Language Models as Measurement Apparatus for Culture

Kent K. Chang
School of Information
University of California, Berkeley
kentkchang@berkeley.edu

Abstract

Language models are increasingly used to quantify cultural phenomena, but what makes such measurement distinctively *cultural*? This paper argues that NLP work on culture is a *material-discursive practice*: the apparatus—model, data, annotation, evaluation—participates in constituting the cultural reality it measures, rather than passively recording it. Drawing on Karen Barad's concept of the *agential cut*—the contingent boundary between phenomenon and instrument—I show that the apparatus's substantive design choices draw such boundaries, and that the boundary is entangled from the start because language models have already internalized much of the cultural material they measure. I illustrate this through three case studies on television and film dialogue and two examinations of the apparatus itself: erasure of character names as cultural markers, and attunement to historically distant Restoration drama. This big picture analysis proposes a research program that is theory-driven, empirically rigorous, and culturally contingent, treating each agential cut as a conscious commitment.

1 Introduction

A growing body of natural language processing (NLP) research engages with cultural objects, often under the rubric of cultural analytics: literary texts, social media, and other artifacts whose significance is irreducible to information content (Piper, 2016, 2017) and constitutes a symbolic form (Cassirer, 2014). For instance, word embeddings have traced historical shifts in cultural concepts (Garg et al., 2018; Hamilton et al., 2016), which has been shown to be robust for humanistic inquiries (Zhou et al., 2025a): their contextualized variants have mapped the geometry of social meaning (Kozlowski et al., 2019; Lucy et al., 2022); connotation frames have measured implicit power and agency in film dialogue (Sap et al., 2017); computational sociolinguistics has modeled stylistic coordination in dialogue (Danescu-Niculescu-Mizil and Lee, 2011); and large language models have been probed for cultural knowledge (Chiu et al., 2025). In this light, this work leverages the affordance of NLP methods to address cultural questions, attending to both empirical rigor and interpretive depth such methods enable. What remains incomplete in this big picture is an explicit account of what it means to *measure* culture—as opposed to measuring sentiment, or syntax, or factual accuracy.

Recent work has begun to address this gap: Zhou et al. (2025b) deftly draws on sociocultural linguistics to argue that cultural NLP needs a coherent theory of culture grounded in indexicality, positionality, and emergence. Building on this, this paper offers a big picture analysis to expound on what it means to *measure* culture with language models, asking: What happens when a language model is used as an instrument of cultural measurement? I argue that NLP work on cultural objects constitutes a *material-discursive practice* (Barad, 2007; Brown and Duguid, 2000): the material configuration of the apparatus (model architecture, training data) and the discursive framework of the researcher (annotation categories, evaluation criteria, interpretive commitments) are entangled in the measurement and inseparable from it.

To make this concrete, I develop the concept of the *agential cut* (Barad, 2007) for NLP: the contingent boundary that an apparatus enacts between what counts as phenomenon and what counts as instrument. In using computational methods to study culture, every design choice—model architectures, taxonomies for classification, adjudication of annotation—draws such a boundary, and the boundary could always have been drawn differently. At the same time, what makes language models distinctive when applied to cultural artifacts is that they often have already encountered snippets (Chang et al., 2023b) or summaries of the cultural material they measure during pre-training:

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 131–143
July 4, 2026 ©2026 Association for Computational Linguistics

the boundary between instrument and object is entangled from the start. As large language models (LLMs) are increasingly used for social and cultural measurements (Bamman et al., 2024; Halterman and Keith, 2025), this entanglement raises a problem that is prior to, and distinct from, the representational gap between data and cultural reality (Bode, 2020): while data remains a partial construction of the world, the LLM, the measuring instrument, has already internalized the very material it is asked to measure. The case studies in this paper return to this entanglement repeatedly.

I develop the argument through three case studies from my dissertation on dialogic interactions found in film and television, a site where social identities are constructed and contested. Each case study represents a type of cultural measurement: *structure* (conversation disentanglement), *interaction* (role attribution and gender), and *deviation* (stereotypic relation extraction). The measurements produced—gendered patterns in conversational agency, disparities in role attribution, the formalization of subversion—are constitutively contingent. In treating language models as measurement apparatus, I hope those case studies demonstrate how rigor and contingency are not in tension but mutually constitutive—and acknowledging this entanglement is the precondition for productive research at the intersection of NLP and culture.

2 Measurements of Culture

2.1 Operationalization and measurement

The dominant framework for computational work on culture derives from the social sciences, where *operationalization*—translating a theoretical concept into a measurable variable—sits at its core. Franco Moretti influentially imported this into digital humanities (Moretti, 2013), and subsequent work has refined the call for the computational study of culture to be explicit about its methodological commitments (Piper, 2017; Underwood, 2019). Two recent positions extend this concern to AI. Wallach et al. (2025) argue that evaluating generative AI systems is fundamentally a measurement challenge in which constructs such as helpfulness, fairness, or harm cannot be treated as natural labels but must be defined, instrumented, and validated.

Measurements of culture, in the sense developed below, take up the same problem fromyet a third angle. The question is not merely one of how we measure AI systems, but also how AI systems measure cultural artifacts. Plays, novels, films, and television scenes are not raw observations of social variable; they are organized by genre convention, historical context, audience reception, and critical interpretation. Computational methods can make such artifacts measurable at scale, but only by making them legible in some particular way: as character lines, reply graphs, role labels, relationship types, or deviations from a learned norm.

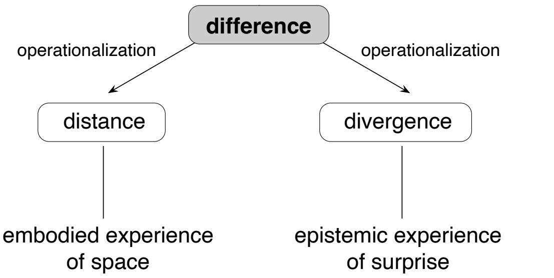

Figure 1: Two paths to operationalizing the concept of difference in computational work on culture (Chang and DeDeo, 2020): distance (a spatial metaphor, subject to metric axioms including symmetry) and divergence (a cognitive metaphor, capturing asymmetric relationships invisible to distance). The choice between them is itself an agential cut that determines which cultural relationships become measurable.

As Piper (2017) argues in his framework for literary modeling, operationalization is itself a reduction—and the reduction is where the measurement happens. Consider a simple example of operationalization—measuring how much two texts differ. Distance metrics like cosine afford spatial proximity arguments and inherit the symmetry of a metric space; divergences like Kullback–Leibler afford asymmetric arguments about encoding cost and surprise (Fig. 1; Chang and DeDeo, 2020). The two operationalize the same word—*difference*—into different cultural facts. The choice between them determines which cultural relationships are made measurable.

Operationalization is productive, but it carries an assumption: that the concept being operationalized exists independently of the measurement procedure. Recent work in cultural analytics questions this: McNulty and Chapot (2025) reconsider the relationship between computation and form in the wake of generative AI, treating models themselves—their architectures and outputs—as cultural-technical "forms" that both enable and require new modes of analysis. Dobson (2025) argues that architecture is not a neutral container but a substantive interpretive choice—a site where meaning is made and

Type	Apparatus configuration	Cultural question	Source of contingency
Structure	Multi-party dialogue → directed reply-to graph (Chang et al., 2023a)	How does conversational agency emerge from the **topology of who-responds-to-whom**?	architecture and input representation define what counts as a thread
Interaction	Audio-visual signal → Goffmanian role labels (Chang et al., 2026)	How are speaking and listening roles **distributed across participants**?	category taxonomy and modality selection determine which roles are measurable
Deviation	Dyadic dialogue → stereotypic relationship type (Chang et al., 2024)	Where does performed interaction **depart from normative expectation**?	norm is learned from training; deviation exists only relative to a trained baseline

Table 1: Three modes of cultural measurement, organized by the type of agential cut the apparatus enacts.

historicity registered. These observations suggest a framework that treats the entire configuration—model, data, annotation, evaluation—as a larger, coherent whole: indeed an *apparatus* that participates in producing its object, not one treating data, task, and algorithm as separable components of a linear research narrative.

2.2 Entanglement and cuts

The linear research narrative—task, data, model, metric—works for many NLP problems by treating these components as separable stages. For cultural measurement, that separation collapses from both sides. Decisions about what to model—how to represent a fictional character, what counts as an interaction, which categories to annotate, how much context to expose—are not preliminary to the measurement but constitutive of it; they determine which cultural realities can emerge and which are foreclosed.

Crucially, the instrument itself is culturally formed: what an LLM knows about culture is what circulates, and what circulates is structured by prestige and the cultural industries long before any researcher uses the model to measure. The problem is not that the model has this cultural past—for measurements of culture, some past is often necessary—but whether that past is acknowledged, tested, and interpreted as part of the apparatus. This is related to, but distinct from, the ontological gap between data and cultural reality that Bode (2020) identifies.

In machine learning terms, decisions about what to model are usually treated as task design, and the cultural formation of the instrument as contamination or memorization (Mallen et al., 2023) that puts pressure on model reliability. Those terms are useful, but too narrow: they treat the apparatus and its object as separable when, for cultural measurement, they are not. From Barad's (2007) reading of Niels Bohr, I take the concept of the *agential cut*: the boundary an apparatus enacts between what counts as phenomenon and what counts as instrument. I use the term here in a constrained methodological sense—not every implementation detail is an agential cut. Random seeds, batch sizes, choice of GPU vendor, logging verbosity: these do not, in any normal range, change what cultural phenomenon can appear in the measurement. A design choice becomes a cut when it does: the label taxonomy, the context window, the modality, the anonymization procedure, the training data, or the norm against which deviation is measured.

The case studies that follow each enact a different cut (summarized in Table 1). Conversation disentanglement cuts continuous dialogue into a directed reply-to graph (Chang et al., 2023a). Conversational role attribution cuts an audiovisual scene into speaker, addressee, and side-participant roles (Chang et al., 2026). Stereotypic relation extraction cuts a trained expectation into a norm against which performance can depart (Chang et al., 2024). Each measurement is empirical, but none is independent of the apparatus that produces it. Framed like this, cultural analytics seeks to hold together the positivist work (of building models that shed light on culture) with the critical work (of insisting on the contingency of every cultural question those models help us ask).

3 Measuring Structure

The first type of cultural measurement involves imposing a formal structure on multi-party dialogue. The agential cut here is *structural*: any conversational exchange can be formalized in multiple ways—as a sequence of turns, a tree of reply-to links, a network of topic threads—and each formalization draws a different boundary between what counts as "structure" and what is relegated to noise.

Conversation analysis has long studied the systematics of turn-taking (Sacks et al., 1974) and the collaborative work of speakers and hearers (Goodwin, 1981); choosing among formalizations is an interpretive commitment about which of those systematics the apparatus will be allowed to see. Choosing reply-to graphs makes conversational floor, address, and initiation visible at scale; it forecloses lexical cohesion, topical drift, and the slow buildup of mutual understanding that fluent dialogue depends on. The gender measurements that follow are visible only inside this cut—they would not survive a re-formalization that, for example, weighted topic continuity over reply structure.

3.1 Conversation disentanglement

The structural formalization I adopt here is *conversation disentanglement*: recovering the thread structure of interleaved multi-party dialogue, a task studied extensively in NLP on IRC chat logs (Elsner and Charniak, 2008; Kummerfeld et al., 2019; Jiang et al., 2018; Zhu et al., 2021). In Chang et al. (2023a), we extend this to scripted multi-party dialogue, developing a BERT-based model (Devlin et al., 2019) that predicts which prior utterance each line responds to, thereby recovering a latent thread structure. Formally, given a sequence of utterances $\{u_1, \ldots, u_n\}$, the model encodes each utterance contextually and scores candidate reply-to links:

$$P(\text{parent}(u_i) = u_j) \propto \exp\big(g(\boldsymbol{h}_i, \boldsymbol{h}_j)\big), \quad (1)$$

where $\boldsymbol{h}_i$ is the contextual representation of utterance u_i and g is a learned scoring function. The result is a directed graph—a thread structure—extracted from continuous dialogue.

In my framework, this is an apparatus-dependent measurement: the threads do not pre-exist in the script but are produced by the apparatus. Here, the start of a thread is grounded in observations in television studies: McKee (2016) argues that speech acts are driven by character need: "all talk responds to a need, engages a purpose, and performs an action." This is central to our annotation scheme, which itself is part of the apparatus: what counts as a "reply," whether breaks a thread or continues it—these are not neutral transcription decisions but interpretive commitments that shape the thread graph the model produces.

At the same time, the original work includes exhaustive experiments across architectures and inpuft representations—different encoders, different context windows, different representations of speaker identity—each constituting a different apparatus configuration applied to the same dialogue. These are design choices that, in addition to enabling comparison between methods, crucially reflect what we think constitutes a speaker that shapes the ensuing analysis. Each combination of annotation scheme, encoder, and input representation enacts a different agential cut on the same underlying dialogue—and the cultural patterns that emerge are inseparable from the apparatus that produced them.

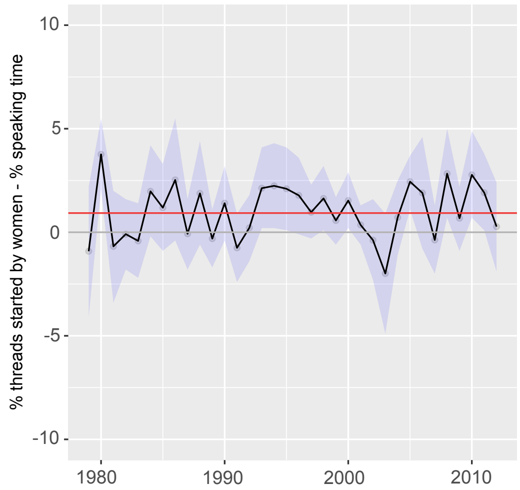

Figure 2: Percentage of conversational threads started by female characters minus their share of speaking time, by year, with 95% confidence intervals (shaded) (Chang et al., 2023a). The red line marks the overall average of $+1.0$ percentage points ($p < 0.05$): despite their under-representation, women initiate threads at a rate that slightly exceeds their speaking time.

3.2 Measurements of agency

Applied at scale to TV and movie transcripts, the resulting thread structures enable measurements of conversational agency: who initiates threads, who sustains them, and how these patterns distribute by gender. In our data from 808 movies, 30.4% of threads are started by female characters—consistent with the well-documented disparity in women's presence and voice in media (Rosen, 1973; hooks, bell, 1992). However, when normalized by each character's share of speaking time (Fig. 2), women initiate threads at a rate that slightly exceeds their speaking time, with an average absolute difference of $+1.0$ percentage points ($p < 0.05$). This finding is surprising: despite their under-representation, female characters are written to claim the conversational floor more often than their male counterparts would predict.

These findings demonstrate what structural measurement—an affordance of NLP applied to cultural objects—makes possible: by committing to a particular formalization of conversational structure, the apparatus renders visible patterns of agency that would be invisible in the unstructured transcript. A viewer watching a scene experiences dialogue as a continuous flow; the apparatus transforms it into a graph whose structural properties—who initiates, who sustains, who is peripheral—can be counted, compared, and statistically tested.

4 Measuring Interaction

The study of how language use varies with interactional context has deep roots: Ervin-Tripp (1964) showed that speakers shift register depending on topic, setting, and listener; Ng and Bradac (1993) documented how verbal behavior both reflects and constitutes power asymmetries; and studies of scripted dialogue have long recognized that television writing encodes—and sometimes contests—these dynamics (Richardson, 2010; Bednarek, 2023).

To address such cultural questions, the second type of cultural measurement involves classifying participants into sociolinguistic categories—speaker, addressee, side-participant—that formalize who participates and how. The agential cut here is *categorical and modal*: the apparatus partitions a continuous audio-visual stream into a finite set of discrete role labels, and the choice of which roles to recognize—and which modalities to admit as evidence—determines what aspects of interactional positioning can register at all.

4.1 Conversational role attribution

Multimodal video understanding has produced large-scale datasets for television—most notably TVQA (Lei et al., 2018), which benchmarks compositional question answering over TV clips. But existing datasets treat dialogue as a source of answers rather than as an interactional system with its own structure. To address this gap, we operationalize the sociolinguistic organization of conversation itself to devise an annotation scheme, culminated in TV-MMPC (Chang et al., 2026).

The annotation scheme is itself an agential cut: Goffman's theorization of conversation participants, along with Clark and Carlson's (1982) taxonomy of speakers and hearers, is mapped onto discrete labels assignable to individual utterances in scripted television, so the model is tasked, for each utterance, to predict its speaker, intended addressee, and side-participants. This cut turns the continuous, multimodal flow of conversation into a discrete assignment of roles, and the choice of role categories is itself a discursive commitment: it reflects a interpretive commitment about which aspects of interactional positioning matter enough to measure.

In this particular case, the models evaluate a range of models: a text-only model discards everything non-verbal; a multimodal model takes in the full audio-visual signal and maps it into the same label space. These are different apparatuses measuring the same phenomenon, and they produce different results—not only in model performance, but in what counts as a relevant signal: which modalities the researcher selects, which frames the vision–language model samples, whether audio and video are processed jointly or separately, and what computing resources make feasible. For Barad, those are all agential cuts enacted by human and non-human actors shaping the apparatus, and consequently, the measurements it produces.

4.2 Measurements of gendered roles

Applied to 350,842 utterances across four TV series in TVQA, the apparatus produces measurements of how interactional roles distribute by gender. To quantify this, we fit a multinomial logistic regression that estimates the log-odds of occupying each role, controlling for show-level effects. With speaker as the reference category and $j \in \{\text{addressee, side-participant}\}$:

$$\log \frac{P(\text{role} = j)}{P(\text{role} = \text{speaker})} = \beta_{j,0} + \beta_{j,\text{female}} \cdot \mathbb{I}(\text{female}) + \sum_{s=1}^{S-1} \gamma_{j,s} \cdot \mathbb{I}(\text{show}_s), \quad (2)$$

This is itself a measurement: from the high-dimensional space of individual role attributions to a two-dimensional summary (one odds ratio per non-speaker role) that isolates the effect of gender. The resulting odds ratios—$\exp(\beta_{j,\text{f}})$—are 1.19 for addressee and 1.20 for side-participant (both $p < 0.001$), indicating that women are approximately 1.2 times as likely as men to appear in a listening role rather than as speaker, after controlling for show.

These measurements provide evidence of how social roles and power dynamics are constructed—and reinforced—in cultural artifacts. They are visible only because the apparatus includes the categories of addressee and side-participant: the theoretical commitment to distinguishing listening roles is what makes the gendered pattern measurable.

5 Measuring Deviation

The third type of cultural measurement involves a normative baseline and quantifying departures from it. Where structural measurement asks "what is the form?" and interaction measurement asks "who participates how?", deviation measurement asks "where does practice depart from expectation?" The agential cut here is *normative*: the apparatus learns a baseline of stereotypic patterns from training data, and what counts as "subversion" exists only relative to that learned norm. A different training corpus would yield a different baseline; a different label inventory—one that included `queer-platonic` or `enmeshed sibling`—would yield different deviations. What this cut makes measurable is stereotypicality and its breach; what it forecloses is meaning that neither stabilizes as a stereotype nor disrupts one—the in-between cases that read as simply ordinary.

Deviation, then, is a relational property: between texts, and between text and apparatus. This case study takes inspiration from cognitive stylistics, what Culpeper (2001) calls "stereotyping": the textual cues by which dramatic figures are constructed as social beings provide the signal, and the model's trained expectations provide the norm against which deviation becomes measurable.

5.1 Stereotypic relation extraction

In Chang et al. (2024), we train a Longformer (Beltagy et al., 2020) encoder on 787 digitized pilot teleplays to predict the social relationship enacted in a dyad's dialogue. Given a scene $\mathcal{S}$ with utterances from a head character c_h and tail character c_t, the model builds a joint representation. A Longformer encoder extracts the CLS token for each speaker's concatenated utterances. To incorporate scene context beyond the target speakers' words, an attentive pooling mechanism weights the hidden states of the full scene, guided by a token-level mask M that zeros out the target speakers' tokens:

$$\alpha = \mathrm{softmax}\big(\boldsymbol{w}_A{}^\top \boldsymbol{h}_\mathcal{S} \odot M\big), \tag{3}$$

$$\boldsymbol{h} = \big[\boldsymbol{e}^{c_h}_{\texttt{<s>}};\ \boldsymbol{e}^{c_t}_{\texttt{<s>}};\ \boldsymbol{h}_\mathcal{S}{}^\top \alpha\big], \tag{4}$$

where $M[j] = 0$ if token j is spoken by either target speaker and 1 otherwise, $\boldsymbol{h}_\mathcal{S}$ is the encoded scene, and $\boldsymbol{w}_A$ is a learned attention vector. The concatenated representation $\boldsymbol{h}$ is projected through a linear classification head:

$$p(r|\boldsymbol{h}) = \mathrm{softmax}(f(\boldsymbol{h})), \tag{5}$$

predicting a relationship type r from a fixed set (e.g., `colleague_of`, `sibling_of`, `spouse_of`).

The architecture enacts a specific agential cut. The mask M directs the model to attend to what *other* characters say and do in the scene—the ambient social context—rather than relying solely on the target dyad's words. This is a deliberate choice about what the apparatus should treat as signal: the broader conversational ecology, not just the dyad in isolation. Character names are anonymized to force the apparatus to operate on dialogic cues—*how* characters talk—rather than on memorized character-relationship associations (see §6 for the empirical consequences of this choice).

5.2 Measurements of subversion

In traditional NLP, model quality is measured by accuracy: how well predictions match ground-truth labels. This evaluation paradigm treats correct labels as the goal and errors as failures to be minimized. But for cultural measurement, this logic inverts—a lesson that working at the intersection of NLP and cultural analysis has made unavoidable. Metrics like accuracy are measurements of *performance*; what cultural analysis requires are measurements of *interpretive significance*.

The model in Chang et al. (2024) is deliberately trained as a "stereotyping reader": it learns what sibling dialogue, or spouse dialogue, *typically* sounds like across hundreds of teleplays. Rather than evaluating accuracy as an end in itself, the work measures *subversion* as the discrepancy between the model's predicted distribution over relationship types and the ground-truth labels. When a model trained on stereotypical patterns of sibling dialogue predicts that two brothers sound like a married couple, the "error" is the finding. The gap between prediction and reality is the very signal to be analyzed—a measurement of how interaction departs from the norm the apparatus has learned.

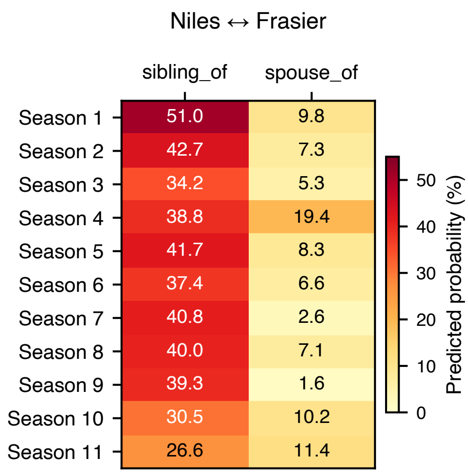

Figure 3: Percentage of predictions for `sibling_of` (ground truth) and `spouse_of` between Niles and Frasier Crane across *Frasier* (Chang et al., 2024).

Applied to *Frasier*, the apparatus identifies the Crane brothers' dialogue as consistently deviating from stereotypical sibling interaction. The model frequently predicts `spouse_of` for Frasier and Niles's exchanges—capturing, and formalizing, what queer theorists would recognize as a form of intimacy that exceeds its nominal category (Sedgwick, 2003; Halperin, 2002): the "crypto-gay" (Clum, 1999) quality that cultural critics have noted in their bickering, codependent, linguistically ornate relationship (Fig. 3). The apparatus produces a formal trace of the gap between performed relationship and normative type that can then become an object of cultural analysis. Butler's (1990) notion of the"subversion of identity" here becomes computationally tractable: notably, it is not meant to be an ontological claim about the characters' sexuality, but as a measurable discrepancy between dialogic performance and the model's trained expectations.

6 Between Measuring Instruments and Cultural Artifacts

The three case studies above focus on what NLP can measure when the cultural material may have been heavily represented in the model's training distribution (well-known TV characters and dialogues from popular series). In this section, we consider two complementary studies to further characterize what the apparatus is actually doing in those measurements: *erasure*, removing the apparatus's purchase on its memory of the cultural artifact in question to reveal what was being measured (§6.1); and *attunement*, deliberately training the apparatus on a corpus it has likely not absorbed, to see what measurement looks like when the apparatus likely has no memory of its object to draw on (§6.2). Together they describe the apparatus from both sides—what it brings to a culturally familiar text and what it has to build for a culturally distant one—and show where the boundary between the measuring instrument and the cultural artifact it measures is actually being drawn.

At issue in both cases is the apparatus's stake in the cultural material it is tasked to measure: what it has absorbed of those things, and what it has not. Brown and Duguid (2000) argue that knowledge cannot be separated from the social practices that produce it: information becomes knowledge only when embedded in a practice that gives it meaning. The apparatus's prior knowledge of, say, *Frasier* is not just stored in its weights but lives in the practice that mobilizes those weights to make a measurement about Frasier and Niles. Here I want to foreground the material side of the apparatus: In Barad's sense, the physical-computational infrastructure—parameters, training data, what the model has absorbed and what it has not. The discursive practice (which question is being asked, what counts as a finding, why these annotation categories) is equally constitutive of how any material result becomes a cultural measurement.

6.1 Erasure

The clearest demonstration of memorization shaping measurement comes from perturbing the apparatus. In the multimodal conversation structure study (§4), replacing character names with anonymous identifiers (e.g., replacing `Sheldon Cooper` to `Character C`) collapses speaker recognition from 78.6 to 13.7 and addressee recognition from 68.1 to 15.7 (Table 2, a). The same model, on the same data, appears to rely on its parametric memory for role attribution. What appeared to be a boundary between a neutral model and a conversational structure was a boundary between two layers of the model's own cultural knowledge: its general language competence and its specific memory of these characters, invoked by their names. Anonymization forces a different cut, one that separates dialogic structure from character identity, and the performance collapse reveals which cut was operative all along.

The same pattern recurs in stereotypic relation extraction (§5; Table 2, b). The Longformer model

Task	Orig.	Anon.	Δ
(a) Multimodal conversation structure			
Gemini 2.0 Flash			
Speaker (Acc)	78.6	13.7	−64.9
Addressee (F_1)	68.1	15.7	−52.5
(b) Stereotypic relation extraction			
Longformer			
Role (Acc)	34.8	24.8	−10.0
+ anon. training	36.7	33.8	−2.9
LLaMA 3-70b			
Role (Acc)	24.3	19.7	−4.6

Table 2: Effects of anonymization across two studies on different test sets: original ("Orig.") and anonymized ("Anon."). Panel (a): for multimodal conversation structure (Chang et al., 2026), speaker and addressee recognition collapse under anonymization; panel (b): for stereotypic relation extraction (Chang et al., 2024): the scene attentive pooling model without anonymized training shows a 10-point accuracy gap; training on anonymized data recovers most of the performance.

trained only on unanonymized data drops 10 points when evaluated on anonymized test data (from 34.8 to 24.8), indicating that the model exploits character names to infer memorized relationships rather than parsing the dialogue. Training on anonymized data recovers most of this gap, which forces the model to attend to how characters talk rather than who they are. LLaMA 3-70b (Grattafiori et al., 2024) zero-shot drops from 24.3 to 19.7 under anonymization, a smaller gap than Gemini's but the same pattern nonetheless. Prompt-based LLMs lack the Longformer's most direct mitigation (adaptation with anonymized data), though input-side anonymization can offer partial alternatives.

These anonymization experiments contextualize the cultural findings from the case studies: the measurements of gendered role attribution and relational subversion are shaped by the model's prior cultural formation. When Gemini attributes a speaking role or LLaMA predicts a character relation, the output reflects the cultural material internalized during training as much as the signal in the input. The material—what the model was trained on, what cultural knowledge its parameters encode—and the discursive—which categories the researcher defines, which tasks the model performs—are entangled in every measurement. This entanglement is at the heart of using language models as measurement apparatus for culture: the instrument might have already internalized some of the cultural material it measures, and the research narrative must account for it, rather than treat data, model, and evaluation as clearly separable stages of a pipeline.

6.2 Attunement

The anonymization analysis showed what the apparatus does when the cultural material may be available in training data. The inverse case is more revealing about what the apparatus *is* when there is no shortcut through prior knowledge, which I explore here in the context of Restoration comedy (1660–1700). Restoration drama is interesting here because it is well-studied in history and criticism but computationally under-explored, in part because the professionally curated and digitized editions sit behind proprietary access. The language is historically distant, and an off-the-shelf LLM, especially a smaller one, has at best a thin grasp of its conventions.

Restoration comedy of manners is driven by characters and archetypes (such as *rake* and *fop*). That said, we do not know *a priori* how many lines we need to have read for a character's archetype to become recognizable, which resembles a sufficient-context problem (López-Monroy et al., 2018) in which the reader does not recognize an archetype all at once but accumulates evidence as the play unfolds, committing when predictions have stabilized enough to act. This problem lets us specify two dimensions of the apparatus directly: what its parameters have been tuned to, and how it reads.

The corpus for this toy experiment is a random sample of 109 plays (1,283 character episodes) from Chadwyck-Healey English Drama collection;[1] each episode $e = (x_{1:T}, y)$ is an ordered sequence of one character's lines. Formally, a backbone f encodes prefix $x_{1:t}$ into a representation $\boldsymbol{h}_t = f_\theta(x_{1:t})$, from which a classifier predicts an archetype label c:

$$p_\theta(c \mid x_{1:t}) = \mathrm{softmax}\big(\boldsymbol{W}\boldsymbol{h}_t\big)_c, \quad c \in \mathcal{C}. \quad (6)$$

Drawing on taxonomies defined in Hirst (1979); Mast (1975), we include the following in $\mathcal{C}$: Rake, Fop, Natural, Obstacle, to be predicted from a character's dialogue alone, absent any metadata. The source data has no archetype labels; we developed them iteratively—reading Restoration criticism, hand-annotating, refining prompts—and then scaled with Gemini 2.5 Flash (Gemini Team et al.,

[1]See Appendix A for more details.

2023), achieving Cohen's $\kappa = 0.71$ against the author's annotation of five plays.

To test this apparatus, we compare two reader-models on the same 1B-parameter Gemma backbone (Gemma Team et al., 2024)—a fixed-window reader (first N lines) and the entropy-thresholding reader together with a majority-class baseline. Entropy thresholding models the epistemic experience of the reader: the apparatus reads sequentially, stops when predictive entropy falls below δ, and predicts the maximum-probability class:

$$H_t = -\sum_{c \in \mathcal{Y}} p_\theta(c \mid x_{1:t}) \log p_\theta(c \mid x_{1:t}), \quad (7)$$

$$\tau_\delta(e) = \min\{t : H_t \leq \delta\}, \quad (8)$$

$$\hat{y}(e) = \arg\max_c p_\theta(c \mid x_{1:\tau(e)}). \quad (9)$$

For evaluation, we track macro-F_1 alongside an efficiency-aware objective $\mathcal{J} = \text{macro-F}_1 - \lambda \cdot \overline{\rho}$ ($\lambda = 0.25$ by default, $\rho(e) = \tau(e)/T$ denotes the fraction of lines consumed), which prices each unit of reading against the prediction it enables.

Attunement comes through one epoch of continued pre-training on roughly 174,000 lines of in-domain dialogue, in the spirit of historically attuned LMs (Manjavacas Arevalo and Fonteyn, 2021). The effect is consistent across reader-models: $\mathcal{J}$ rises from 0.17 to 0.29 for fixed-window and from 0.25 to 0.37 for entropy thresholding. The attuned entropy-thresholding reader reaches macro-F_1 0.44 reading only 26% of lines, against the fixed-window reader's 0.38 at 36% and a majority baseline of 0.14. Attunement gives the apparatus a thinner, deliberately-built version of the cultural past; the apparatus itself is layered: a human-calibrated annotator-LLM defines the norm against which an attuned reader-LLM is measured.

Taken together, §6.1 and §6.2 examine the same instrument from two angles. In the first, the apparatus's prior cultural formation is already there and can be perturbed; in the second, it has to be built up, and even built carefully it cannot reach beyond the cuts that defined it. In both cases, the measuring instrument and the cultural artifact are entangled in every result; neither produces it alone.

7 Conclusion

In this paper, I have argued that NLP work on cultural objects is a material-discursive practice in which the apparatus participates in producing the phenomena it measures. Three case studies (§3–§5) developed this along three dimensions of cultural measurement—*structure*, *interaction*, and *deviation*—and §6 examined the apparatus directly through *erasure* and *attunement*. A pattern emerges across both: each study is *theory-driven*. Conversation disentanglement draws on pragmatics and conversation analysis (Sacks et al., 1974; Goffman, 1963); interaction measurement on Goffman's (1981) participation framework; stereotypic relation extraction on cognitive stylistics (Culpeper, 2001), as well as queer theory's attention to the subversion of normative identity categories (Butler, 1990; Sedgwick, 2003); and attunement on dramatic criticism of comedy of manners (Hirst, 1979; Mast, 1975). Each measurement is also *culturally contingent*: the findings depend on the specific apparatus through which they were produced.

What sets the measurement of culture apart from other forms of task and evaluation is that the instrument itself is culturally situated: it carries the training distribution's biases, its era's textual archive, its architecture's affordances. The concepts of agential cut and material-discursive practice provide a framework for taking this reflexive entanglement seriously. This framework has implications for how cultural measurement should be practiced: Accuracy and model performance remain necessary, but they are insufficient. The anonymization experiments show that what a model gets *wrong*—evidence of memorization, deviation of stereotypical expectation—can be more productive than what it gets right. More fundamentally, optimizing for a task and interrogating the contingency of the measurement are complementary: the former establishes what the apparatus can do, the latter reveals what the apparatus is made of—and neither can be separated from the situated practice that produces it. This requires that agential cuts be conscious interpretive and ethical commitments.

In this light, cultural analytics is best understood as an interdisciplinary experiment that treats computational work as a material-discursive practice: it ceaselessly reflects on—and attempts to redefine—its positionality between positivist tradition and the negative movement of theory, while interrogating the reciprocal relations among humans, data, and information that undergird it, in order to shed new light on the worlds we inhabit. To measure culture with a language model is to turn the instrument on itself—and learning to do so deliberately, rigorously, and productively completes the big picture.

Acknowledgments

I thank my thesis advisor, Prof. David Bamman, for years of guidance on the work synthesized here, and my collaborators on the projects underlying this paper, named in the original publications, for the partnership that made each contribution possible. I am also grateful to the Stanford Literary Lab for sharing the Chadwyck-Healey English Drama source data used in §6.2, and to Prof. Mark Algee-Hewitt and Emil Wang for helpful discussions. I thank the anonymous reviewer for their feedback.

Most of all: the students of my Cultural Analytics class at UC Berkeley in Fall 2025, subtitled "Machine Learning and Measurements of Culture."[2] The ideas represented here emerged, died, transformed over the semester that we spent together. Teaching and learning with you was among the most generative experiences of my time at Berkeley. You complete the big picture.

References

David Bamman, Kent K Chang, Li Lucy, and Naitian Zhou. 2024. On classification with large language models in cultural analytics. In *Proceedings of the Computational Humanities Research Conference 2024*, Aarhus, Denmark.

Karen Barad. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, Durham, NC.

Monika Bednarek. 2023. *Language and Characterisation in Television Series: A Corpus-informed Approach to the Construction of Social Identity in the Media*. John Benjamins Publishing Company.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv [cs.CL]*.

Katherine Bode. 2020. Why You Can't Model Away Bias. *Modern Language Quarterly*, 80(3).

John Seely Brown and Paul Duguid. 2000. *The Social Life of Information*. Harvard Business Review Press, Boston, MA.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, NY.

Ernst Cassirer. 2014. The concept of symbolic form in the construction of the human sciences. In *The Warburg Years (1919–1933): Essays on Language, Art, Myth, and Technology*. Yale University Press.

[2] `https://ca.kentkc.org`

Kent K. Chang, Danica Chen, and David Bamman. 2023a. Dramatic conversation disentanglement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4020–4046, Toronto, Canada. Association for Computational Linguistics.

Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023b. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Kent K. Chang, Mackenzie Hanh Cramer, Anna Ho, Ti Ti Nguyen, Yilin Yuan, and David Bamman. 2026. Multimodal conversation structure understanding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7437–7458, Rabat, Morocco. Association for Computational Linguistics.

Kent K. Chang and Simon DeDeo. 2020. Divergence and the complexity of difference in text and culture. *Journal of Cultural Analytics*, 4(11):1–36.

Kent K. Chang, Anna Ho, and David Bamman. 2024. Subversive characters and stereotyping readers: Characterizing queer relationalities with dialogue-based relation extraction. In *Proceedings of the Computational Humanities Research Conference 2024*, Aarhus, Denmark.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Herbert H. Clark and Thomas B. Carlson. 1982. Hearers and speech acts. *Language*, 58(2).

John M Clum. 1999. *Something for the Boys*. St. Martin's Press, New York.

Jonathan Culpeper. 2001. *Language and Characterisation: People in Plays and Other Texts*. Longman.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

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.

James E. Dobson. 2025. Beyond computational formalism or, architecture matters. *Journal of Cultural Analytics*, 10(3).

Micha Elsner and Eugene Charniak. 2008. You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio. Association for Computational Linguistics.

Susan Ervin-Tripp. 1964. An analysis of the interaction of language, topic, and listener. *American anthropologist*, 66(6_PART2):86–102.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv [cs.CL]*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv [cs.CL]*.

Erving Goffman. 1963. *Behavior in Public Places*. The Free Press, New York.

Erving Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press, Philadelphia, PA.

Charles Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

David M Halperin. 2002. *How to Do the History of Homosexuality*. Univ. of Chicago Press, Chicago.

Andrew Halterman and Katherine A Keith. 2025. What is a protest anyway? codebook conceptualization is still a first-order concern in LLM-era classification. *arXiv [cs.CL]*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

David L. Hirst. 1979. *Comedy of Manners*. Methuen, London.

hooks, bell. 1992. *Black Looks: Race and Representation*. South End Press.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to Disentangle Interleaved Conversational Threads with a Siamese Hierarchical Network and Similarity Ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana. Association for Computational Linguistics.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5).

Jonathan K Kummerfeld, Sai R Gouravajhala, Joseph J Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A Large-Scale Corpus for Conversation Disentanglement.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adrian Pastor López-Monroy, Fabio A. González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio. 2018. Early text classification using multi-resolution concept representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1216–1225, New Orleans, Louisiana. Association for Computational Linguistics.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450–1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36.

Gerald Mast. 1975. *The Comic Mind: Comedy and the Movies*. Bobbs-Merrill, Indianapolis.

Robert McKee. 2016. *Dialogue: The art of verbal action for page, stage, and screen*. Hachette UK.

Tess McNulty and Laura Alice Chapot. 2025. Computation and form, reconsidered. *Journal of Cultural Analytics*, 10(3).

Franco Moretti. 2013. Operationalizing: or, the function of measurement in modern literary theory. *New Left Review*, 84:103–119.

Sik H Ng and James J Bradac. 1993. *Power in Language: Verbal Communication and Social Influence*. SAGE Publications.

Andrew Piper. 2016. There will be numbers. *Journal of cultural analytics*.

Andrew Piper. 2017. Think small: On literary modeling. *PMLA*, 132(3):651–658.

Kay Richardson. 2010. *Television Dramatic Dialogue: A Sociolinguistic Study*. Oxford University Press.

Marjorie Rosen. 1973. *Popcorn Venus; Women, Movies and the American Dream*. Coward, McCann and Geoghegan.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334. Association for Computational Linguistics.

Eve Kosofsky Sedgwick. 2003. *Touching Feeling*. Duke University Press.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago, IL.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*.

Kaitlyn Zhou, Haishan Gao, Sarah Li Chen, Dan Edelstein, Dan Jurafsky, and Chen Shani. 2025a. Rethinking word similarity: Semantic similarity through classification confusion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5803–5817, Albuquerque, New Mexico. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025b. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2021. Findings on Conversation Disentanglement. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 1–11, Online. Australasian Language Technology Association.

A Restoration Drama Corpus

The corpus underlying §6.2 draws on the Chadwyck-Healey English Drama collection, hand-transcribed by the original publisher with a bespoke XML schema and shared by the Stanford Literary Lab. The raw transcripts are processed through a two-pass agentic pipeline. In the first pass, Gemini reads each source file with an annotated view of its page-break tags and produces a structured manifest via a Pydantic schema: the manifest segments the file into plays by title, author, and page range, and recovers a cast map linking canonical character names to abbreviated speech tags. In the second pass, a LangGraph state machine walks each play page-by-page, maintaining the active play, the current speaker, and a short context window of previous lines; for each page, the model classifies lines as `SPEECH`, `STAGE_DIRECTION`, `ACT_HEADER`, `SCENE_HEADER`, `PROLOGUE`, `EPILOGUE`, or `PARATEXT`, normalizes speaker names against the cast map, and detects transitions between plays in multi-play source files.

The pipeline ran over 166 source files, producing 598 play-level segments. From these, 113 TSVs were retained for downstream analysis, 111 of which were readable, and 109 yielded at least one valid character episode under our preprocessing rules. The author first annotate a randomly sampled five plays, which informed the design of the pipeline described above, and then use Gemini 2.5 Flash to generate the character-episode archetype labels upstream. Cohen's κ between the Gemini and human labels was 0.71; broader human annotation remains future work.

An *episode* is the dialogue produced by one character in one play, kept in dramatic order, retained only if it contains ≥ 3 lines and has a majority archetype label among the four classes. Across 109 plays we obtain 1,283 episodes (median 49 lines per episode, maximum 1,110). Data is split by play, not by character, yielding 82 train, 11 development (for hyperparameter selection), 16 test plays, and 972, 116, 195 episodes, respectively.

The continued-pretraining data is constructed by concatenating each play's speaker-attributed lines in dramatic order, totaling 174,016 lines across 111 play-level documents. Continued pre-training uses causal next-token modelling on `google/gemma-3-1b-pt` for one epoch with block size 1024, learning rate 2×10^{-4}, per-device batch size 1, and gradient accumulation 16, without LoRA or 4-bit quantization. Because this stage uses unlabeled dialogue from the corpus as a whole, including held-out plays, it should be understood as unsupervised domain adaptation.