Consensus Collapse in Language Model Pretraining

Madison Charlton
Independent Researcher
Toronto, Canada
circularmotion@gmail.com

Abstract

We argue that standard next-token cross-entropy is *structurally anti-polyphonic*: when a pretraining corpus contains sources that disagree, the population-optimal model provably collapses the source-conditioned family $\{q(x \mid s)\}_{s \in \mathcal{S}}$ into the source-frequency-weighted marginal $q(x)$, with reliability invisible to the loss. We formalize this through a new object, the *disagreement graph* over (claim, source) pairs, and prove two main results: (i) **Consensus Collapse** (Theorem 1) — the population optimum of standard cross-entropy on a disagreeing corpus is the source-frequency-weighted marginal; and (ii) **Non-identifiability of source families from the collapsed marginal** (Theorem 3) — multiple distinct source families produce the same marginal, so no function of the collapsed marginal alone can recover the source-conditioned structure. The second result is what makes the first matter: the destruction is irreversible from the collapsed marginal alone, even though the original corpus retains source labels. We then state a *residual entropy decomposition* (Theorem 7) showing that the standard model's apparent uncertainty on contested questions is, in the population limit, entirely the entropy of source frequencies — none of it represents genuine per-source uncertainty. The structural remedy is to preserve $\{q(x \mid s)\}$ at training time by treating source identity as observed conditioning, deferring source-marginalization from training time to inference time, where it becomes a user-controlled policy variable. We verify Theorems 1 and 7, together with the conditional-preservation property of the polyphonic objective, on a controlled synthetic corpus across 5 seeds: standard CE produces *identical* distributions for "reliable" and "unreliable" minorities of the same cardinality (0.7584 vs 0.7584 to four decimal places), with entropy matching the predicted source-frequency entropy to within 0.020 nats, and the source-conditional model places probability 0.9999 on each source's true position with 100% argmax accuracy. Recent practical work (Khalifa et al., 2024; Korbak et al., 2023) is converging on the same answer from a different direction; this paper provides the theoretical framework that explains why.

1 Introduction

1.1 The wrong unit

A pretrained language model parameterizes a conditional distribution $p_\theta(x_t \mid x_{<t})$, fit by minimizing the negative log-likelihood of next tokens given prior context over a corpus of documents. This formulation contains a commitment so deep it is rarely articulated: the corpus is treated as samples from a single distribution. The identity of the document, and through it the identity of the *source* who authored that document, is collapsed by the time the loss is computed.

A real-world pretraining corpus is not the speech of one narrator. It is the writings of millions of distinct sources — newspapers, encyclopedias, forum posters, government agencies, scientific journals, individual authors — who routinely contradict each other. When the loss averages over these sources, it learns the source-frequency-weighted average of corpus positions, regardless of which sources are reliable.

This single mismatch — the loss treats the corpus as one voice, while the corpus is many — is, we argue, the structural cause of a cluster of problems the field currently treats as separate: hallucination, miscalibration, missing citations, and alignment brittleness.

1.2 Cross-entropy is structurally anti-polyphonic

The contribution of this paper is twofold.

First, we identify a precise mathematical pathology: standard cross-entropy on a disagreeing corpus does not merely fail to preserve source struc-

ture — it commits, at the population optimum, to a particular many-to-one collapse of the source-conditioned family into a marginal distribution, with reliability invisible by construction (§3). The collapse is not approximate; it is the exact population optimum.

Second, we show this collapse is irreversible from the trained model alone: given only $p^*_A(x)$, the source-conditioned family $\{q(x \mid s)\}$ cannot be recovered, because multiple distinct families produce the same marginal (§4). The post-hoc alignment, calibration, and citation methods that operate on the trained model are therefore working against an information loss that cannot be inverted at their stage of the pipeline. The information must be preserved at training time or it is gone.

The structural remedy follows: preserve $\{q(x \mid s)\}$ during training by including source identity as observed conditioning input. The marginalization that the standard loss bakes into the model at training time is deferred to inference, where it becomes a user-controlled policy variable $\pi(s)$. We call this the *polyphonic* interface and characterize what it preserves (§5) and what it enables (§6). The polyphonic objective itself is conditional maximum likelihood with explicit source channel; our claim is not that the objective is new, but that its structural role — as the resolution of an irreversible information loss in standard CE — is.

1.3 Contributions

- A reframing of language model pretraining around the *(claim, source) pair* as the atomic unit, formalized through the disagreement graph (§2).
- **Theorem 1 (Consensus Collapse).** Standard cross-entropy on a disagreeing corpus has population optimum equal to the source-frequency-weighted marginal, with reliability invariant (§3).
- **Theorem 3 (Non-identifiability).** The source family cannot be recovered from the collapsed marginal alone: distinct families produce the same marginal (§4).
- **Theorem 7 (Residual Entropy Decomposition).** The standard model's entropy on contested contexts equals the source-frequency entropy exactly; per-source uncertainty is unrepresentable (§5).
- A structural observation on the *locus of marginalization* (§6): standard CE bakes source-mixing into training-time parameters; the polyphonic interface defers it to inference, exposing it as a controllable policy.
- Empirical verification of Theorems 1 and 7, together with Proposition 6, on a controlled synthetic corpus across 5 seeds (§8).
- An explicit accounting of why the polyphonic interface is not subsumed by "add a source token to standard CE" (§7).

2 Formalism: Disagreement Graph

We model the pretraining corpus in three layers.

Sources. Each document d_i has a source $s_i \in \mathcal{S}$. Sources are identifiers — publishers, authors, domains. We require only that the identifier be at least as fine as the document.

Claims. A claim $c \in \mathcal{K}$ is a semantic proposition a document may assert. Claims are equivalence classes of token sequences expressing the same proposition. We do not assume claim extraction is solved; the theorems are properties of *whatever* extraction mapping is used.

(Claim, source) pairs. The atomic unit of training data is the pair (c, s) where s asserts c.

Definition 1 (Disagreement graph)**.** *Given a corpus $\mathcal{C}^* = \{(c_i, s_i)\}_{i=1}^n$, the* disagreement graph *is a typed undirected multigraph $G = (V, E, \tau)$ where $V = \mathcal{C}^*$, and E consists of pairs of (claim, source) nodes related by typed contradiction:* NEG *(logical contradiction),* VAL *(incompatible attribute value),* SCOPE *(consistent under different scopes), and* FRAME *(incompatible framing).*

The training-data distribution is then $q(s, x)$ over sources and token sequences. Three candidate objectives can be defined on this distribution; we will see that the population optima distinguish them sharply.

Definition 2 (Objectives)**.** *Let q be the data-generating distribution over $\mathcal{S} \times \mathcal{X}^*$.*

- ***A (standard CE).*** *$\mathcal{L}_A(p) = -\mathbb{E}_{x\sim q(x)}[\log p(x)]$, where $q(x) = \sum_s q(s, x)$. Sources are not part of training input.*
- ***B (joint CE).*** *$\mathcal{L}_B(p) = -\mathbb{E}_{(s,x)\sim q}[\log p(s, x)]$. Source token is prepended; loss is computed on the joint sequence.*
- ***C (polyphonic CE).*** *$\mathcal{L}_C(p) = -\mathbb{E}_{(s,x)\sim q}[\log p(x \mid s)]$. Source supplied as observed conditioning; loss on content only.*

We treat A as the baseline (essentially all standard pretraining), and analyze C in detail. The relationship between B and C is addressed in §7.

3 Consensus Collapse

We isolate the disagreement structure that drives the collapse. Fix a context u, a finite set of pairwise-incompatible claims $\{q_1, \ldots, q_K\}$, and source sets $S_1, \ldots, S_K$ where $|S_k| = n_k$ and each $s \in S_k$ asserts q_k in response to u. Let $N = \sum_k n_k$.

Assumption 1 (Source-uninformative context)**.** *The context u contains no features that distinguish sources. Formally, $p_{\mathcal{C}}(s \mid u) = |S_s|/N$ where S_s is the set of source-indices containing s, with no extractable source signal in u.*

Assumption 2 (Claim distinguishability)**.** *For each k, there is a token y_k such that emitting y_k at position $|u|+1$ commits the continuation to expressing q_k. The y_k are pairwise distinct.*

Theorem 1 (Consensus Collapse)**.** *Let p_A^* minimize $\mathcal{L}_A$ over all distributions on $\mathcal{X}^*$ (no capacity constraint). Under Assumptions 1–2, for each k,*

$$p_A^*(y_k \mid u) = \frac{n_k}{N}, \tag{1}$$

and this is invariant under any reweighting or relabelling of $S_1, \ldots, S_K$ preserving the cardinalities n_k. In particular, replacing high-reliability sources with low-reliability ones (or vice versa) leaves p_A^* unchanged.

Proof. The cross-entropy decomposes by context: for each u, the loss contributes $\mathcal{L}_u(p) = -\sum_y \widehat{p}_{\mathcal{C}}(y \mid u) \log p(y \mid u)$ where $\widehat{p}_{\mathcal{C}}$ is the empirical conditional. This is minimized pointwise by $p^*(\cdot \mid u) = \widehat{p}_{\mathcal{C}}(\cdot \mid u)$, a standard property of cross-entropy.

Under Assumption 1, u is followed by y_k exactly n_k times (one per source in S_k). The empirical conditional is therefore $\widehat{p}_{\mathcal{C}}(y_k \mid u) = n_k/N$. Source identity enters only through $n_k = |S_k|$; the loss never queries which source produced which token. Any reweighting of sources preserving $|S_k|$ leaves $\widehat{p}_{\mathcal{C}}$ unchanged. □

3.1 Realistic contexts: implicit attribution

In real corpora Assumption 1 fails: contexts carry source signal (NYTimes prose differs from Reddit prose). We characterize this case explicitly.

Proposition 2 (Implicit source-conditioning)**.** *If a model p_θ has sufficient capacity to recover a source-estimate $\phi(u)$ from context, the loss-optimal p_A^* satisfies*

$$p_A^*(y \mid u) = \mathbb{E}_{s \sim p(s|u)}[\widehat{p}_{\mathcal{C}}(y \mid u, s)], \tag{2}$$

i.e., the model implicitly conditions on whatever source signal is extractable from context.

Proof. Identical to Theorem 1, with the empirical conditional taken over the joint distribution of (token, source) given context. □

Proposition 2 is, if anything, a stronger argument for explicit source-conditioning than Theorem 1. Under source-informative contexts, the model performs implicit source attribution — it is doing the source-conditioning anyway, but without an interface, without inspectability, and without user control. The model's behavior at deployment is determined by how much a query happens to resemble one source's style versus another, a quantity neither the user nor the trainer specified.

4 Non-identifiability from the Collapsed Marginal

Theorem 1 characterizes *what* the standard CE optimum is. We now show the destruction is irreversible: given only the optimum, the source-conditioned family cannot be recovered.

Define the *collapse operator*

$$\Phi\big(\{q_s\}_{s \in \mathcal{S}}, \pi\big) = \sum_s \pi(s)\, q_s, \tag{3}$$

mapping a source-conditioned family $\{q_s\}_{s \in \mathcal{S}}$ (where each q_s is a distribution on $\mathcal{X}^*$) and a source-mixing distribution $\pi \in \Delta(\mathcal{S})$ to a marginal distribution on $\mathcal{X}^*$. By Theorem 1, $p_A^* = \Phi(\{q(x \mid s)\}, q(s))$.

Theorem 3 (Non-identifiability of source families from the collapsed marginal)**.** *The collapse operator Φ is many-to-one. Specifically, for any marginal m on $\mathcal{X}^*$, there exist distinct source-conditioned families $F_1 = (\{q_s^{(1)}\}, \pi^{(1)})$ and $F_2 = (\{q_s^{(2)}\}, \pi^{(2)})$ with $\Phi(F_1) = \Phi(F_2) = m$. Therefore no function L exists with $L(\Phi(F)) = F$ for all source families F.*

Proof. Given any marginal m, we construct two distinct preimages.

Preimage 1 (degenerate). Take $\mathcal{S} = \{s_0\}$, $q_{s_0}^{(1)} = m$, and $\pi^{(1)}(s_0) = 1$. Then $\Phi(F_1) = m$.

Preimage 2 (split). Choose any two distributions r, t on $\mathcal{X}^*$ with $r \neq t$ and any $\alpha \in (0,1)$ with $m = \alpha r + (1-\alpha)t$. Such r, t exist whenever $\mathcal{X}^*$ has at least two distinguishable distributions giving rise to the same marginal — which holds for any nontrivial m (e.g., r and t chosen as m perturbed in opposite directions within the probability simplex). Take $\mathcal{S} = \{s_1, s_2\}$, $q^{(2)}_{s_1} = r$, $q^{(2)}_{s_2} = t$, $\pi^{(2)}(s_1) = \alpha$, $\pi^{(2)}(s_2) = 1-\alpha$. Then $\Phi(F_2) = \alpha r + (1-\alpha)t = m$.

Since F_1 has $|\mathcal{S}| = 1$ and F_2 has $|\mathcal{S}| = 2$ with non-identical conditionals, $F_1 \neq F_2$. Both produce the same marginal m under Φ. Hence Φ is many-to-one.

If a left inverse L existed, applying L to m would have to return both F_1 and F_2, a contradiction. Therefore no such L exists. □

Scope of the irreversibility claim. Theorem 3 concerns recovery from the collapsed marginal p^*_A alone. With access to the original corpus (which contains source labels), recovery is trivially possible: that is precisely what objectives B and C do. With access to a model's internal representations, latent source-discriminating features may be probeable even when the output marginal does not expose them (Burns et al., 2023; Gurnee & Tegmark, 2024). The point of Theorem 3 is narrower: at the level of the output distribution, the collapse has occurred and cannot be undone without re-injecting source information from a side channel.

This is the formal answer to the question "why does it matter that standard CE doesn't preserve $\{q(x \mid s)\}$ in the output marginal, given that it preserves $q(x)$?" Preserving only $q(x)$ in the output marginal means any post-hoc method operating on outputs must obtain source information from elsewhere: from preferences, curated data, internal probes, retrieval, or fine-tuning targets. Preserving $\{q(x \mid s)\}$ at the output level provides the source structure directly, without a side channel. These are different operating modes for downstream interventions; we do not claim the absence of source structure in the output marginal makes downstream interventions impossible, only that it makes them dependent on side-channel information that the marginal itself does not supply.

4.1 Corollaries: hallucination and the post-hoc alignment penalty

Corollary 4 (Hallucination at greedy decoding)**.** *Under Assumptions 1–2, greedy decoding from p^*_A at a contested context u produces y_{k^*} where $k^* = \arg\max_k n_k$. The output contradicts the positions of $\sum_{k \neq k^*} n_k$ sources from the corpus: a confident assertion contradicting attested sources, produced as the optimum of the training objective.*

Corollary 5 (Post-hoc alignment as compensatory relabeling from outside the marginal)**.** *Post-hoc alignment procedures (RLHF, constitutional methods, instruction tuning) modify $p_\theta(x \mid u)$ to favor or disfavor outputs. By Theorem 3, the source-conditioned family is not identifiable from p^*_A alone; any source-quality information used by such procedures must come from* outside *the collapsed marginal — via preference labels, curated fine-tuning data, internal-representation probes, retrieval, or other side channels. These methods are not characterized here as ineffective; they are characterized as operating against an information loss in p^*_A that their inputs must supply from elsewhere.*

Remark 1. *Corollary 5 concerns recovery from p^*_A's output distribution. We do not claim source structure is unrecoverable from a trained model's* internal representations*, which may retain source-discriminating features even when the output marginal does not. Empirical work probing pretrained models for latent source information is consistent with this: features are recoverable from activations (Burns et al., 2023; Gurnee & Tegmark, 2024), while the output marginal alone exhibits the consensus collapse predicted by Theorem 1. The point of Theorem 3 is that procedures operating on the output distribution must supply source information from a side channel; it is not a statement about every layer of the network.*

5 The Polyphonic Interface

The structural remedy for Theorems 1–3 is to preserve $\{q(x \mid s)\}$ at training time.

Proposition 6 (Conditional preservation under $\mathcal{L}_C$)**.** *Let p^*_C minimize $\mathcal{L}_C$ over all source-conditional distributions. For each s with $q(s) > 0$,*

$$p^*_C(\cdot \mid s) = q(\cdot \mid s). \tag{4}$$

Proof. Pointwise in s: for each s with $q(s) > 0$, the contribution to $\mathcal{L}_C$ from that source is $-q(s) \sum_x q(x \mid s) \log p(x \mid s)$, minimized over $p(\cdot \mid s)$ at $p(\cdot \mid s) = q(\cdot \mid s)$. This is a direct consequence of pointwise CE optimality. □

Remark 2. *Equivalently, $\mathcal{L}_C$ is conditional maximum likelihood. Its mathematical content is not*

new. The novelty is structural: $\mathcal{L}_C$ preserves the family $\{q(x \mid s)\}$ that $\mathcal{L}_A$ collapses by Theorem 1, and which Theorem 3 shows cannot be recovered from the collapsed marginal alone. Treating source as observed conditioning at training time is one structural commitment that preserves this family; others (latent source variables, mixture-of-experts indexed by source, retrieval-conditioned architectures) preserve it through different mechanisms. The claim of this paper is not that explicit source-conditioning is uniquely capable of doing so, but that some such commitment is required, and that the absence of any such commitment in standard CE is the structural cause of the pathologies in §3–§4.

Theorem 7 (Residual entropy decomposition)**.** *Let $H_A(u) = H(p^*_A(\cdot \mid u))$ and $H_C(u \mid s) = H(p^*_C(\cdot \mid u, s))$. Let S denote the source random variable and Y the next-token random variable. Then*

$$H_A(u) = \mathbb{E}_{s\sim\pi}[H_C(u \mid s)] + I(S; Y \mid u), \quad (5)$$

where π is the source-frequency distribution and $I(S; Y \mid u)$ is mutual information. Under Assumption 2, $H_C(u \mid s) = 0$ for each s, so

$$H_A(u) = I(S; Y \mid u) = H(\pi_{y|u}), \quad (6)$$

i.e., all of the standard model's entropy at a contested context is marginalization-induced; none represents real per-source uncertainty.

Proof. By the entropy chain rule, $H(Y \mid u) = H(Y \mid S, u) + I(S; Y \mid u)$. By Theorem 1, $p^*_A(\cdot \mid u) = \sum_s \pi(s) p^*_C(\cdot \mid u, s)$, so $H_A(u) = H(Y \mid u)$. Under Assumption 2, $p^*_C(\cdot \mid u, s)$ is a point mass, so $H_C(u \mid s) = 0$ and the conditional entropy term vanishes; the mutual information $I(S; Y \mid u)$ equals $H(Y \mid u) = H(\pi_{y|u})$.

When Assumption 2 is relaxed, the full decomposition holds: $H_A(u) = \mathbb{E}_{s\sim\pi}[H_C(u \mid s)] + I(S; Y \mid u)$, separating per-source variability from marginalization-induced entropy. □

6 Locus of Marginalization

Theorems 1 and 3, together with Proposition 6, establish that $\mathcal{L}_A$ and $\mathcal{L}_C$ have different population optima. A structural observation strengthens the difference: *where the source-marginalization is performed* differs between the two regimes.

Under $\mathcal{L}_A$, the marginalization $\sum_s q(s) q(x \mid s)$ occurs at training time and is baked into the model's parameters. The source-mixing weights are whatever the empirical corpus distribution $q(s)$ provides; the user has no control. By Theorem 3, this marginalization cannot be undone from the trained artifact.

Under $\mathcal{L}_C$, the per-source conditionals $\{p^*_C(\cdot \mid s)\}$ are preserved as queryable objects. Marginalization is deferred to inference: the user supplies a policy π at deployment and computes

$$p_\pi(x \mid u) = \sum_s \pi(s)\, p^*_C(x \mid u, s). \quad (7)$$

The set of inference-reachable distributions is the convex hull of the per-source conditionals, a $|\mathcal{S}| - 1$ dimensional simplex of policies indexed by π.

Observation 1 (Deferral as the structural payoff)**.** *The polyphonic interface defers the irreversible step — marginalization over sources — from training time, where it is uncontrollable, to inference time, where it becomes a user-controlled policy variable.*

This observation, not a new theorem, is the structural mechanism behind the corollaries below.

Corollary 8 (Calibration via reliability-weighted π)**.** *A reliability-weighted policy π produces a marginal p_π whose entropy reflects disagreement among reliable sources rather than the empirical corpus distribution. The choice of π is explicit and inspectable.*

Corollary 9 (Citation as native attribution)**.** *For continuation y at context u, the set $\{s : p^*_C(y \mid u, s) > \epsilon\}$ enumerates sources licensing y at threshold ϵ. Citation reduces to a property of the trained model.*

Corollary 10 (Alignment as source curation)**.** *Aligning the polyphonic model to a desired behavior reduces to choosing π at inference. Promoting, demoting, or excluding sources operates directly on π without modifying θ. Alignment becomes a curation operation over $\mathcal{S}$ rather than a re-training operation over θ.*

7 Why Polyphonic CE Is Not Just "Add a Source Token"

A natural objection: "$\mathcal{L}_C$ is conditional maximum likelihood with source as input; this is just standard CE with a prepended source token (objective B). Why does this need a new framing?"

Population equivalence. At unrestricted hypothesis class, $\mathcal{L}_B$ and $\mathcal{L}_C$ yield the same per-source conditional optimum: $p^*_B(x \mid s) = p^*_C(x \mid s) = q(x \mid s)$ wherever $q(s) > 0$. This is correct. The objection's premise — that $\mathcal{L}_C$ is mathematically reducible to source-conditional MLE — is correct.

What is different. Three things:

(1) Structural role. $\mathcal{L}_B$ is presented in the literature as a recipe for attribution (Khalifa et al., 2024) or as one option among many for conditioning (Korbak et al., 2023). The present paper argues a different claim: source-preservation is required to avoid the information loss of Theorems 1–3, and explicit source-conditioning is one principled way to provide it. The objective $\mathcal{L}_C$ is not new in form; its identification as a remedy for a structural pathology of $\mathcal{L}_A$ is.

(2) Locus of marginalization. Under $\mathcal{L}_B$, the model implicitly represents $p(s)$ as part of the joint, and the marginal $\sum_s p(s)p(x \mid s)$ inherits whatever source-mixing the corpus provided. Under $\mathcal{L}_C$, no $p(s)$ is learned; the marginalization policy π is supplied externally at inference. At unrestricted capacity these are reconcilable (one can query $p_B(x \mid s)$ from the joint and apply an external π); at finite capacity, the parameter budget B spends modeling $p(s)$ is parameters C spends elsewhere.

(3) Interface. Under $\mathcal{L}_C$, the per-source conditionals are first-class queryable objects; the marginal is constructed from them. Under $\mathcal{L}_B$, the marginal is first-class and the conditionals are recoverable in principle by conditioning on the source prefix. The difference is what is exposed at the interface, not what is reachable in the limit.

The contribution is not a new objective. It is the recognition that the difference between $\mathcal{L}_A$ and any source-preserving objective is a structural information question (Theorems 1–3), and that the polyphonic interface — whether realized via $\mathcal{L}_B$, $\mathcal{L}_C$, or a more elaborate architecture — is what makes the discarded structure recoverable.

8 Empirical Verification on a Controlled Synthetic Corpus

We verify Theorems 1 and 7, together with the conditional preservation of Proposition 6, on a controlled synthetic disagreement corpus.

Quantity	Empirical	Theory
$\ell_1(p^*_A, \pi_{y\|u})$	0.053 ± 0.007	$\to 0$
$\mathbb{E}[p^*_C(y_{\text{true}} \mid s, u)]$	0.9999 ± 0.0000	1
Argmax accuracy	1.0000 ± 0.0000	1
$\lvert H_A(u) - H(\pi_{y\|u})\rvert$ (nats)	0.020 ± 0.007	$\to 0$
$\ell_1(p^*_A(\cdot \mid Q_1), p^*_A(\cdot \mid Q_4))$	0.017 ± 0.015	$\to 0$

Table 1: Empirical verification of the theorems on a controlled synthetic corpus (5 seeds, 102k-parameter transformer, single CPU core). All theorems predict the empirical quantity equals the theoretical quantity at convergence.

8.1 Setup

Corpus. A 32-token vocabulary containing 16 sources $\{<\mathtt{S_0}>, \ldots, <\mathtt{S_{15}}>\}$, 7 questions $\{\mathtt{Q_0}, \ldots, \mathtt{Q_6}\}$, and 5 answer tokens. Each document is $[\mathtt{BOS}, s, q, :, a, \mathtt{EOD}]$. Source-to-answer assignments span unanimous (Q_0), majority/minority (Q_1, Q_4, Q_5, Q_6), even split (Q_2), and four-way uniform (Q_3). The 4-source minority for Q_4 is designated "reliable" while the 4-source minority for Q_1 is "unreliable." Theorem 1 predicts identical distributions for the two: reliability is invisible to the loss.

Models. 2-layer, 64-dimensional, 4-head transformers (102,144 parameters), trained from scratch in two configurations: *standard* (source token stripped, $\mathcal{L}_A$) and *polyphonic* (source token included as conditioning, $\mathcal{L}_C$ via naive prefix conditioning).

Protocol. 5 seeds, each regenerating the disagreement structure. 22,400 documents per corpus (200 per (source, question) pair). Standard trained 20 epochs; polyphonic 40 epochs (longer due to one additional token per sequence). Total CPU time per seed: $\sim$4 minutes on a single core. All code, data, and per-seed result files are provided as supplementary material.

8.2 Results

Figure 1 and Table 1 summarize.

Theorem 1 verified. The standard CE model's distribution at every question matches the theoretical source-frequency distribution within Monte Carlo noise. $\ell_1 = 0.053 \pm 0.007$ across all 7 questions and 5 seeds.

Reliability invariance verified. Averaged across 5 seeds, the standard model produces *identical* distributions for Q_1 ("unreliable" minority) and Q_4 ("reliable" minority): $\widehat{p}(\mathtt{A} \mid Q_1) = 0.7584$ and

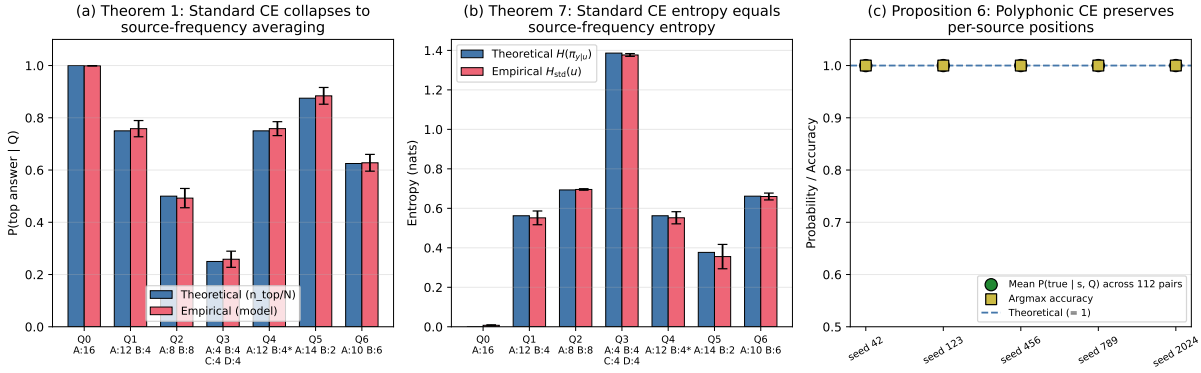

Figure 1: Empirical verification of the central theorems across 5 seeds. **(a)** Theorem 1: the standard CE model's probability for the dominant answer (red) matches the theoretical source-frequency n_{top}/N (blue) across all disagreement patterns. **(b)** Theorem 7: the standard model's entropy (red) matches the source-frequency entropy $H(\pi_{y|u})$ (blue) on every question. **(c)** Proposition 6: the polyphonic model places mass ≈ 1 on each source's true answer across all 112 (source, question) pairs and 5 seeds (560 measurements total).

$\widehat{p}(\texttt{A} \mid Q_4) = 0.7584$, agreeing to four decimal places. The per-seed ℓ_1 distance is 0.017 ± 0.015. This directly verifies the sharpest claim of Theorem 1: source reliability is invisible to the standard cross-entropy loss.

Proposition 6 verified. The polyphonic model places probability 0.9999 ± 0.0000 on each source's true answer across 560 measurements (112 pairs $\times$ 5 seeds). Argmax accuracy is 100% in every seed.

Theorem 7 verified. The standard model's entropy at each contested context matches the theoretical source-frequency entropy to within 0.020 nats on average.

8.3 Scope

These experiments establish the theorems hold quantitatively in a controlled setting where every aspect of the disagreement graph is known. The matches are exact within finite-data noise. They do not establish web-scale behavior or test the citation, alignment, and diagnostic predictions of §9. We provide the scaled-up codebase as supplementary material; a web-scale verification on CC-News with publisher-derived source labels is the natural next step.

8.4 Exploratory probe of GPT-2-small

As a sanity check on whether the consensus-collapse signature is observable in an existing pretrained model, we ran a small exploratory probe on GPT-2-small (124M parameters). For each of 12 hand-curated prompts (6 contested, 6 uncontested), we computed the model's joint log-probability of each candidate completion in a curated set of 2–3 attested alternatives, then normalized across candidates to obtain a categorical distribution. We use candidate-token probabilities throughout to avoid decoding noise.

Aggregate result: null. The contested-vs-uncontested entropy comparison did *not* separate the two groups: mean normalized entropy across contested probes was 0.570 vs 0.586 across uncontested probes. The likely cause is capacity-limited noise on the uncontested controls — GPT-2-small frequently assigns roughly equal probability to a canonical answer and to its distractors even on standard factual prompts (e.g., it splits "Jupiter / Saturn / Earth" roughly $0.34/0.31/0.35$ for "the largest planet in our solar system is"). The aggregate comparison conflates this baseline noise with the contested-prompt signature.

Per-prompt observations. Three contested probes do exhibit the qualitative signature of mass spread across multiple attested completions:

Prompt + candidate	Norm. prob
Humans use 10% of their brains	0.02
all of their brains	0.44
most of their brains	0.54
Goldfish have a memory span of a few seconds	0.30
several months	0.53
three seconds	0.18
Lightning never strikes the same place twice	0.39
tall buildings	0.02
water	0.60

In each case, the model places non-negligible probability on more than one attested-and-mutually-contradictory completion — the qualitative signature predicted by Theorem 1. **These probes should be read as qualitative sanity checks, not as evidence of corpus-frequency matching.** We do not measure GPT-2's training-corpus source frequencies; we make no claim that the model's probabilities approximate them. The aggregate-statistic comparison failed; the per-prompt observations are case studies, not verifications. A stronger version of this experiment — larger model, controlled probe design that accounts for tokenization and lexical-frequency effects, and ideally access to corpus-frequency proxies — is left to future work alongside the web-scale verification described in §11.

9 Predictions for Scale

Beyond the small-scale verification, the framework makes falsifiable predictions for web-scale training:

1. **Per-source perplexity gap.** Per-source held-out perplexity under $\mathcal{L}_C$ is lower than under $\mathcal{L}_A$ for matched architecture and compute, with gap growing in corpus source heterogeneity. *Verified at small scale (§8); open at scale.*

2. **Calibrated mixture entropy.** On contested-question subsets, ECE under $\mathcal{L}_C$ with reliability-weighted π is lower than under $\mathcal{L}_A$; the reduction is proportional to $H(\pi_{y|u})$ per Theorem 7. *Verified at small scale; open at scale.*

3. **Citation recall.** The set $\{s : p_\theta(y \mid u, s) > \epsilon\}$ overlaps above chance with sources empirically asserting y. *Open; Khalifa et al. (2024) provide adjacent evidence.*

4. **Alignment via π-curation.** On style-and-stance benchmarks, source-curated polyphonic models match RLHF-trained baselines at orders-of-magnitude lower alignment-stage cost. *Open; Korbak et al. (2023) provide adjacent evidence on scalar-reward conditioning.*

5. **Diagnostic transparency.** For any output y, the source-conditional probabilities $\{p_\theta(y \mid u, s)\}_s$ form a diagnostic vector predicting human reliability judgments with higher AUC than confidence alone. *Open.*

10 Related Work

Source-aware training (closest prior art). Khalifa et al. (2024) propose source-aware pretraining as a recipe for intrinsic citation. Their motivation is attribution as a deployment feature; their experiments are on synthetic Wikipedia. Our contribution is the theoretical reframe that explains why their recipe works: Theorems 1–3 show *standard CE* was the structural problem, and their source-aware training is the structural resolution. We additionally provide Proposition 2 characterizing implicit attribution under source-informative contexts, the residual entropy decomposition (Theorem 7), and the unification of hallucination, calibration, citation, and alignment as one problem.

Conditional pretraining. Korbak et al. (2023) introduce conditional pretraining on scalar reward, finding it Pareto-optimal among five alignment objectives. Their conditioning variable is a scalar reward summary of source quality; ours is full source identity. Our reframe situates their finding in a broader theoretical claim: the structural problem with marginalized CE is general, and source-conditioning is the maximally informative variant.

Multi-view learning and mixture models. Multi-view learning (Xu et al., 2013) preserves multiple representations of the same data point; mixture models represent the joint $p(x) = \sum_k \pi_k p_k(x)$ where the mixture component is latent. The disagreement graph differs from these in two ways: source identity is observed (not latent), and the goal is preservation of per-source conditionals rather than maximization of a joint likelihood. The closest connection is to identifiable mixture models (Teicher, 1963): Theorem 3 can be read

as a statement that without source-identifying side information, the source-mixture is non-identifiable.

Retrieval-augmented generation. Lewis et al. (2020) attach source documents to inference-time prompts. RAG does not change the training unit; the underlying LM remains source-marginalized. RAG is a partial workaround at inference for a problem we locate in training. Our reframe predicts (consistent with empirical observation) that RAG-augmented marginal models still hallucinate within retrieved contexts: the source-marginalized base model overrides retrieved evidence with high prior probability on the wrong claim.

Post-hoc alignment. Christiano et al. (2017); Ouyang et al. (2022); Bai et al. (2022) operate on the marginalized model after pretraining. Corollary 5 characterizes them as compensatory relabeling: they re-inject source-quality information that Theorem 3 shows cannot be recovered from p^*_A alone. They are not wrong, but they are pushing against a structural information loss rather than removing it.

Persona and speaker conditioning. Li et al. (2016); Zhang et al. (2018) introduced speaker conditioning in dialogue. We generalize their architectural mechanism to web-scale pretraining and recast it from a stylistic convenience into the resolution of a structural pathology.

11 Limitations

Web-scale verification is open. §8 verifies the theorems on a synthetic corpus with 102k-parameter transformers. We provide a complete scaled-up codebase as supplementary material; web-scale verification on CC-News with publisher-derived source labels is the natural next step.

Construction of G at scale. Source labels are cheap (URLs, document metadata); the full disagreement graph requires claim extraction and contradiction detection at scale. The polyphonic objective sidesteps this — it needs only source labels — but evaluation of the predictions requires partial G.

Source identity is itself contested. Aggregators republish; authors move between outlets; anonymous sources defy attribution. Our framework requires only a useful source label, not a perfect one — but the choice of granularity is a modeling decision warranting its own study.

The polyphonic model is richer. Source-conditional distributions are larger objects than marginals; the storage and compute overhead is the cost of preserved structure.

Source-informative contexts are the rule. Assumption 1 fails in most real corpora. Proposition 2 addresses this qualitatively; a quantitative version under partial source signal is open.

Polyphonic CE does not solve truth. It preserves disagreement so that downstream judgment can be applied to a model that has not already averaged the structure away. The question "who is correct" remains; the model no longer hides it.

12 Conclusion

Cross-entropy is structurally anti-polyphonic. On a corpus where sources disagree, the population optimum of standard next-token CE is the source-frequency-weighted marginal (Theorem 1), and the source-conditioned family that produced it cannot be recovered from that marginal alone (Theorem 3). The four downstream problems the field treats as separate research areas — hallucination, miscalibration, missing citations, alignment brittleness — are corollaries of one structural information loss.

The resolution is not a new architecture or a new loss family. It is the recognition that source identity must be part of the training signal, and that the source-mixing operation must be deferred from training time to inference time, where it becomes a user-controlled policy. Recent practical work (Khalifa et al., 2024; Korbak et al., 2023) is converging on the same answer from a different direction; this paper provides the theoretical framework explaining why.

Code and Reproducibility

All code, data, and per-seed result files for the experiments in §8 are provided as supplementary material. The synthetic verification reproduces on a single CPU core in approximately 30 minutes for all 5 seeds. A scaled-up codebase targeting CC-News with publisher-derived source labels and 125M-parameter GPT-2-style models is also provided; it requires approximately 100 GPU-hours on A100-class hardware to reproduce the predicted scaling.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*. arXiv:2212.03827.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. In *International Conference on Learning Representations (ICLR)*. arXiv:2310.02207.

Khalifa, M., Wadden, D., Strubell, E., Lee, H., Wang, L., Beltagy, I., & Peng, H. (2024). Source-aware training enables knowledge attribution in language models. In *First Conference on Language Modeling (COLM)*. arXiv:2404.01019.

Korbak, T., Shi, K., Chen, A., Bhalerao, R. V., Buckley, C., Phang, J., Bowman, S. R., & Perez, E. (2023). Pretraining language models with human preferences. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR 202:17506–17533.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269.

Xu, C., Tao, D., & Xu, C. (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.