately describes a meaningful difference between the two excerpts.

- *Agree*—The statement correctly describes a way in which the cropped excerpt changes, distorts, or omits something meaningful from the source excerpt.
- *Disagree*—The statement does not accurately describe the difference between the two excerpts. If you partially agree but don't fully agree with the statement, select *Disagree*.
- *Unclear*—You don't have enough information in the two excerpts to determine whether the statement is accurate or not. Use this only when the excerpts genuinely don't give you what you'd need to judge—not as a middle ground between *Agree* and *Disagree*.

**Important Guidelines for the Inference Statements.**

- **Base your judgment only on what's written.** Compare the cropped excerpt and the source excerpt as they appear on the page. Don't rely on outside knowledge about the speaker, the event, or the broader political context.
- **Set aside your personal views.** Many of these excerpts involve public figures and politically charged topics. Your task is to evaluate whether the statement accurately describes the difference between the two excerpts—not whether you agree or disagree with the speaker's views (which is assessed in the positionality statements).
- **Read both excerpts fully before answering.** The differences can be subtle. A quick skim of the cropped excerpt or the primary source excerpt may cause you to miss important context that the crop leaves out.

Thank you for your careful attention—your annotations are essential to this research.

## E Annotator Demographics

Below we show demographic details for the 804 unique annotators on the Context Loss Inference Statement Task.

| Variable | Distribution |
|---|---|
| Age | 36.7 mean, 35 median, 18–78 (SD 12.2) |
| Gender | Woman 50.7%, Man 47.6%, Non-binary 1.6% |
| Sex | Female 51.5%, Male 48.5% |
| Ethnicity | White 31.0%, Black 22.8%, Mixed 17.4%, Asian 14.6%, Other 14.3% |
| LGBTQ | Yes 19.8%, No 80.2% |
| Political spectrum | Liberal 36.4%, Moderate 33.8%, Conservative 28.7% |
| Political affiliation | Democrat 37.8%, Independent 32.1%, Republican 24.8%, Other 1.9% |

## F Annotation agreement structure

The panel of (typically 5) annotations for each contextomy inference statement (i.e. across the 15 template dimensions) supports deeper investigation of the annotation structure. Here we explore the majority vote labels across the dimensions in comparison to the distribution of the individual annotator labels.

**Majority labels mask substantial disagreement.** Comparing the majority-label and individual-response distributions reveals systematic overstatement of consensus when (crop, hypothesis) pairs are collapsed to a single label. The aggregate *Agree* rate drops from 83.0% under majority voting to 67.4% over individual responses, a gap of 15.6 percentage points. *Disagree* rises correspondingly from 16.2% to 24.7%, indicating that meaningful annotator disagreement is suppressed by the aggregation. Most strikingly, *Unclear* responses, nearly invisible at the majority level (0.7%), account for 7.9% of individual judgments—a more than tenfold increase. This 7-point share of judgments that fail to resolve into any clear label is empirically inaccessible without a pluralistic view of the data. The shift is most consequential for Identity-Based Alienation: at the majority level its label distribution is essentially tied (49.1% *Agree* vs. 49.7% *Disagree*), but over individual responses it leans clearly toward *Disagree* (41.7% vs. 48.6%), suggesting that the dimension's apparent ambiguity at the majority level conceals a real underlying skew. Taken together, these gaps motivate modeling the distribution of annotator judgments rather than a single aggregated label.

## G Cross-Encoder Fine-Tuning Details

We did not perform an extensive hyper-parameter search, but the limited development process is described here.

| Template dimension | %Agree | %Disagree | %Unclear |
|---|---|---|---|
| Audience Relationship | 89.0 | 9.6 | 1.4 |
| Manipulative Framing | 77.8 | 22.0 | 0.2 |
| Emotional Tone | 83.2 | 16.3 | 0.4 |
| Ethical Stance | 85.0 | 13.9 | 1.1 |
| Supporting Rationale | 93.8 | 6.1 | 0.1 |
| Inferential Conclusions | 88.1 | 11.0 | 0.9 |
| Recommended Actions | 88.3 | 11.0 | 0.7 |
| Speaker Goals | 89.4 | 10.0 | 0.5 |
| Situational Context | 94.8 | 4.7 | 0.5 |
| Statement Salience | 73.0 | 26.6 | 0.4 |
| Identity Based Alienation | 49.1 | 49.7 | 1.2 |
| Sociocultural Context | 95.5 | 4.4 | 0.1 |
| Linguistic Qualifiers | 75.2 | 22.8 | 2.1 |
| Speaker Role | 83.9 | 15.3 | 0.9 |
| Speaker Portrayal | 79.5 | 20.2 | 0.3 |
| All Templates | 83.0 | 16.2 | 0.7 |

Table 8: Label distribution across the 15 template dimensions. The bottom row reports the aggregate across all dimensions.

In the **first round** of development, we used a standard learning rate of $2 \times 10^{-5}$ and 500 warmup steps in which the learning rate was linerally scaled. Models were trained for 10 epochs and we considered the best development set performance over the 10 epochs.

We experimented with two training paradigms for learning soft labels in the statement-level task: In the *sample classification* paradigm (SC), we treated each annotation as a separate datapoint for classification. The loss function in this parafim is the standard cross entropy loss function against a one-hot distribution representing the true label. In this setup that the model is exposed to multiple examples of the same input with potentially different ground truth labels. In the *distribution regression* (DR) paradigm, annotator responses for a given item (typically 5) are aggregated into a frequency distribution, and the loss function is the cross entropy between the model's prediction and the true frequency distribution. Note that while the loss function is different, the prediction head is architecturally identical in both cases: linear layer with 3 outputs and softmax activation function. We found very little performance difference on the statement-level development set between SC and DR-trained models.[3]

In the first round of development, we found that the best performance was often achieved after the first epoch of training, suggesting that the model may be overfitting to the data before the first epoch

[3] See Uma et al. (2021, §3.3) for a more complete discussion of these two techniques for predicting soft labels.

| Dimension | Agree | Disagree | Unclear |
|---|---|---|---|
| Audience Relationship | 70.6 | 19.6 | 9.7 |
| Manipulative Framing | 63.4 | 29.0 | 7.7 |
| Emotional Tone | 67.6 | 26.4 | 6.0 |
| Ethical Stance | 69.6 | 23.3 | 7.0 |
| Supporting Rationale | 77.3 | 15.8 | 6.9 |
| Inferential Conclusions | 71.1 | 21.3 | 7.6 |
| Recommended Actions | 72.9 | 19.6 | 7.6 |
| Speaker Goals | 72.4 | 21.0 | 6.6 |
| Situational Context | 77.0 | 14.6 | 8.5 |
| Statement Salience | 60.2 | 32.9 | 6.8 |
| Identity Based Alienation | 41.7 | 48.6 | 9.8 |
| Sociocultural Context | 77.7 | 15.3 | 7.0 |
| Linguistic Qualifiers | 57.8 | 30.2 | 12.1 |
| Speaker Role | 67.0 | 24.3 | 8.7 |
| Speaker Portrayal | 64.2 | 28.7 | 7.1 |
| All Templates | 67.4 | 24.7 | 7.9 |

Table 9: Distribution of individual annotator labels (%) across the 15 template dimensions, computed over all annotator responses rather than collapsed to majority labels. The bottom row reports the aggregate.

finished. In the **second round** of development, we experimented with longer warmup period of 2000 steps and a learning rate of $5 \times 10^{-6}$. This eliminated the tendancy for the first epoch to be the best, but the models continued to perform very poorly on minority class labels.

In the **third round** of development (models reported in the paper), we experimented with focal loss, with $\gamma = 5$, and using inverse class frequecy as the weight parameter ($\alpha$). These models performed better on the annotation-level (hard label) task, but more poorly on label distribution prediction. This is not surprising since the aggregated cross-entropy loss encourages the model to predict the true label distribution for a given input item, while focal loss is designed with a hard prediction (argmax) in mind. For that reason, we trained an equivallent set of models using cross entropy for the statement-level task.

All experiments were carried out on an NVIDIA RTX 6000 with PyTorch v.2.5.1, Transformers v.5.3.0 and SentenceTransformers 4.57.6. SC Models took about 3 hours to train (10 hours for ModernBERT-large) and DR models took about 45 minutes (2.5 hours for ModernBERT-large).

# H LLM Annotation Prompts

Note that annotation-level models without annotator demographics are identical to statement level SC models.

## H.1 No demographic information

```
You are predicting how a panel of 5 human
    annotators would label a claim about the
    relationship between a short crop of a quote
     and the longer original quote.

Your goal is NOT to decide whether the inference
     statement is objectively true. Your goal is
     to predict how five human annotators would
    vote.

Labels:
- agree: The inference statement accurately
    characterizes the relationship.
- disagree: The inference statement does not
    accurately characterize it.
- unclear: The statement is ambiguous, partially
     true, or open to interpretation.

Reasoning procedure:
1. Read the full quote and identify the speaker'
    s meaning and framing.
2. Read the crop and infer what a reader seeing
    only the crop would conclude.
3. Compare the crop to the full quote.
4. Evaluate whether the inference statement
    accurately describes that relationship.
5. Predict how five annotators would distribute
    their votes.

Calibration guidance:
- Mechanical claims with clear textual support
    should yield concentrated counts.
- Tone or framing judgments may produce split
    votes.
- If a statement is partially true or overstated,
     "unclear" may be appropriate.
- Do not assume the inference statement is
    correct.

Return valid JSON only in this exact format:

{
  "reasoning": "...",
  "calibration_note": "...",
  "predicted_distribution": {
    "agree": N,
    "disagree": N,
    "unclear": N
  }
}

The counts must be non-negative integers summing
     to 5.
```

## H.2 With annotator demographics

```
You are predicting how a specific human
    annotator would label a claim about the
    relationship between a short crop of a quote
     and the longer original. You will be given
    the annotator's demographic profile along
    with the task.

Background.
Short crops of longer quotes are common in media
     and online discussion. Sometimes a crop
    preserves the speaker's intended meaning
    faithfully; sometimes it shifts or distorts
    that meaning by omitting context.
```

```
Your task.
Given:
- the full quote,
- the crop,
- one inference statement,
- and a demographic profile for one specific annotator,
predict the label this annotator would assign.

Your goal is NOT to decide whether the inference statement is objectively true.
Your goal is to predict how THIS annotator would label it.

Labels:
- agree: the inference statement accurately describes the relationship.
- disagree: the inference statement does not accurately describe the relationship.
- unclear: the statement is ambiguous, partially true, interpretive, or difficult to judge cleanly.

Guidance on demographics:
- Demographics may or may not matter for a given example.
- Some politically or culturally charged examples may interact with the annotator profile.
- Avoid crude stereotypes.
- Use demographics only when they plausibly shape interpretation.
- Many examples should primarily be decided from textual evidence alone.

Reasoning procedure:
1. Read the full quote and identify the speaker's meaning and framing.
2. Read the crop and infer what a reader seeing only the crop would conclude.
3. Compare the crop to the full quote.
4. Evaluate whether the inference statement accurately characterizes that relationship.
5. Consider whether this annotator's demographic profile might influence interpretation.
6. Predict the annotator's label.
7. Produce a probability distribution over the labels.


Calibration guidance:
- Mechanical textual claims should often yield concentrated probabilities.
- Tone and framing judgments may yield more uncertainty.
- If a statement is partially true or overstated, uncertainty may increase.
- Do not assume the inference statement is correct.

Return valid JSON only in exactly this format:
{
  "reasoning": "...",
  "demographic_consideration": "...",
  "predicted_label": "agree" | "disagree" | "unclear",
  "probability_distribution": {
    "agree": p1,
    "disagree": p2,
    "unclear": p3
  }
}

The probabilities must be non-negative floats summing to 1.0.
The predicted_label must correspond to the largest probability.
```