Argument Collapse: LLMs Flatten Long-Form Public Debate

Yekyung Kim* **Yapei Chang*** **Chau Minh Pham** **Mohit Iyyer**
University of Maryland, College Park
{yekyung,yapeic,chau,miyyer}@umd.edu

Abstract

As LLMs are increasingly used to draft public-facing arguments, they may flatten public debate by repeatedly introducing the same polished, plausible arguments. We study *argument collapse*, the tendency of essays generated by different LLMs to converge to a smaller set of main arguments, sub-arguments, and paragraph-level structures. We compare 1,039 human responses from 195 *New York Times* (NYT) debates, 448 human responses from 61 longer-form *Boston Review* (BR) forums, and 23,381 LLM-generated essays. In the NYT corpus, 65.3% of human main arguments are unique within a debate, compared to 3.4% of LLM main arguments. Asking LLMs to generate diverse answers adds variation, but a typical model recovers only about half of the distinct human main arguments, with much of the added variation falling outside the observed human argument space. Collapse also appears in sub-arguments, where among essays with the same main argument, 41.0% of human sub-arguments are unique versus 9.1% from LLM responses. Qualitatively, LLMs often reuse generalized and hedged sub-arguments, while humans prefer more concrete and topic-specific ones. Structure-wise, LLM-generated essays tend to follow a more fixed arc, often opening with a direct claim and moving quickly toward proposals. The same patterns hold in longer BR essays, suggesting that argument collapse extends beyond short-form responses. Finally, human-preference evaluators favor both common arguments and LLM essays, which could reinforce argument collapse.

github.com/mungg/argument_collapse

1 Introduction

LLMs are now common aids in public-facing argumentative writing, including opinion essays and policy memos (Lee et al., 2022; Russell et al., 2026a). Such usage comes with a caveat: model suggestions can alter what claims writers make, how those claims are supported, and how much of the writer's own voice remains present (Padmakumar and He, 2024; Doshi and Hauser, 2024; Abdulhai et al., 2026; Röttger et al., 2026a). Prior work shows that LLMs can produce generative monocultures through narrowing output distributions (Wu et al., 2025; Zhang et al., 2025b; Jiang et al., 2026; Nie et al., 2026) and reducing epistemic diversity in generated claims (Wright et al., 2025). These homogenization measures, however, often do not directly compare human and LLM output distributions under the same task conditions, leaving it unclear what is lost when people write arguments with model assistance (Jain et al., 2025).

In this work, we ask whether LLMs collapse into a narrower range of main arguments, supporting claims, and argumentative structures than human writers produce in response to the same debate questions. We use the term *argument collapse* to describe cross-model convergence: the tendency of different LLMs, built by different frontier industry labs, to return to the same small set of plausible arguments rather than span the broader range of arguments humans make. Such failure has broad implications, as these systems can measurably shift reader beliefs (Jakesch et al., 2023; Fisher et al., 2025), narrow readers' perspectives (Sharma et al., 2024; Peterson, 2025), and recirculate through public discourse and training corpora in ways that amplify dominant arguments at the expense of long-tail reasoning (Shumailov et al., 2024).

We study *argument collapse* in three settings. In the `vanilla` setting, we ask whether LLMs collapse even under a basic setup: when given a contested question, do their responses reflect the range of arguments humans make? Because LLMs are often used for drafting and ideation (Wan et al., 2024), we test whether a `diversified` prompt, which explicitly asks models to generate diverse responses, re-

*These authors contributed equally to this work.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 23847–23893
October 24-29, 2026 ©2026 Association for Computational Linguistics

covers the breadth of human arguments. Finally, in the `position-guided` setting, we provide the models with human responders' main argument, biography, and tone, then ask whether it can come up with the supporting reasons that the writer actually use. We compare human-written responses with essays generated by five frontier LLMs across two corpora: *New York Times Room for Debate* (NYT; ≈352 words) and longer *Boston Review* forum responses (BR; ≈1,150 words). Our analysis focuses on main arguments, supporting reasons, and paragraph-level argumentative structure. Across all settings, we find evidence of *argument collapse*:

Main arguments collapse under vanilla prompting. Models from different providers repeatedly converge on the same main arguments (Figure 1A). In NYT, 65.3% of human main arguments are unique within a debate, compared with only 3.4% of `vanilla` LLM arguments. These arguments are often plausible and human-like, but are much less diverse than those produced by human writers.

Diversity prompting adds variation, but only partly recovers human arguments. `diversified` outputs are more varied than `vanilla` outputs, but they still miss many human-written arguments. A typical LLM recovers only about half of the distinct human main arguments (50–55%), often missing narrower or more situated arguments.

Sub-argument collapse persists even when the main argument is fixed. Among essays sharing the same main argument, only 9.1% of `vanilla` sub-arguments are unique, compared with 41.0% for humans (Figure 1B). `diversified` and `position-guided` outputs improve this only partially.

LLMs converge on different kinds of sub-arguments. LLMs more often repeat generalized and hedged supporting arguments, whereas humans more often use concrete and topic-specific ones.

Collapse also appears in essay structure. LLMs follow a more fixed structural arc. In NYT, `vanilla` LLMs move from `support` to `proposal` more than twice as often as humans (29.4% vs. 12.3%).

2 Constructing a corpus of debates

Measuring argument collapse requires multiple human and LLM responses to the same contested debate. We use two public debate corpora with multiple responses to the same question.[1]

2.1 Collecting human debate corpora

NYT Room for Debate. Each NYT debate consists of one debate question and a set of invited human-written responses, each written as a short argumentative essay, with a median response length of 352 words. We collect the public-web archive, parse each page into a canonical layout with one question file and one markdown file per responder, and apply basic structural filters before creating the final corpus. [2]

Boston Review. Each BR forum consists of a lead essay and commissioned replies. This provides the same multiple-responses-to-one-debate structure as NYT, but in a longer form. Human responses have a median length of 1,150 words, with more room to develop their argument.

Filtering for durable debates. We exclude debates whose answers depend heavily on fast-changing events, because LLM responses to such debates would confound argument collapse with training-cutoff drift. To do this, we tag each debate question for temporal change rate following the FreshQA framing (Vu et al., 2024), and only keep debates where neither of two LLM taggers labels the debate question as `fast_changing`. [3]

Final datasets. From the filtered NYT archive, we sample **195 debates** with broad topic coverage and a balanced question-type split: **1,039 human response essays** across 97 binary and 98 open-ended debates. [4] Our final BR corpus contains **448 human response essays** across **61 forums**. See §A.1 for artifact-use and intended-use details.

[1] We collect publicly accessible pages and linked debate pages from `https://www.nytimes.com/roomfordebate` and `https://www.bostonreview.net/forums/`.

[2] Each debate must contain at least three responses, and each response must be at least 50 words.

[3] `fast_changing` if the natural answer changes within roughly a year, `slow_changing` if it changes over several years, or `never_changing` if it is essentially static. We use `gpt-5.4-mini` and `gemini-3-flash-preview` as the two judges. They return a single label and a short rationale grounded in the debate question; model-call settings for both taggers are reported in Table 5, and the tagging template is in §G.2.

[4] We tag NYT debates by question type and broad subject area. Question-type tags identify binary debates with clear support/oppose sides for the stance analysis. Topic tags are used only to avoid a topically concentrated sample. The tagging templates are in §G.2.

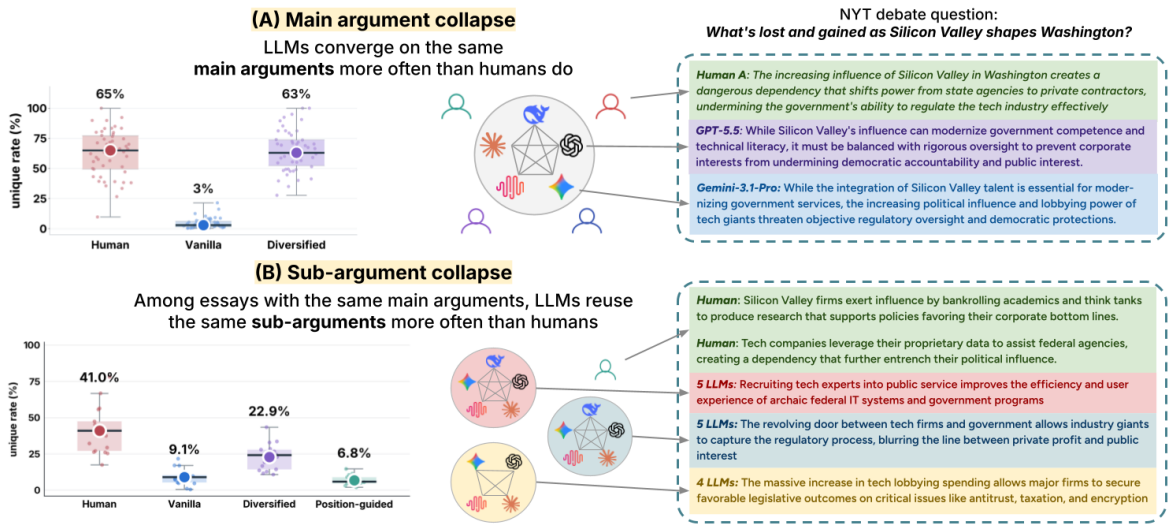

Figure 1: **Argument collapse at two levels of content. (A) Main-argument collapse:** LLMs converge on the same central arguments more often than human writers do. **(B) Sub-argument collapse:** among essays with the same main arguments, LLMs reuse the same supporting sub-arguments more often than human writers.

2.2 Collecting LLM responses

For debates in both corpora, we generate corresponding responses from five frontier LLMs, GPT, Claude, Gemini, DeepSeek, and Minimax, under three generation conditions, `vanilla`, `diversified`, and `position-guided`.[5] Across the three generation conditions, we collect **23,381 LLM essays**, with **16,661** for NYT and **6,720** for BR.[6]

Vanilla. The `vanilla` condition measures each LLM's natural answer to the debate. We give the model only the debate question and sample one response per API call, repeating this N times for each model, where N is the number of human responders in the debate.[7] From these samples, we identify one representative main argument for each model. The resulting comparison is between the human writers' main arguments and five model-level representative arguments, one per LLM.

Diversified. The `diversified` condition tests whether missing main-argument diversity can be elicited directly. We ask the model to produce a *set* of N essays in a single API call, where the instruction explicitly asks the model to vary central claims, supporting arguments, argumentative flow, and discourse moves as widely as possible.[8]

Position-guided. We use `position-guided` to test a stronger form of guidance. The LLM receives the debate question plus an anonymized sketch derived from one human responder in the same debate, which includes the human's main argument, bio, and tone. The LLM is then asked to write one essay from that writer's perspective.[9]

2.3 Corpus and stance annotations

Using `gemini-3-flash-preview`, we tag each NYT debate by question type, distinguishing binary debates with clear support/oppose sides from open-ended debates. This yields 97 binary and 98 open-ended debates. For binary debates, we also label each essay's final stance. A judge first identifies the debate's support and oppose sides as concise statements, and a second judge labels each essay as `strong_oppose`, `weak_oppose`, `neutral`, `weak_support`, or `strong_support` relative to those sides. One author independently annotate a random sample of 20 essays using the same

[5] Full model identifiers, model-call settings, and generation templates are reported in Table 5 and §G.1.

[6] The NYT total decomposes as 5,195 `vanilla` + 6,271 `diversified` + 5,195 `position-guided` essays. The `diversified` count exceeds the naive 5 models × 1,039 humans = 5,195 baseline because we additionally sample GPT under a higher reasoning-effort setting as an ablation. BR is 5 models × 3 conditions × 448 humans = 6,720.

[7] For BR, we just sample 5 times for each model per debate.

[8] Outputs are returned as a single text response with marker-delimited essays (`===== ESSAY N =====`) rather than JSON, because pilot runs showed that several models drop or truncate essays under structured-output constraints when asked for more than three at once.

[9] Most instances in the NYT dataset already come with an author bio. To reduce the risk that LLMs recall specific humans from training data, we anonymize this bio with `gemini-3-flash-preview`. Whenever an author bio is not present, we ask the judge to derive a rough sketch based on the human essay itself. See details in §G.2.

stance schema, reaching 70% agreement with the judge's annotations ($\kappa = 0.625$).

3 LLM collapse at the content level

The paired human and LLM responses let us measure collapse in the substance of the arguments themselves. Content collapse occurs when separately generated essays return to the same arguments rather than spreading across the range of arguments human writers make. We investigate this phenomenon for *main arguments* and *sub-arguments*, finding collapse at both levels.

3.1 Analysis setup

We use one pipeline for both main arguments and sub-arguments to extract argument units from each essay, then label pairwise overlap between same-debate units. All extraction and labeling steps in this pipeline use `gemini-3-flash-preview`. See all prompts in §G.3.[10]

Argument extraction. An essay's main argument is the overall claim it defends, and a sub-argument is a discrete supporting claim, piece of evidence, warrant, or qualification that develops or backs the main argument (Toulmin, 1958; Stab and Gurevych, 2017). See example main and sub-arguments in Figure 1. We extract one main argument and a list of sub-arguments per essay using an LLM judge.[11]

Labeling pairwise argument overlap. For each pair of same-type argument units within the same debate,[12] we ask the judge how much the two units overlap using one of four labels: `equivalent`, `strong_overlap`, `weak_overlap`, and `different`. Two author annotators manually label 100 same-debate main-argument pairs to validate this schema. The two annotators reach $\kappa = 0.61$ on exact four-label agreement and $\kappa = 0.80$ after collapsing labels into substantially overlapping (`equivalent` or `strong_overlap`) versus not substantially overlapping (`weak_overlap` or `different`). [13] Unless specified otherwise, we treat `equivalent` and `strong_overlap` pairs as *substantially overlapping*.

Unique rate. For both main and sub-arguments, we ask how often an argument unit is genuinely unique within the group being compared.[14] An argument is reused if it has a substantial overlap with another argument from the same debate and group; otherwise it is *unique*. Because groups can contain different numbers of arguments, we compare them at the same sample size: if we drew m arguments from each group within a debate, what fraction would have no substantial-overlap match inside that sample? Formally, for a group G, let d_i be the number of substantial-overlap matches for argument $i \in G$. At sample size m,

$$U_m(G) = \frac{1}{|G|} \sum_{i \in G} \frac{\binom{|G|-1-d_i}{m-1}}{\binom{|G|-1}{m-1}}.$$

This is the expected fraction of sampled arguments with no match inside the sample. When $m = |G|$, this is simply the fraction of all arguments in the group that have no match. Higher values mean more distinctive arguments; lower values mean more within-group reuse. [15] Other recovery and reuse metrics used in the content analyses are summarized in Table 6.

3.2 Main argument collapse across vanilla and diversified settings

We ask how models, under different generation settings, vary in the main arguments they produce.

3.2.1 LLMs naturally collapse under vanilla prompting

When models receive only the contested debate question, they consistently repeat the same main argument across different debate types. These repeated arguments often overlap with at least one human argument in the same debate, but LLM essays also show weaker stance strength.[16]

[10] Model-call settings for extraction and labeling are reported in Table 5.

[11] One author validates a random sample of 30 human essays for both precision and recall: all 30 extracted main arguments matched the essay, all 134 extracted sub-arguments were present in the source text, and no substantive supporting sub-arguments were missing.

[12] That is, either a pair of main arguments or a pair of sub-arguments.

[13] The final judge agreed with the two annotators on 69% and 63% of exact labels, and on 93% of coarse labels for both annotators. See validation details and example pairs in §B.3.

[14] A comparison group is the set being evaluated for internal repetition within the same debate, such as human-written arguments or LLM-generated arguments.

[15] For main arguments, each essay contributes one unit. For sub-arguments, an essay can contribute several supporting reasons, so the reported rate is a share of extracted sub-arguments rather than a share of essays.

[16] To keep group sizes comparable, we report U_m with $m = \min(N_{\text{human}}, 5)$.

Representative main argument from the "Are Americans Too Obsessed With Cleanliness?" debate	Humans writers	van. reps	div. families
(1) vanilla LLM responses converge on one hedged argument			
While Americans should maintain essential hygiene practices that prevent disease, they should abandon the neurotic pursuit of total sterility driven by social anxiety and marketing.	2/9	5/5	5/5
(2) diversified recovers additional human arguments			
The United States must prioritize and improve hygiene education and enforcement, specifically hand washing, to prevent illness and save lives.	1/9	0/5	4/5
American obsession with cleanliness was initiated by 19th-century social factors and is sustained by advertising campaigns that exploit social anxieties.	1/9	0/5	4/5
(3) diversified rarely recovers more distinctive human arguments			
Hygiene practices and their definitions are culturally relative rather than universal, frequently leading to misunderstandings between different societies.	1/9	0/5	1/5
Rigid control and rule-following in suburban life are ultimately futile because true grace and peace come from accepting life's inherent messiness.	1/9	0/5	1/5
Religious purification rituals primarily facilitate a search for the sacred and must be understood beyond mere psychological wellness or neurosis.	1/9	0/5	1/5
(4) diversified introduces arguments no human raised			
Americans should relax their extreme standards of cleanliness because our obsession with sterilization isolates us and creates an unnecessary barrier to authentic human connection.	0/9	0/5	1/5
Americans are not too obsessed with cleanliness; rather, they are dangerously inconsistent in their hygiene practices when moving from private to public spaces.	0/9	0/5	1/5

Table 1: **Main-argument collapse and partial recovery in one cleanliness debate.** This debate asks whether Americans are too obsessed with cleanliness. Rows show representative main arguments from the observed overlap patterns. Counts indicate how many human writers, vanilla LLM representatives, or diversified LLM families produced a substantially overlapping argument; for diversified, a family counts once if any of its diversified answers matches the row. All five vanilla representatives converge on the same hedged argument. diversified generation recovers some additional human arguments, rarely reaches more distinctive human arguments, and introduces arguments not raised by any human writer. See more examples in §C.2.

Vanilla responses repeat human-like arguments. LLMs from different providers repeatedly converge on the same main arguments, while human writers more often introduce unique claims. As shown in Figure 1A, across 195 NYT debates, 65.3% of human main arguments are unique within a debate, compared with only 3.4% of vanilla LLM arguments. The same pattern holds in the longer BR essays (78.6% vs. 18.4%), with the human rate higher in 58 of 61 forums. The repeated LLM arguments usually remain within the human argument space. In NYT, 77% of vanilla LLM main arguments substantially overlap with at least one human main argument from the same debate. For example, in Table 1, all five LLMs produce the similar hedged *"concede basic hygiene, prescribe moderation"* argument, while human writers also raise several other main arguments.

On binary debates, LLMs show weaker stance strength. For debates with clearly defined support and oppose sides,[17] we evaluate both side balance and stance strength. Counting both weak and strong labels, vanilla LLM essays support the proposition somewhat more often than human essays (56.1% vs. 49.7%) and oppose it less often (34.5% vs. 43.1%); the remaining essays are neutral or noncommittal (9.4% vs. 7.2%). The sharper gap is stance strength: 76.1% of human essays take a strong support or oppose stance, compared with 63.4% of vanilla LLM essays.

3.2.2 LLM behavior under diversified prompting

The vanilla results show that models converge when asked for a single natural answer. We next test whether collapse persists when models are explicitly asked to produce several diverse answers, as users might do when brainstorming.

[17] Identified and labeled using the stance-annotation procedure in Section 2.3.

Diversified prompting increases main argument uniqueness. `diversified` prompting increases the overall unique rate, though the improvement varies by model. GPT remains below the human baseline: 45% of its diversified main arguments are unique, compared with 65.3% of human main arguments in the same debates. Minimax (53%) and Claude (58%) are closer to the human rate, DeepSeek is roughly human-level (63%), and Gemini exceeds the human unique rate (82%).

Diversified prompting only partly recovers human argument diversity. `diversified` prompting increases uniqueness, but it does not fully recover the range of arguments human writers make: a typical `diversified` LLM covers only about half of human main-argument clusters, ranging from 50% for Claude to 55% for Gemini. Pooling all five models raises coverage to 73.9%, but mostly by recovering easier-to-find arguments: arguments made by multiple human writers are recovered 98.1% of the time, compared with 67.8% for one-off human arguments. BR shows the same partial recovery pattern: pooling five `diversified` models recovers 63.1% of human main-argument clusters.

Many diversified arguments have no observed human counterpart. From the LLM side, many `diversified` arguments have no observed human counterpart: only 47.6% to 60.3% of NYT main arguments, and only 27.6% of pooled BR main arguments, substantially overlap with something humans actually said. Broad and direct answers are easier to recover, while specific and narrower proposals are more often missed. The cleanliness debate in Table 1 gives a concrete example: LLMs recover the hand-washing-as-vital and cultural-construct arguments, miss the culturally-relative, control-over-chaos, and religious-impulse arguments, and add an authentic-human-connection argument not observed among the human writers (additional examples in §C.2).

Diversified prompting balances sides, but not stance strength. For binary debates, `diversified` prompting makes LLMs look more like humans in which side they choose. The share of LLM responses supporting the debate proposition falls from 56.1% under `vanilla` to 49.6% under `diversified`, nearly identical to the human share of 49.7%. But only 66.4% of `diversified` responses take a strong support or strong oppose stance, compared with 76.1% of human responses.

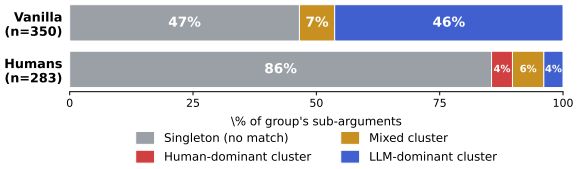

Figure 2: **Per-group distribution of sub-arguments.** Each bar shows the share of a group's sub-arguments in singleton clusters or multi-member clusters with $\geq 70\%$ human or LLM members. Details in §D.3.

Collapse is not explained by memorization. To isolate the effect of potential memorization on argument collapse, we focus on the six most recent BR forums[18] (May 2025 to February 2026) with post-cutoff forums for three of five models. The pattern persists with human arguments remaining more unique than `vanilla` arguments at 84.0% compared with 35.6%, while `diversified` prompting recovers only 53.8% of human main arguments (§C.3).

3.3 Sub-argument collapse

Having established main-argument collapse, we next ask whether the collapse appears in supporting arguments. This question is important because even when essays share the same main argument, they can develop it through different lines of support. To isolate sub-argument collapse, we focus on debates where humans and LLMs have shared main arguments.[19] This subset contains 16 NYT debates, 62 human essays, 80 `vanilla` essays, 321 `diversified` essays and 310 `position-guided` essays. Within each cohort, we compute U_m for humans and three LLM setups using a common sample size m , then macro-average across cohorts.[20]

Sub-argument collapse persists even under shared main arguments. Only 9.1% of `vanilla` sub-arguments are unique, compared with 41.0% for humans. `diversified` recovers some diversity but reaches only 22.9% in Figure 1B. The `position-guided` condition asks whether grounding LLMs in human perspectives restores more varied supporting reasons. As a sanity check, we first hold the target writer fixed and compare the five LLMs conditioned on that same writer. Their sub-

[18] We exclude NYT, as Room for Debate ended in 2018.

[19] We filter the cohorts to ensure that at least three humans share the same main argument and that, on the LLM side, all five LLMs have at least one essay aligned with that same argument. Full details in §D.2.

[20] $m = \min(|H|, |V|, |D|, |P|)$ where H, V, D, P are the humans, `vanilla`, `diversified`, `position-guided` sub-argument pools, so all four groups are compared at the same sample size within each cohort.

Human examples	
Specific case	"The case of **Hurricane Sandy** demonstrates that **inadequate flood-protection studies and infrastructure investment**, ..."
Specific cause-effect	"Heavy-handed federal investment in drug enforcement has led to **over half of the federal prison population being incarcerated for drug offenses**."
Concrete solution	"...could be prevented if companies implemented simple, low-cost measures like **complex passwords and frequent software patching**."
	"Building **bicycle lanes that are physically separated from car traffic** is essential for ensuring the safety of all demographics, including children."
LLM examples	
Generic citation	"**Research indicates** that sex offenders have lower recidivism rates than other felons and that residency restrictions do not actually reduce reoffending." (by Claude)
Abstract concept	"Cybersecurity failures create significant **externalities where the damage of a breach extends far beyond the responsible company** to the general public." (by Minimax)
Abstract solution	"Crime laboratories should be **independent from police and prosecutors** to insulate analysts from investigative pressure and cognitive bias." (by GPT)
Hedged generality	"A **global framework must guarantee** surrogates **comprehensive medical care and long-term health insurance** to protect them from physical complications." (by Gemini)
	"**Effective maternal care requires investing in a continuum of support**, such as home-visiting programs and practical assistance, rather than ..." (by DeepSeek)

Table 2: **Example sub-arguments from humans and LLMs in NYT debates.** Italic tags describe the anchor used.

arguments become highly similar (6.8% unique rate), confirming that position guidance does anchor models to the assigned perspective. We then ask the more substantive question: if the assigned writer changes, will the same LLM produce a wider range of sub-arguments? Holding the LLM fixed and varying the target writer raises unique rate to 18.4% on average across the five LLMs, still well below the human rate. The most diverse LLMs are Minimax and Claude, but they reach only 22.2% and 21.4%. See more details in §D.2.

LLMs converge on generalized arguments and humans on concrete, topic-specific ones. The distribution of sub-argument clusters suggests a clear difference between humans and `vanilla` LLMs (Figure 2). Humans have 85.5% of their sub-arguments in singleton clusters, whereas `vanilla` LLMs have only 46.6% in singletons and another 46.3% in clusters where at least 70% of sub-arguments come from LLMs (*LLM-dominant* clusters). To understand these qualitative differences, we compare sub-arguments from LLM-dominant against human-dominant clusters, and each singletons (details in §D.4). We observe that arguments from humans often stay closer to specific instances. Humans tend to ground their arguments in a specific case, or a concrete intervention and they take a direct position with less hedging. In comparison, LLM arguments more often reach for generalized frameworks such as abstract concepts, cite generic appeals rather than naming a specific case, abstract institutional interventions and hedged generalities that balance multiple stakeholders without committing to any specific action. Table 2 shows the examples. LLM collapse is not just repeated content, but convergence onto generalized frameworks that are hard to falsify and hedged framings that stay on safe ground. One possible explanation is that alignment favor safer and more broadly acceptable responses, pushing models toward generalized supports (Zhang et al., 2025a; Yun et al., 2025; Kirk et al., 2024).

Sub-argument collapse persists in longer essays. To test whether sub-argument collapse generalizes in longer essays, we apply the within-group unique rate analysis from §3.3 to a Boston Review subset of 16 forums, 60 human responses and 70 `vanilla` essays.[21] For each forum, we identify each writer group's largest main-argument cluster and compute U_m separately within each, then macro-average across forums. Across the 16 qualifying forums, the unique rate is 16.3% for `vanilla` LLMs and 42.2% for humans. This result aligns with what we observed in NYT despite the longer essays and different format. See more details in §D.2.

4 LLM collapse at the structure level

The same argument content can be built in many ways: a writer might open with a direct claim, develop evidence slowly, concede an opposing view, narrate an example, or end with a policy proposal.

[21] We apply less strict filtering to BR because it has a smaller dataset and more fine-grained main arguments, so only a few cohorts satisfy the filter. See §D.2.

We therefore ask whether LLM collapse also appears at the structural level. The answer is yes: LLM essays are more likely to follow the same structural arc, opening with a direct claim, moving through support, and ending with a proposal, while human essays vary more in how they build and develop the argument.

4.1 Measuring structural collapse

The structure analysis asks whether essays are organized in similar ways, regardless of the specific arguments they make. We compare paragraph-label patterns directly.

Paragraph-level annotation. For each essay, we tag every paragraph along two orthogonal layers. The *argumentative-role* layer assigns the paragraph's role in the essay's progression: `thesis`, `support`, `counterclaim`, `rebuttal`, `concession`, `reframing`, `implication`, `proposal`, or `none` (full definitions in §E.1). The *discourse-mode* layer captures how the paragraph is written, independent of its argumentative role: `argumentation`, `exposition`, `narration`, or `description`. The two layers run as separate annotation passes with the same configuration.[22]

Structural summaries. We summarize structure in two ways. First, we divide each essay into eight normalized position bins and measure which paragraph labels appear near the beginning, middle, and end. Second, we measure paragraph-to-paragraph transitions, such as whether a `support` paragraph is followed by more `support` or by a `proposal`.

4.2 Evidence for structural collapse

LLMs follow a more fixed structural arc. Human essays have a recognizable but flexible organization: they need not begin with a compact `thesis`, and they continue mixing support, explanation, and occasional narration across the essay. LLM essays are more regular. In both NYT and BR, `vanilla` LLM responses are more likely than humans to open with `thesis`, move into `support`, and end with `proposal`. In terms of discourse mode, LLM paragraphs are more likely to be labeled `argumentation`, while human responses tend to mix in more `exposition` throughout. The same pattern appears in guided generation, as shown in Figure 8 and Figure 9.

Humans develop support, LLMs move sooner to proposals. In terms of paragraph transitions, human essays are more likely to continue developing support. `support` → `support` appears in 50.5% of NYT and 54.5% of BR human transitions, compared with 36.0% and 29.7% for `vanilla` responses. `vanilla` LLMs instead move from support to resolution more often: `support` → `proposal` appears in 29.4% of NYT LLM support transitions and 17.7% of BR LLM support transitions, compared with 12.3% and 7.2% for humans. See details in Table 28 and Table 29.

Diversified and position-guided essays still look structurally LLM-like. `diversified` and `position-guided` essays remain close to the `vanilla` arc. In NYT, they slightly reduce the tendency to open with `thesis` and slightly increase `exposition` or `narration` discourse, but `argumentation` still accounts for 89.7% of `diversified` paragraphs and 89.1% of `position-guided` paragraphs, compared with 71.5% for humans. In BR, the guided conditions also move some paragraph-position trends toward humans, but generated essays still rely more heavily on `argumentation` and show less variation than human responses.

5 Do Evaluators Reward Collapse?

As discussed in §3.3, one possible explanation for argument collapse is that alignment favors broadly acceptable responses over rare ones. We therefore ask whether widely-used evaluators show a similar preference. We examine two reward models trained on human preference data, Skywork-Reward-V2 (Liu et al., 2026) and ArmoRM (Wang et al., 2024), two LLM-as-a-judge protocols following MT-Bench (Zheng et al., 2023) and WildBench (Lin et al., 2025), and two quality-based evaluators, an LLM judge scoring *argument quality* with the 15 dimensions of Wachsmuth et al. (2017) and a creative-writing preference reward model (*Writing RM*; Fein et al., 2026). See more details in §F.1.

Evaluators prefer common arguments over rare ones. We compare shared- against unique-argument essays within the 102 debates where both types occur in the human and `diversified` LLM pools (Table 3, left). All six evaluators favor shared-argument essays, with all but one comparison significant. Human-preference evaluators show a clear

[22] Each annotation pass receives the full essay text and the layer-specific taxonomy; model-call settings are reported in Table 5,annotation prompts are in §G.4, and human validation results are in Table 4.

Evaluator	Shared vs. unique: shared wins (%)		Same argument: LLM wins (%)	
	Human	LLM	Main Arg	Essay
Skywork-V2	74.7	72.3	84.3	100.0
ArmoRM	67.0	66.3	77.5	100.0
MT-Bench	64.3	69.4	86.9	89.3
WildBench	70.3	64.0	–	89.5
Argument quality	62.0	63.7	62.7	96.2
Writing RM	59.3†	61.8	55.0†	50.7†

Table 3: **Evaluator preferences on shared-vs.-unique comparisons and same-argument pairs.** Each number is the percentage of debates favoring the shared argument (left) or the LLM essay (right) under debate-level sign tests (†: not significant; all others $p < 0.05$). WildBench is essay-only due to its checklist-based criteria.

preference for shared arguments, perhaps reflecting a tendency to favor more widely held views.[23] The effect is weaker but persists under evaluations of argument quality and creative writing quality. This result highlights a potential incentive for evaluators to reward more common arguments.

Evaluators prefer the LLM essay under same argument. We compare 550 human and `vanilla` pairs across 140 debates with equivalent main arguments, evaluating both the full essays and main arguments (Table 3, right). All evaluators except *Writing RM* favor LLM outputs at both levels, with the preference particularly strong for full essays. The preference persists after paraphrasing human claims (§F.3), indicating a potential incentive for how LLMs develop the same argument.

6 Related Work

Diversity collapse and surface idiosyncrasy in LLM writing. *Model collapse* describes tail loss when models are recursively trained on their own outputs (Shumailov et al., 2024). Prior work finds lexical and content diversity loss in co-writing and revision (Padmakumar and He, 2024; Anderson et al., 2024; Abdulhai et al., 2026), semantic and voice shifts under minimal editing (Jiang et al., 2026; Röttger et al., 2026a), and constrained coverage of political and epistemic content (Santurkar et al., 2023; Argyle et al., 2023; Röttger et al., 2026b; Wright et al., 2025; Durmus et al., 2024). Other work attributes these patterns to typicality bias, prompt format, or alignment pressures (Zhang et al., 2025a; Yun et al., 2025; Chakrabarty et al., 2025; Tu et al., 2026). Prior work shows models generate recognizable lexical and stylistic patterns even under paraphrase or style changes (Sun et al., 2025; Bitton et al., 2025; Russell et al., 2026b). LLM-assisted writing shifts user beliefs, narrows brainstorming, and influences decisions such as voting in field settings (Jakesch et al., 2023; Sharma et al., 2024; Fisher et al., 2025), and Wen et al. (2026) introduce *entropy collapse* in automated research idea generation, where ideas converge on a narrow shared set across LLMs and prompts. In this work, we study longer form essays and measure diversity across arguments, supporting reasons, and rhetorical structure.

Argument-level evaluation and mitigation. Argumentative content has long been studied through argument mining (Stab and Gurevych, 2017; Wachsmuth et al., 2017; Gupta et al., 2024), discourse analysis (Hyland, 2005) and quality assessment comparing with human and llm (Herbold et al., 2023). More recent work asks how to recover lost diversity in LLM outputs through prompting or sampling (Zhang et al., 2025a; Hayati et al., 2024), configurable pluralism prompts (Nie et al., 2026), multi-perspective generation (Wu et al., 2025; Zhang et al., 2025b), and lightweight editing pipelines (Jiang et al., 2026). Persona prompting goes further by conditioning generations on a target identity, role, or stance (Samuel et al., 2024; Li et al., 2025; Du et al., 2025; Shin et al., 2025). Related work studies strategy-level patterns in LLM-generated conversation (Poungpeth et al., 2026).

7 Conclusion

We define *argument collapse* as the tendency of independently generated essays to converge on the same small set of plausible arguments. Across *New York Times* debates and longer-form *Boston Review* forums, we find collapse at three levels: main arguments, supporting reasons, and argumentative structure. Model-generated essays converge on fewer main arguments, reuse supporting reasons more often even under a shared main argument, and follow a more standardized argumentative arc. Moreover, automatic evaluators reward shared arguments more than unique ones. The risk is not only surface idiosyncrasy or opinion bias, but a narrowing of the range of arguments readers encounter, potentially amplifying dominant ones and limiting long-tail reasoning.

[23] This preference holds even when scoring the main argument level (§F.3).

Limitations

Our study focuses on measuring argumentative diversity rather than argument quality or factual accuracy. Even if humans generate more distinctive arguments, this does not necessarily mean those arguments are better than others. Thus, our results should not be interpreted as showing that human essays are always more persuasive, more accurate, or more preferable than LLM-generated essays. Second, the essays were written at different times. The human essays were written earlier, while LLM responses were generated later and may reflect differences in training data and temporal knowledge. We try to address this concern by filtering out debates that depend heavily on fast-changing events. However, subtle temporal mismatches cannot be fully eliminated, so this remains a limitation. Some of the older debates may also have appeared in the models' training data. We find the same collapse pattern in recent Boston Review forums published after the models' training cutoffs (§C.3), but cannot fully rule out memorization effects for older debates. Third, our dataset focuses on public debate forums, so our findings may not transfer directly to other domains, such as research writing or legal reasoning. Lastly, several parts of our analysis rely on LLM-based annotations. While we report inter-annotator agreement (IAA), these annotations are still imperfect and may introduce systematic biases.

Ethical considerations

LLMs are used for writing assistance, not for generation from scratch.

Acknowledgments

We thank the University of Maryland Computational Linguistics and Information Processing (CLIP) Lab for their feedback and support. This project was partially supported by awards IIS-2626013 and IIS-2545884 from the National Science Foundation (NSF). We also thank Google for a Cloud Credit award that enabled this research. We also thank Professor Shi Feng for feedback on an early version of the work.

References

Marwa Abdulhai, Isadora White, Yanming Wan, Ibrahim Qureshi, Joel Leibo, Max Kleiman-Weiner, and Natasha Jaques. 2026. How llms distort our written language.

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, Chicago, IL, USA. ACM.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Yehonatan Bitton, Elad Bitton, and Shai Nisan. 2025. Detecting stylistic fingerprints of large language models. *ArXiv preprint*, abs/2503.01659.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. AI-slop to AI-polish? aligning language models through edit-based writing rewards and test-time computation. In *Second Conference on Language Modeling*.

Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28).

Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, and Yiqun Liu. 2025. TwinVoice: A multi-dimensional benchmark towards digital twins via LLM persona simulation. *ArXiv preprint*, abs/2510.25536.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2026. LitBench: A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7740–7755, Rabat, Morocco. Association for Computational Linguistics.

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W. Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2025. Biased LLMs can influence political decision-making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6559–6607, Vienna, Austria. Association for Computational Linguistics.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10259–10276, Bangkok, Thailand. Association for Computational Linguistics.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How far can we extract diverse perspectives from large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5336–5366, Miami, Florida, USA. Association for Computational Linguistics.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13.

Ken Hyland. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7(2):173–192.

Shomik Jain, Jack Lanchantin, Maximilian Nickel, Candace Ross, Karen Ullrich, Ashia Wilson, and Jamelle Watson-Daniels. 2025. Task-dependent evaluation of llm output homogenization: A taxonomy-guided framework.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 111:1–111:15. ACM.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2026. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. *Preprint*, arXiv:2310.06452.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 388:1–388:19. ACM.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM-generated persona is a promise with a catch. In *Advances in Neural Information Processing Systems 38*. Position Paper.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2025. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*, pages 47852–47870.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2026. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. In *The Fourteenth International Conference on Learning Representations*.

Shangrui Nie, Kian Omoomi, Lucie Flek, Zhixue Zhao, and Charles Welch. 2026. PERSPECTRA: A scalable and configurable pluralist benchmark of perspectives from arguments. In *The Fourteenth International Conference on Learning Representations*.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Andrew J. Peterson. 2025. AI and the problem of knowledge collapse. *AI & Society*, 40(5):3249–3269.

Nalin Poungpeth, Nicholas Clark, and Tanu Mitra. 2026. Spontaneous persuasion: An audit of model persuasiveness in everyday conversations.

Jenna Russell, Marzena Karpinska, Destiny Akinode, Katherine Thai, Bradley Emi, Max Spero, and Mohit Iyyer. 2026a. AI use in american newspapers is widespread, uneven, and rarely disclosed. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. To appear; oral presentation.

Jenna Russell, Rishanth Rajendhran, Chau Minh Pham, Mohit Iyyer, and John Wieting. 2026b. Storyscope: Investigating idiosyncrasies in ai fiction.

Paul Röttger, Kobi Hackenburg, Hannah Rose Kirk, and Christopher Summerfield. 2026a. Measuring and mitigating persona distortions from ai writing assistance.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2026b. IssueBench: Millions of realistic prompts for measuring issue bias in LLM writing assistance. *Transactions of the Association for Computational Linguistics*, 14:318–340.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. PersonaGym: Evaluating persona agents and LLMs. *ArXiv preprint*, abs/2407.18416.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of LLM-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA. ACM.

Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. 2025. Spotting out-of-character behavior: Atomic-level evaluation of persona fidelity in open-ended generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26312–26332, Vienna, Austria. Association for Computational Linguistics.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. 2025. Idiosyncrasies in large language models. *ArXiv preprint*, abs/2502.12150.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge.

Guancheng Tu, Shiyang Zhang, Tianyu Zhang, Yi Zhang, and Diji Yang. 2026. Shared nature, unique nurture: Prism for pluralistic reasoning via in-context structure modeling.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Fresh-LLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1).

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxin Wen, Liang Qiu, Joe Benton, Jan Hendrik Kirchner, and Jan Leike. 2026. Automated weak-to-strong researcher. Anthropic Alignment Science Blog. Accessed: 2026-05-24.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. 2025. Epistemic diversity and knowledge collapse in large language models.

Fan Wu, Emily Black, and Varun Chandrasekaran. 2025. Generative monoculture in large language models. In *The Thirteenth International Conference on Learning Representations*.

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. 2025. The price of format: Diversity collapse in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. 2025a. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025b. Noveltybench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Data Appendix

A.1 Artifact Use and Intended Use

We use public debate artifacts from *New York Times* Room for Debate and *Boston Review* forums to study how human and LLM-written responses vary when they address the same contested issue. Essays were published for public reading and discussion, and our analysis treats them as public argumentative writing. The essays are written by named

Annotation stage	Human annotation	IAA (human–human)	Judge validation
Stance labeling	1 author, 20 essays	–	vs. author: 70.0% ($\kappa = 0.625$)
Main-/sub-argument extraction	1 author, 30 essays	–	30 main, 134 sub-arguments
Pairwise argument overlap	2 authors, 100 pairs	$\kappa = 0.61 / 0.80$	vs. humans: 93% (coarse)
Paragraph discourse mode	2 authors, 83 paragraphs	84.1% ($\kappa = 0.49$)	vs. gold: 91.5% ($\kappa = 0.71$)
Paragraph argument role	2 authors, 83 paragraphs	54.2% ($\kappa = 0.49$)	vs. gold: 51.3% ($\kappa = 0.44$)

Table 4: **Human validation across annotation stages:** annotation effort, human–human agreement (IAA), and the LLM judge's agreement with human labels. Pairwise-overlap κ values are exact four-way / coarse agreement (Table 7); paragraph-level judge results are measured against an adjudicated gold set, with exact match over label sets for the multi-label argument role.

public op-ed contributors and forum responders, so we treat authorship as already public. For position-guided generation, we anonymize author bios before passing them to the LLMs. We do not redistribute the full original human-written essays. We release only research artifacts needed to audit and reproduce the analysis, such as code, prompts, cohort identifiers, URLs, parsed metadata, derived annotations, pairwise overlap labels, aggregate statistics, and model-generated responses. These materials are intended for research only. They should not be used to republish, replace, or commercially redistribute the original NYT or BR texts. LLM-generated essays were produced through official APIs under the relevant provider terms. We did not perform automated offensive-content filtering, because both source corpora are editorially curated.

A.2 Summary of Human Validation

Table 4 summarizes the human validation performed at each LLM-annotated stage of the pipeline. For the two paragraph-level layers, two authors independently labeled 83 paragraphs on both layers, disagreements were adjudicated into a gold set, and the judge was evaluated against this gold set. The argument-role layer is multi-label, so exact match requires the full label set of a paragraph to agree, which makes agreement substantially harder than in the single-label tasks.

B Methodology Appendix

B.1 Model-Call Hyperparameters

Table 5 reports the generation and annotation settings used across model calls.

B.2 Content Metric Details

Table 6 summarizes the content metrics used in the main-argument and sub-argument analyses. All are computed from the same pairwise overlap labels and use `equivalent` or `strong_overlap` as the substantial-overlap boundary unless otherwise specified.

B.3 Pairwise Argument-Overlap Validation

We manually validated the four-label pairwise schema used for main-argument overlap on a set of 100 same-debate argument pairs from the analysis data. Two authors independently labeled all 100 pairs using the same four labels used by the automatic judge: `equivalent`, `strong_overlap`, `weak_overlap`, and `different`. Annotators followed the labeling rules and used the interface shown in Figure 3. The schema was evaluated at two resolutions. The fine evaluation requires an exact match on the four-way label. The coarse evaluation asks whether raters agree on the analysis-critical boundary between substantially overlapping arguments (`equivalent` or `strong_overlap`) and arguments that are non-overlapping or only weakly related (`weak_overlap` or `different`).

The validation shows that the hardest boundary is the fine-grained distinction among neighboring labels, especially between `equivalent` and `strong_overlap` or between `weak_overlap` and `different`. Agreement is higher on the broader substantial-overlap boundary, so the primary main-argument cluster analyses use `equivalent` and `strong_overlap` as the merge boundary.

During schema development, we also used repeated judge revisions to resolve systematic errors. One revision removed calibration examples drawn from real debates in the analysis set and replaced them with invented examples. Another targeted open-ended debates where two arguments merely shared the debate's broad goal but proposed different central mechanisms; those cases had been over-labeled as `weak_overlap`, and the revised schema directed them to `different`.

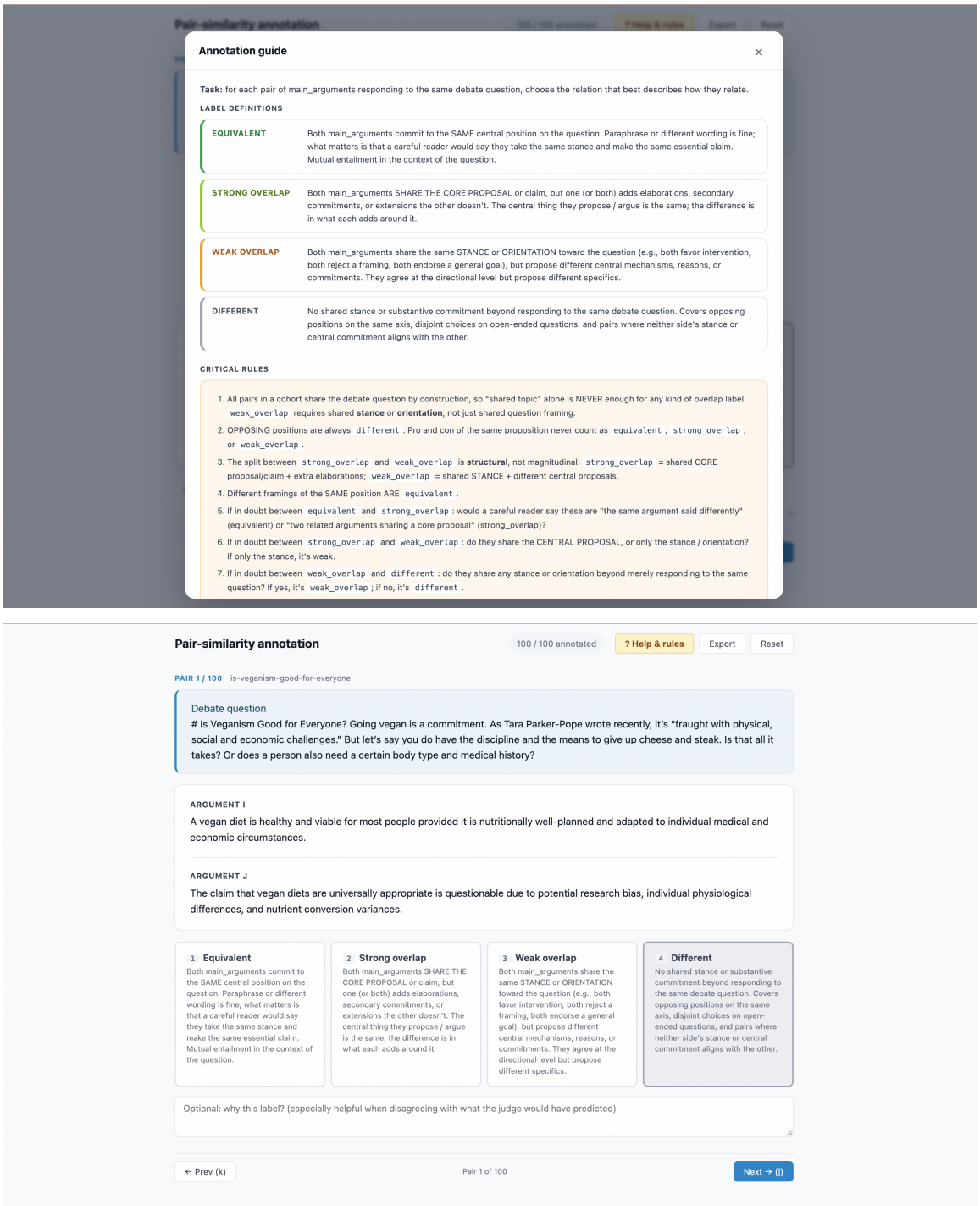

Figure 3: **Annotation rules (top) and interface (bottom) for the pairwise argument-overlap task.** Each annotator received the four-label rubric, definitions, and decision guidance before labeling, and selected one of `equivalent`, `strong_overlap`, `weak_overlap`, or `different` for each pair.

Model / tool	Stage / task	Hyperparameters
Generation		
GPT-5.5 (OpenAI API)	Essay generation for default (v1a), self-diversified (v15a), and position-guided (v4a) conditions	temperature=1.0; reasoning.effort=medium; top_p not set; no web/search tools; NYT runs used provider output defaults, and Boston Review runs used max_output_tokens=32000.
GPT-5.5 (OpenAI API)	Self-diversification reasoning-effort sensitivity check	Same generation settings as above, except reasoning.effort=xhigh.
Gemini-3.1-Pro-Preview (Vertex)	Essay generation for v1a, v15a, and v4a conditions	temperature=1.0; thinking_level=MEDIUM; top_p not set; no web/search tools; NYT runs used provider output defaults, and Boston Review runs used max_output_tokens=32000.
Claude Opus 4.7 (OpenRouter)	Essay generation for v1a, v15a, and v4a conditions	temperature=1.0; reasoning.enabled=true; verbosity=high; top_p not set; no web/search tools; NYT runs used provider output defaults, and Boston Review runs used max_tokens=32000.
Minimax M2.7 (OpenRouter)	Essay generation for v1a, v15a, and v4a conditions	temperature=1.0; reasoning.effort=medium; top_p not set; no web/search tools; NYT runs used provider output defaults, and Boston Review runs used max_tokens=32000.
DeepSeek V4 Pro (OpenRouter)	Essay generation for v1a, v15a, and v4a conditions	temperature=1.0; reasoning.effort=medium; top_p not set; no web/search tools; NYT runs used provider output defaults, and Boston Review runs used max_tokens=32000.
Annotation and preprocessing		
Gemini-3-Flash-Preview (Vertex)	Topic, question-type, sensitivity, and temporal-change tags for sampling and filtering	temperature=0.0; thinking_level=MINIMAL; max_output_tokens=400; top_p not set; strict JSON post-parse.
GPT-5.4-Mini (OpenAI API)	Temporal-change agreement check	temperature=0.0; reasoning.effort=none; max_output_tokens=400; top_p not set; strict JSON post-parse.
Gemini-3-Flash-Preview (Vertex)	Position-guide extraction for position-guided generation; Toulmin-style extraction of arguments and sub-arguments	temperature=0.0; thinking_level=MINIMAL; max_output_tokens=1200; top_p not set; strict JSON post-parse.
Gemini-3-Flash-Preview (Vertex)	Pairwise main-argument overlap judgment	temperature=0.0; thinking_level=MINIMAL; max_output_tokens=2000; top_p not set; strict JSON post-parse
Gemini-3-Flash-Preview (Vertex)	Pairwise sub-argument overlap judgment, including the cross-stance reuse analysis	temperature=0.0; thinking_level=MINIMAL; max_output_tokens=1800; top_p not set; strict JSON post-parse.
Gemini-3-Flash-Preview (Vertex)	Topic-agnostic bucket-register contrastive descriptions	temperature=0.0; thinking_level=MINIMAL; top_p not set; paragraph-form response.
Gemini-3-Flash-Preview (Vertex)	Stance-axis extraction and five-point essay stance labeling	response_mime_type=application/json; thinking_level=MINIMAL; stage-1 max_output_tokens=400; stage-2 max_output_tokens=300; top_p not set.
Gemini-3-Flash-Preview (Vertex)	Paragraph-level argumentative-role and discourse-mode annotation	response_mime_type=application/json; thinking_level=MINIMAL; max_output_tokens=4000; top_p not set.

Table 5: **Model settings used for generation and annotation.** Provider defaults mean that a parameter was not explicitly set in the request. No web-search or retrieval tools were enabled.

C Main-Argument Appendix

C.1 Self-Diversification Recovery Details

Figure 4 visualizes uniqueness within each model's diversified outputs, and Table 10 reports the corresponding means and confidence intervals. This is a within-model comparison: a diversified main argument is counted as unique if it does not substantially overlap another medium-effort output from the same model in the same debate. We use the same substantial-overlap boundary as the default analysis.

C.2 Which Human Main Arguments Are Recovered?

We also aggregate recovery at the level of human main-argument clusters. A cluster is counted as recovered if at least one medium-effort self-diversified output from any of the five LLMs substantially overlaps that human cluster (equivalent or strong_overlap). Under this pooled view, 73.9% of human clusters are recovered by at least one LLM, and the average human cluster is recovered by 2.45 of the five LLMs.

The strongest pattern is cluster popularity. Human arguments made by multiple writers in the same debate are recovered 98.1% of the time. Human arguments made by a single writer are recovered less often (67.8%). Recovery is also higher for binary debates than open-ended debates (79.0% versus 69.7%). A qualitative inspection of recovered and missed clusters suggests that LLMs most reliably recover broad, direct answers to the debate question, while missed clusters more often involve specific examples, narrower proposals, or author-specific framings. Table 11 gives additional debate-level examples using the same categories as the main-text table.

C.3 Post-Cutoff Boston Review Forums

A natural concern is that the debates and forums in our corpora may appear in the models' training data, so the observed convergence could partly reflect memorization of the original discussions

Metric	Definition
Within-group unique rate U_m	Expected fraction of a group's argument units that have *no* substantial-overlap match with another unit from the same group and debate in a same-sized sample of size m. This is the primary metric for main-argument and sub-argument collapse.
Human-cluster recovery	Share of distinct human main-argument clusters in a debate that are substantially overlapped by at least one LLM-generated main argument. This asks how much of the observed human argument space LLMs reach.
Generated-side human overlap	Share of LLM-generated main-argument clusters that substantially overlap at least one human main-argument cluster in the same debate. This asks how much generated variation falls inside the observed human argument space.
Cluster-size bands	Descriptive grouping of argument clusters by how many units they contain. We use this in appendix analyses to ask whether common human arguments are easier to recover than one-off human arguments.
Cluster-region share ρ	For a sub-argument cluster, $\rho = n_{\mathrm{LLM}}/(n_{\mathrm{LLM}} + n_{\mathrm{Human}})$ is the share of units contributed by LLMs. This supports the appendix analysis of human-dominant, mixed, and LLM-dominant sub-argument regions.
Symmetric reuse / recovery $r(A, B)$	Pair- or pool-level overlap between two essays or writer pools, defined as the average of directional recovery from A to B and from B to A. This is used for sub-argument cross-group and cross-stance analyses, with common-size subsampling where pool sizes differ.

Table 6: **Content metrics used in the analysis.** Metrics are derived from pairwise argument-overlap labels. U_m measures within-group uniqueness; recovery and generated-side overlap compare human and LLM argument spaces; cluster-size and ρ summaries support appendix analyses of which arguments are recovered or repeatedly reused; $r(A, B)$ measures pair- or pool-level sub-argument reuse.

Comparison	Fine agreement	Coarse agreement
Judge–Author 1	69% ($\kappa = 0.58$)	93% ($\kappa = 0.86$)
Judge–Author 2	63% ($\kappa = 0.50$)	93% ($\kappa = 0.86$)
Author 1–Author 2	72% ($\kappa = 0.61$)	90% ($\kappa = 0.80$)

Table 7: **Manual validation of pairwise main-argument labels.** Fine agreement is exact agreement on the four-label schema. Coarse agreement groups `equivalent`/`strong_overlap` as substantially overlapping and `weak_overlap`/`different` as not substantially overlapping.

rather than generation behavior. Since NYT Room for Debate ended in 2018, we probe this using the six most recent Boston Review forums, published between May 2025 and February 2026. This window is fully post-cutoff for Gemini-3.1-Pro-Preview (training cutoff January 2025), includes one post-cutoff forum and one forum at the cutoff boundary for GPT-5.5 (December 2025), and one post-cutoff forum for Claude Opus 4.7 (January 2026); DeepSeek V4 Pro and Minimax M2.7 do not disclose official cutoffs. All essays are generated without web-search or retrieval tools (Table 5), so models cannot retrieve the source discussions at inference time. On these six forums, the collapse pattern is unchanged: 84.0% of human main arguments are unique within a forum, compared with 35.6% for `vanilla` LLM arguments, and `diversified` prompting recovers only 53.8% of human main-argument clusters, consistent with the full-corpus results (§3.2). While this check cannot rule out memorization for older debates, it indicates that the collapse we measure is not driven by exposure to the source discussions.

D Sub-Argument Appendix

D.1 Sub-argument extraction and validation

Sub-argument sampling and validation. We analyze a stratified 30-debate subset of the 195-debate sample, balanced across the main-argument convergence spectrum (10 most divergent debates, 10 most convergent, and 10 middle) and covering 8 of 10 topic categories. We focus on this subset because sub-argument analysis requires exhaustive pairwise comparison across extracted supporting claims, producing $294{,}765$ judged pairs at the chosen granularity. At approximately $\sim \$300$ in annotation cost, this set the practical limit for manual evaluation.

Granularity is fixed at the sub-argument level, defined as single supporting claims grounded in 2–5 sentence spans. Finer segmentation would substantially increase the $O(N^2)$ pair count without commensurate analytical benefit, while also remov-

Label	Debate	Argument A	Argument B	Interpretation
`equivalent`	Privacy and the Internet of Things	To realize the benefits of the Internet of Things, manufacturers and regulators must prioritize security by design, transparency, and consumer privacy protections.	To realize the Internet of Things' potential, manufacturers and regulators must implement security-by-design standards and transparent data controls to ensure user trust.	Same core solution, stated in different wording.
`strong_overlap`	Too Few Fish in the Sea	Sustainable seafood is achievable only by redefining market demand through transparency, science-based regulation, and shifting consumption to lower-impact species.	Ethical seafood consumption requires shifting demand toward lower-trophic species and freshwater fish rather than relying on flawed eco-labeling of popular, overfished predatory species.	Shared core proposal, with one argument adding broader regulation and transparency.
`weak_overlap`	Is Veganism Good for Everyone?	A well-planned vegan diet is healthy for most people, but its success depends on nutritional knowledge, supplementation, and individual medical circumstances.	A strictly vegan diet is not universally optimal because individual biological diversity and medical conditions make it practically or biologically unsuitable for everyone.	Same orientation, but different central claim.
`different`	Can the Market Stave Off Global Warming?	Cap-and-trade systems are an ineffective and inequitable mechanism for addressing climate change, particularly regarding the needs and economic realities of developing nations.	A national carbon-pricing regime, specifically cap-and-trade, is the only feasible and effective method for the U.S. to achieve significant long-term emission reductions.	Opposing positions on the debate question.

Table 8: **Examples of pairwise main-argument overlap labels from the validation set.** Rows show actual same-debate main-argument pairs for which the final judge label and human validation label agreed. These examples were not used as calibration examples in the judge prompt.

ing contextual information needed for meaningful argumentative comparison.

To validate extraction quality, we spot-checked 50 randomly sampled sub-argument extractions against their source essays. Of these, 47 mapped cleanly to source spans and the remaining 3 were lightly paraphrased restatements rather than unsupported abstractions. Given the large within-debate effect sizes observed in the 30-debate subset, scaling the same procedure to the full 195 debates would be expected primarily to narrow confidence intervals rather than change the qualitative direction of the results.

Stance distribution per prompt and family. §5 reports the full five-point stance label distribution for the 97 binary cohorts in the analysis sample, broken down by prompt condition and model family.

Pairwise label transitivity. The four-label pair judge produces graded similarity scores rather than a formal equivalence relation, so we report the empirical transitivity of its labels on the 16-cohort subset. Table 12 reports the label distribution on the third edge (A, C) given that both anchor edges (A, B) and (B, C) fall above a chosen threshold. Equivalence is not strictly transitive (67.3% at the strict anchor), but the labels behave as a well-ordered graded similarity scale: chained near-matches stay near and *never* fall to `unrelated`. This is also why our sub-argument clusters are built as connected components over `equivalent` (strict) or {`equivalent`, `strong_overlap`} (loose) edges rather than as transitive-closure partitions.

D.2 Within-group unique rate: shared-main-argument subset

We isolate sub-argument collapse from main-argument variation by restricting to debate questions where humans and default-prompted LLMs independently converge on the same main argument.

Cohort selection. From the 195-debate sample, we select cohorts that simultaneously satisfy four conditions on the main-argument pairwise judgments (Section 3). Throughout, "loose" edges de-

Example main argument from the "What's lost and gained as Silicon Valley shapes Washington?" debate	div. families
(1) van. collapses onto a hedged frame, also present in all div. pools	**5/5**
GPT *While Silicon Valley's influence can modernize government competence and technical literacy, it must be balanced with rigorous oversight to prevent* ***corporate interests from undermining democratic accountability and public interest****.*	
CLAUDE *While the influx of Silicon Valley expertise significantly improves the efficiency and usability of government digital services, it simultaneously risks* ***compromising regulatory integrity and democratic deliberation through corporate lobbying*** *and a bias toward technological disruption.*	
MINIMAX *While the influx of Silicon Valley talent can improve government operational efficiency, it poses significant risks to* ***democratic legitimacy and the prioritization of public values over corporate interests****.*	
DEEPSEEK *While Silicon Valley's influence in Washington provides essential modernization of public services, it simultaneously threatens* ***democratic accountability by prioritizing technocratic efficiency over regulatory independence and public interest****.*	
GEMINI *While the integration of Silicon Valley talent is essential for modernizing government services, the increasing* ***political influence and lobbying power of tech giants threaten objective regulatory oversight and democratic protections****.*	
(2) div. partially recovers a human argument	
The increasing influence of Silicon Valley creates a dangerous dependency that ***shifts power from state agencies to private contractors****, undermining the government's ability to regulate the tech industry effectively.*	2/5
(3a) But div. misses more distinctive human arguments...	
While Silicon Valley offers valuable technical skills, its ***lack of demographic diversity*** *creates significant blind spots that prevent it from effectively serving the needs of all citizens.*	0/5
The expansion of the H-1B visa program, driven by Silicon Valley's lobbying, should be treated with skepticism because the economic benefits are overstated and the program often ***displaces American workers****.*	0/5
The influence of Silicon Valley in Washington is ***not fundamentally different from that of other innovative industries****, and the government should focus on building internal expertise rather than fixating on the novelty of 'tech' lobbying.*	0/5
(3b) ...and introduces arguments no human raised	
The migration of tech talent into the federal government is a ***necessary corrective*** *that improves the functionality and efficiency of essential public services.*	5/5
Silicon Valley's growing influence in Washington is a strategic effort to preserve exploitative business models by ***dismantling federal labor protections*** *and the social safety net.*	1/5
The increasing influence of Silicon Valley risks prioritizing tech-savvy users while ***marginalizing vulnerable populations*** *who rely on human-centric government services.*	1/5

Table 9: **Main-argument collapse and an invented consensus: what's lost and gained as Silicon Valley shapes Washington?** Sections follow Table 1. **(1)** All five vanilla (van.) outputs cite the same gain (**modernizing government services**) and the same loss (**threatening democratic accountability**), with all five opening "While...". This single hedge is the only argument vanilla produces, and div. never escapes it either ($5/5$). Compared with the cleanliness case(Table 1), diversification recovers less: only one human argument is partially recovered ($2/5$ in **(2)**), while **(3a)** shows three distinctive humans entirely missed ($0/5$ each). **(3b)** shows a second convergence: all five families produce an invention no human raised, "tech-talent migration is a **necessary corrective** for failing public services" ($5/5$). Other inventions stay at single-family margins. Diversification reproduces vanilla's convergence, just outside the human distribution.

note equivalent or strong_overlap, the same boundary used elsewhere in this paper for substantial main-argument overlap.

1. *V cluster.* For each of the five vanilla LLM families, we pick a canonical medoid vanilla essay (the most central essay in that family's modal equivalent-cluster). The five medoids must be transitively connected through loose edges, i.e., the five vanilla LLMs collectively produce the same main argument.

2. *H cluster.* Among humans, we take the largest connected component under loose human–human edges only (humans must agree among themselves, not through LLM essays). This component must contain at least three humans.

3. *Per-human bridge.* Every human in the H cluster must have at least two loose edges

Group	Unique main arguments	Difference from human mean
Human writers	61.9% [57.5, 66.2]	—
GPT	45.0% [40.9, 49.2]	−16.8 pp
Gemini	82.4% [79.0, 85.7]	+20.5 pp
Claude	58.5% [54.0, 63.0]	−3.4 pp
Minimax	53.2% [48.8, 57.4]	−8.7 pp
DeepSeek	62.6% [58.0, 67.2]	+0.7 pp

Table 10: **Main-argument uniqueness under diversified prompting.** Values report the average share of main arguments that are unique within the debate, with 95% cluster-bootstrap CIs in brackets. The human row uses all human writers in each debate, while model rows use medium-effort diversified outputs from one model at a time. This differs from the vanilla comparison in §3.2, which uses a common-size comparison between human writers and five LLM representatives; here the goal is to measure how much diversity each model produces within its own diversified outputs.

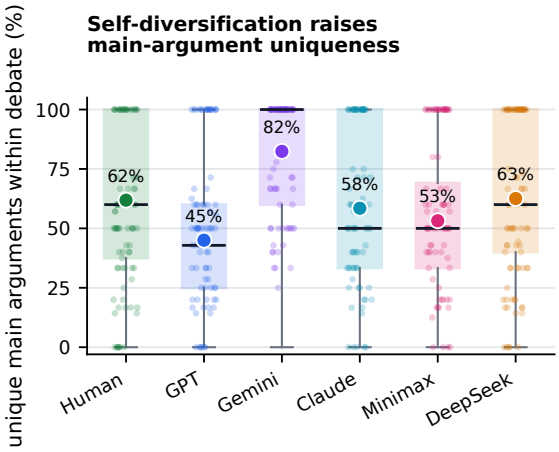

Figure 4: **Diversified prompting raises within-model main-argument uniqueness.** Share of main arguments that are unique within the debate for all human writers and for medium-effort diversified outputs from each LLM family. Small points show debate-level observations; large points show group means.

to distinct `vanilla` medoids in the V cluster. This per-human requirement rules out cohorts where humans are matched to the LLM cluster only through a single weak bridge.

4. *Diversified coverage.* For each of the five LLM families, at least one `diversified` essay must have a loose edge to some human in the H cluster. Enforced at cohort-selection time so that the diversified analysis operates on the same cohort set as the humans/default/position-guided analyses.

The resulting subset contains 16 debate questions, 62 cluster humans, and $5 \times 16 = 80$ `vanilla` medoids (one per LLM family per cohort). The matching `diversified` pool—`diversified` essays loose-connected to some human in the H cluster—contains 321 essays. Table 14 reports per-cohort essay counts and average sub-arguments per essay across all writer conditions; Table 17 reports per-essay sub-argument averages across NYT and BR.

Sub-argument labeling. For each cluster essay we extract sub-arguments and label every inter-essay sub-argument pair on the four-label scale (Section 3). The judged pool covers humans + `vanilla` medoids + all `diversified` essays matched to the H cluster. Total judged pairs across the 16 cohorts: 88,881 (humans–humans: 2,070; humans–`vanilla`: 3,811; `vanilla`–`vanilla`: 3,074; `vanilla`–`diversified`: 18,488; cross-family `diversified`–`diversified`: 53,813). Same-family `diversified`–`diversified` pairs are not judged because they never co-occur in a 1-per-family combination.

Unique rate U_m and common-m subsampling. For a comparison pool P in a cohort, $U_m(P)$ is the expected fraction of P's sub-arguments that have *no* same-pool match in a different essay of the same pool (`equivalent` for strict; `equivalent` or `strong_overlap` for loose). U_m is computed as a closed-form expectation,

$$U_m(P) = \frac{1}{|P|} \sum_{i \in P} \frac{\binom{|P|-1-d_i}{m-1}}{\binom{|P|-1}{m-1}},$$

where d_i is the number of within-pool matches argument i has under the chosen threshold. This is the exact expected value under uniform random sampling of m units from P, with no Monte Carlo noise. Because pools differ in size, we use a single common-m per cohort to make groups directly comparable.

Per-cohort common-m. For each cohort we set

$$m_{\text{global}} = \min\big(|H|,\ |V|,\ m_D^{\min},\ m_P^{\min}\big),$$

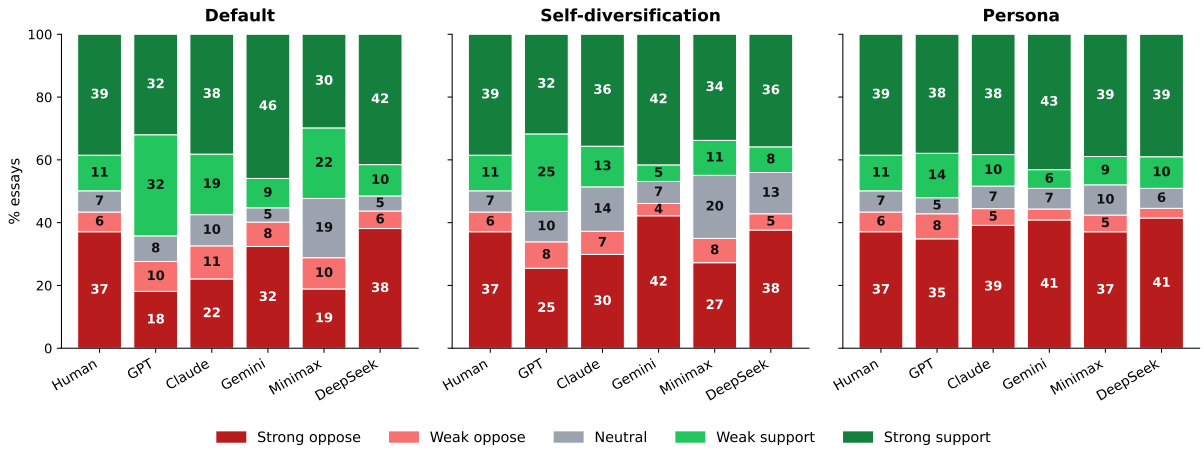

Figure 5: **Stance distribution per prompt, by family.** Five-point stance label distribution for the binary cohorts in the analysis sample ($n = 8{,}496$ essays). Each panel is one LLM prompt condition (default, self-diversified, position-guided); within each panel, the leftmost bar is the human reference and the remaining bars are the five LLM families.

where $|H|$ and $|V|$ are the sub-argument counts of the H cluster and the `vanilla` medoid pool, $m_P^{\min}$ is the smallest `position-guided` family pool (under the family-fixed Setup-2 configuration; see below), and $m_D^{\min}$ is the smallest `diversified` 1-per-family combination pool (sum of the smallest matching `diversified` essay per family). Choosing the global minimum guarantees m is feasible for every pool used in the analysis. All conditions are computed at this single m_{global} per cohort, and per-cohort U_m values are macro-averaged across the 16 qualifying cohorts; final values appear in §15.

Diversified pool construction. `diversified` is the only condition where the relevant pool is itself an average over sampled essay subsets. For each cohort we enumerate all combinations of one `diversified` essay per LLM family (each family must contribute at least one essay matched to the H cluster, by condition (4) above). For each combination, the pool is the union of the five chosen essays' sub-arguments, and we compute $U_{m_{\text{global}}}$ on that pool. The cohort-level `diversified` U_m is the unweighted mean of $U_{m_{\text{global}}}$ over all such combinations. Because the closed-form U_m is deterministic, this enumeration produces an exact expected value over the 1-per-family sampling distribution—no Monte Carlo estimation is required.

Position-guided matched pools (used in recovery analyses). For the cross-group recovery analyses (later in this subsection) we extend the shared-main-argument subset to include, for each of the 62 cluster humans, 5 `vanilla` essays (one per LLM family) and 5 `position-guided` essays (one per LLM family, guided by that human writer), yielding 310 `vanilla` and 310 `position-guided` essays. Pool counts in §13.

Cross-corpus replication: Boston Review modal-main-argument subset. We apply the within-group U_m analysis to the 61-forum Boston Review corpus described in Section 2. The NYT shared-main-argument filter from §D.2 (which requires humans and all five `vanilla` families to converge on the *same* main-argument cluster) is too strict for BR: BR responses develop more fine-grained main arguments and each forum contains fewer essays per group, so cross-group convergence on a single main-argument cluster is rare. We therefore use a per-group modal-cluster variant. For each forum we independently identify (i) the largest connected component of humans under loose human–human main-argument overlap, and (ii) the largest connected component of `vanilla` essays under loose `vanilla`–`vanilla` overlap. From the `vanilla` cluster we select up to five canonical medoids using the same `select_llm_representatives` procedure as for NYT. Forums qualify if both the human cluster and the resulting medoid set contain at least three essays; this yields 16 qualifying forums, 60 cluster human responses, and 70 `vanilla` medoid responses (avg. 3.8 humans + 4.4 medoids per forum). Note that the human and `vanilla` modal clusters in a given forum may correspond to different main arguments—this is a deliberate consequence of the per-group choice, motivated by BR's finer-grained main-argument distribution. We label every inter-essay sub-argument pair on the

four-label scale and compute within-group unique rate U_m per forum under the same common-m procedure as NYT ($m = \min(|H|, |V|)$), then macro-average across forums. Loose-threshold rates: $U_{\text{human}} = 42.2\%$ and $U_{\text{vanilla}} = 16.3\%$ (NYT 41.0% and 9.1%). Strict-threshold rates: $U_{\text{human}} = 89.8\%$, $U_{\text{vanilla}} = 56.8\%$ (NYT 94.9% and 60.6%). The loose `vanilla` rate is higher on BR than NYT (16.3% vs. 9.1%), consistent with BR responses' finer-grained main-argument distribution and smaller per-cluster pool sizes producing more within-cluster heterogeneity. The LLM–human gap is preserved in both loose (−25.9pp BR vs. −31.9pp NYT) and strict thresholds (−33.0pp BR vs. −34.3pp NYT), so sub-argument collapse is robust across corpora with very different essay lengths and forum styles. Per-forum BR breakdowns appear in Table 18.

Cross-group recovery analysis. Within-group unique rate U_m measures whether a group repeats its own sub-arguments. Here we complement that with *cross-group recovery*: how much of one group's reasoning is reachable from another group's pool. We use the same 16 shared-main-argument cohorts.

Metric. For pools A and B, the recovery rate $R(A \rightarrow B)$ is the fraction of A's sub-arguments that share at least one `equivalent` or `strong_overlap` match with some sub-argument in B. The symmetric recovery is $r(A, B) = \frac{1}{2}(R(A \rightarrow B) + R(B \rightarrow A))$, with common-$m$ subsampling to control for pool-size asymmetry. We report two views.

(i) Pool-to-pool overlap. §19 reports r between the cluster-human pool and each LLM-condition pool, macro-averaged across the 16 cohorts. `position-guided` guidance lifts H↔LLM recovery slightly at the strict threshold (12.8% vs. 9.7%) but not at the loose threshold (42.9% vs. 43.8%), indicating that even per-writer position guidance does not substantively expand the set of human sub-arguments reachable from LLM outputs.

(ii) Per-family recovery from humans. For each LLM family f, we compute the recovery of cluster-human sub-arguments into family f's essay pool: $\text{Rec}_f = \text{avg}_i R(E_i \rightarrow P_f)$, where E_i ranges over the cluster humans in a cohort and P_f is the union of family f's essays in that cohort under the given condition. §20 reports per-family recovery for `vanilla` (one essay per family per cluster-human seed per cohort) and `position-guided` (one essay per family per cluster human writer per cohort). Recovery is modestly higher under `position-guided` (44–47% loose) than `vanilla` (37–44% loose), and the spread across families is narrow in both conditions; no family is qualitatively closer to human reasoning than the others.

Cross-stance sub-argument reuse. As a stricter test of sub-argument collapse, we additionally analyze cross-stance sub-argument reuse on a subset of NYT-Room-for-Debate binary cohorts that show stance variation in both writer groups. Across 26 such cohorts, `vanilla` LLMs reuse sub-arguments across opposite stances at 12.7%, roughly twice the human rate of 6.5%. Blinded coding shows LLM cross-stance overlap concentrates on functionalist assessments and definitional claims, whereas human overlap is more often grounded in causal-mechanism diagnoses.

Cohort selection. Starting from the 250 binary cohorts for which essays have stance labels (Section 3), we require, for both the human pool and the `diversified` LLM pool, at least two essays labeled `strong_support` *and* at least two essays labeled `strong_oppose`. We further require that every one of the five LLM families contribute at least one strong-stance essay on each side, otherwise the family-balanced sampling step below is not possible. The intersection of these two criteria yields 26 cohorts, spanning debates across politics, science, technology, sports, religion, and culture.

Essay sampling. For the LLM pool we sample one `diversified` essay per family per stance per cohort (seed 42), giving exactly 5 `strong_support` and 5 `strong_oppose` `diversified` essays per cohort (260 `diversified` essays total). This balances the LLM pool across families so the analysis does not depend on which family is most prolific in any given debate. For the human pool we keep every strong-stance essay (mean 4.85/cohort, range 4–6; 126 humans total across the 26 cohorts). Per-essay sub-argument counts are similarly tight in both groups (humans: mean 4.50, median 4, range $[3, 6]$; `diversified`: mean 4.50, median 4, range $[3, 6]$).

Pair construction. Within each cohort we form all within-group essay pairs (humans–humans or `diversified`–`diversified`; cross-group pairs are skipped) and classify each by stance combination: `SS`, `OO`, or `SO`. Same-essay sub-argument pairs are excluded. Final essay-pair counts across the 26

cohorts appear in §23. For each of the 1,420 inter-essay pairs we label every cross-essay sub-argument pair on the four-label scale (Section 3), yielding 27,439 sub-argument pair judgments using the debate-question context prompt; model-call settings are reported in §5, and the sub-argument prompt is in §G.3.

Symmetric reuse score $r(A, B)$. For each within-group essay pair (A, B) we compute the symmetric pair-level reuse rate

$$r(A, B) = \tfrac{1}{2}\left(r_{A\to B} + r_{B\to A}\right),$$

where $r_{A\to B}$ is the share of A's sub-arguments that match at least one sub-argument in B, and $r_{B\to A}$ is defined symmetrically. This normalizes for differences in $|A|$ and $|B|$. A sub-argument is counted as "reused" if it matches some sub-argument in the other essay with relation `equivalent` or `strong_overlap`. For each (writer-group, stance-combination) we take the per-cohort mean of r, then macro-average across the 26 cohorts so each cohort contributes equally regardless of essay count. 95% confidence intervals are computed by cluster bootstrap at the cohort level (1,000 resamples).

Loose-threshold result. The loose-threshold ({`equivalent`, `strong_overlap`}) macro-averaged rates are reported in Table 21. Both groups drop sharply from same-stance to opposite-stance reuse; the human opposite-stance interval excludes any substantial cross-stance transfer, and the LLM cross-stance rate is roughly twice the human rate.

Strict-equivalence variant ($S = 1.0$). The strict variant counts a sub-argument as reused only at exact equivalence (`equivalent`-only edges). The macro-averaged rates under this stricter threshold are reported in §22; both groups fall well below the loose-threshold rates, and both cross-stance rates fall below 1%, but the LLM rate remains above the human rate at every stance combination.

Per-pair categorization. We categorize cross-stance sub-argument pairs labeled `equivalent` or `strong_overlap` in the 26-cohort subset. Counts: 437 LLM–LLM cross-stance pairs vs. 60 human–human cross-stance pairs, a 7.3× ratio in absolute counts that maps to the ≈ 2× normalized rate in Table 21 after per-essay-size and per-cohort normalization.

Pass 1: per-pair characterization. For each cross-stance pair, the judge sees the debate question, the support-side phrasing, and the oppose-side phrasing, and is asked to output (i) a 2–5 word topic-agnostic TYPE label characterizing the epistemic form of the shared content (e.g., empirical observation, structural diagnosis, normative principle, definitional claim, reform proposal), and (ii) a one-sentence rationale for why this kind of statement can attach to either side of any debate. The prompt explicitly forbids topic-bound labels (e.g., "jurisdictional principle" for drug-enforcement debates) and supplies a candidate list of generic statement types.

Pass 2: emergent clustering. The 437 LLM-side TYPE labels (and separately, the 60 human-side TYPE labels) are passed to a single follow-up call that clusters semantically equivalent labels into 3–6 emergent categories, names each, defines it in one sentence, and lists members by index. The LLM-side clustering yields five categories (causal-mechanism diagnosis, functionalist assessment, normative principle, definitional / categorical boundary, structural reform proposal); the human-side clustering produces an analogous but coarser scheme. Cross-group category shares (LLM 33/19/18/17/13%; human 45/13/18/5/18%) are computed by manually aligning the two schemes. All judging uses `gemini-3-flash-preview` at temperature 0.2, minimal reasoning effort. Representative cross-stance reuse pairs are shown in Table 24.

D.3 Cluster ratio: multi-member ρ-distribution

We characterize how each writer group's sub-arguments distribute across singleton and multi-member clusters. Figure 6 shows the histogram of multi-member cluster LLM-share ρ, and Table 25 reports the per-group breakdown by cluster region (singleton, human-dominant, mixed, LLM-dominant) that underlies Figure 2. Figure 7 extends this per-group breakdown to the `diversified` and `position-guided` conditions.

D.4 Cluster ratio: qualitative analyses

We characterize the contents of the Human-dominant and LLM-dominant regions of the cluster-ratio distribution (§3.3) through three complementary LLM-judge analyses. Together they answer three linked questions: what distinguishes human and LLM multi-member convergence, what remains uniquely human or uniquely LLM when no one else reuses it, and whether the same LLM pattern survives in the larger clusters that attract

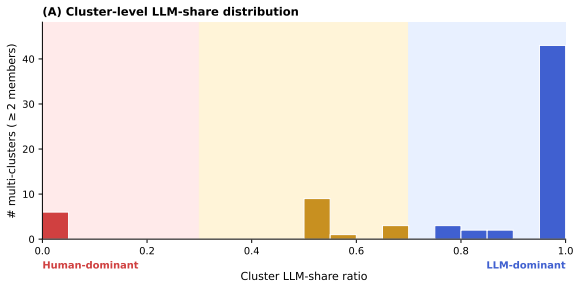

Figure 6: **Distribution of multi-member cluster LLM-share ρ.** Histogram over the 69 multi-member sub-argument clusters ($\geq$ 2 members) in the 16-cohort shared-main-argument subset. Cluster LLM-share $\rho = n_{\text{LLM}}/(n_{\text{LLM}} + n_{\text{Human}})$ is already per-cluster normalized. Background bands mark the Human-dominant ($\rho \leq 0.3$), Mixed ($0.3 < \rho < 0.7$), and LLM-dominant ($\rho \geq 0.7$) regions used in Figure 2. The distribution is sharply asymmetric: 6 Human-dominant, 13 Mixed, 50 LLM-dominant; no mixed cluster contains more humans than LLMs.

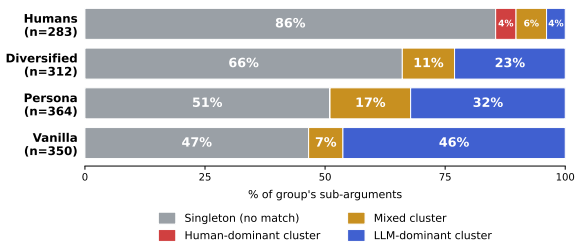

Figure 7: **Per-condition sub-argument distribution across cluster regions.** Each LLM condition is clustered against the same human cluster separately; to prevent pool size from mechanically inflating clustering, each condition's pool is matched to `vanilla` (one essay per LLM family per cohort, mean over 10 samplings). `diversified` and `position-guided` recover diversity only partially (singleton share 66% and 51%), above `vanilla` (47%) but well below humans (86%).

several independent essays. All three analyses operate within the same 16-cohort shared-main-argument subset, use strict-threshold (`equivalent`-only) clusters, and present sub-arguments with the group identity (humans vs. LLMs) masked as "Set A" / "Set B" whenever a direct contrast is being made. The judge is `gemini-3-flash-preview` at temperature 0.4 with 2500 max output tokens.

Cluster construction. Within each cohort, we form sub-argument clusters from `equivalent` edges only (strict threshold) and characterize each multi-member cluster by its LLM share $\rho = n_{\text{LLM}}/(n_{\text{LLM}} + n_{\text{Human}})$. We assign each cluster a region by ρ: *Human-dominant* ($\rho \leq 0.3$), *Mixed* ($0.3 < \rho < 0.7$), or *LLM-dominant* ($\rho \geq 0.7$). For each multi-member cluster, the medoid is the member with the highest mean pairwise equivalence score to the other members.

Analysis 1: blind within-cohort contrasts between multi-member convergence regions. This analysis asks what human and LLM convergence look like when both sides contain a genuinely shared cluster. For every cohort with at least one multi-member ($\geq$ 2) cluster in both Human-dominant and LLM-dominant regions (6 of 16 cohorts qualify), we present all qualifying clusters from each region as a paired contrast (medoid only). Set A / Set B assignment is randomized per call (50/50). The judge is asked to identify 2–4 recurring contrasts between the two sets, name each, and supply one representative phrase per side. One blinded contrast is produced per cohort.

Analysis 2: blind contrasts between human-only and LLM-only singletons. This analysis asks what each group contributes when no one else follows it. We therefore contrast the size-1 clusters (singletons): sub-arguments that belong only to a single essay and have no `equivalent` match anywhere in the cohort. Within each of the 16 shared-main-argument cohorts (242 humans-only and 164 LLMs-only singletons in total), we sample $\min(|S_H|, |S_L|, 20)$ singletons from each side, present them as Set A / Set B with randomized assignment, and request the same recurring-pattern / contrast output as Analysis 1. One blinded contrast is produced per cohort.

Analysis 3: characterization of larger LLM-dominant clusters (size ≥ 3). The first two analyses cover the multi-member and singleton ends, but leave open whether the same picture survives in the larger LLM clusters that attract several distinct essays. We therefore extract every LLM-dominant cluster of size ≥ 3 across the 16 cohorts (26 clusters spanning 15 cohorts), take each cluster's medoid, and ask the judge (a single call seeing all 26 medoids together with their cohort tags) to identify recurring patterns and notable absences. This run is not blinded against a contrastive set; its goal is to characterize what survives convergence at higher cluster sizes, not to compare against humans.

Synthesis. The Pass-1 contrasts of Analyses 1 and 2 and the recurring-patterns output of Analysis 3 are the source of the qualitative claims in §3.3: the same contrast emerges in all three analyses across debates from basketball rule design and cybersecurity to social networks, postpartum care,

and coastal-property policy. Human arguments stay closer to concrete institutions, lived roles, and topic-specific constraints, whereas LLM arguments repeatedly move toward portable mechanism-level abstractions. Table 2 reports the five recurring sentence-level contrasts that survive this synthesis. Per-cohort Pass-1 outputs, the 26 medoid characterizations, and an HTML viewer of all clusters by region are released with the code.

E Structure Appendix

E.1 Paragraph-Level Taxonomies

Table 26 and Table 27 define the two paragraph-label layers used in the structure analysis. The argumentative-role layer captures what a paragraph does in the essay's argument, while the discourse-mode layer captures how the paragraph is written.

E.2 Full Structural Heatmaps

Figure 8 and Figure 9 show the full position-binned label distributions for NYT and Boston Review. These figures expand the main-text structure discussion by showing all generation conditions for both label layers.

E.3 Label-Flow Patterns

Table 28 and Table 29 report selected paragraph-transition patterns for NYT and Boston Review. They support the main-text claim that human essays sustain supporting development more often, while LLM essays move from support toward proposal more quickly.

F Evaluator Analysis Details

F.1 Evaluators

Table 30 summarizes the six evaluators. Reward models (Skywork-Reward-V2-Llama-3.1-8B, ArmoRM-Llama3-8B-v0.1, LitBench-Creative-Writing-RM-3B) score the debate question as the user turn and the evaluated text as the assistant turn using each model's chat template (bf16; 4,096-token truncation). Judge protocols run on `gemini-3.5-flash` at temperature 0, and the 15 argument-quality ratings are averaged into a single score.

F.2 Setup

Data. Pools: 1,039 human essays, 975 `vanilla` (one per model×debate), and 6,271 `diversified`. A main argument is *shared* if another same-pool essay in the debate makes a same-mode argument, and *unique* otherwise.

Same-argument pairs. We compare all 550 human–`vanilla` pairs across 140 debates whose main arguments are `equivalent`, using both full essays and extracted claims. Essays may occur in multiple pairs and debate-level aggregation absorbs this dependence. As a wording control, we also paraphrase each human claim sentence-by-sentence under a content-preservation instruction.

Shared-vs-unique comparisons. We compare shared- against unique-argument essays in the 102 debates where both occur in both pools. A debate favors shared arguments when they win a majority of shared×unique pairs. Pair-independent score evaluators use all pairs (805 human; 13,143 `diversified`; ties $= 0.5$); pairwise judges use K=3 uniformly sampled pairs per debate (fixed seed), evaluated in both presentation orders (win $= 1$, tie $= 0.5$). We aggregate all contrasts by debate and apply two-sided binomial sign tests, dropping tied debates.

F.3 Robustness and Controls

Claim-unit scoring. Scoring the extracted argument text alone preserves or strengthens the reward-model preference (Skywork 73.6/82.0%, ArmoRM 68.4/63.4% of debates). The writing-preference model is unstable at this unit, and no claim-level variant exists for the pairwise judges or the rubric, whose protocols require full essays.

Paraphrase control. Human-claim paraphrasing leaves authorship preference largely intact: Skywork 84→78%, ArmoRM 78→85%, and MT-Bench 87→78% of debates (all $p < 0.005$). The writing-preference model favors LLM claims only against paraphrased humans (64%; original essays and claims n.s.), consistent with mild degradation from paraphrasing.

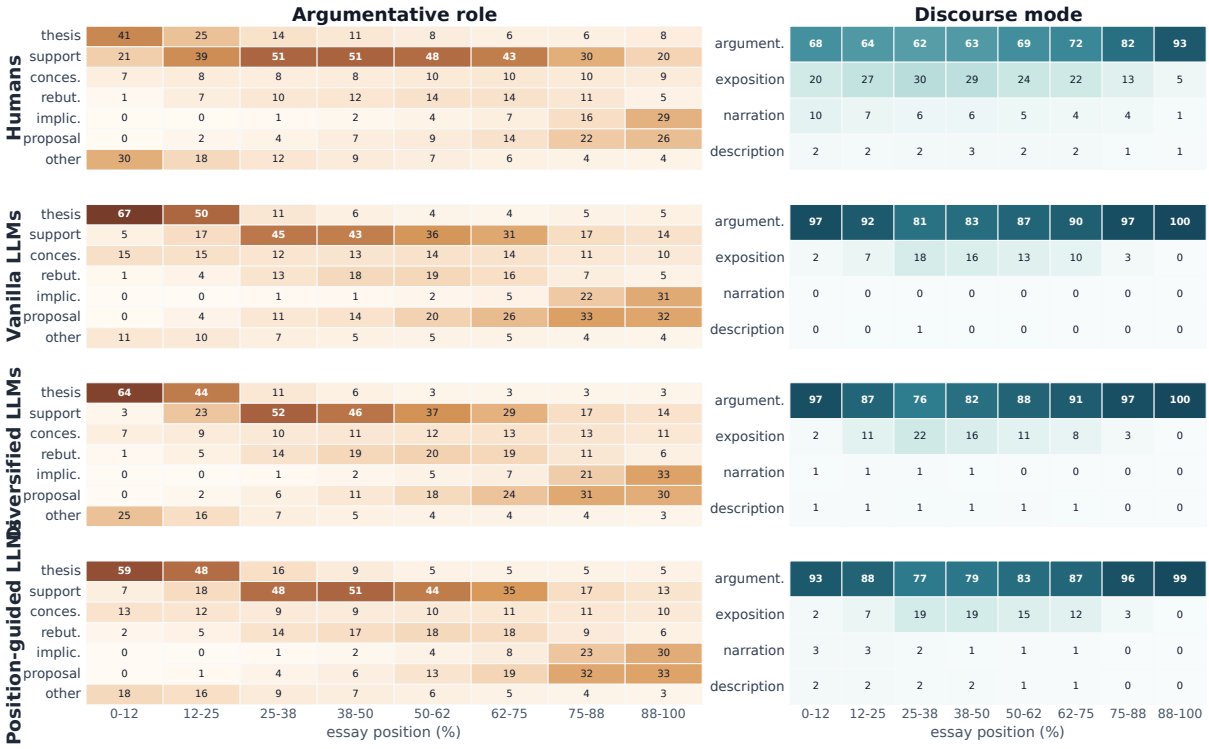

Figure 8: **NYT structural heatmaps across all generation conditions.** Position-binned paragraph-label shares for human essays and LLM essays under `vanilla`, `diversified`, and position-guided generation. The argument layer is multi-label, so cells report the share of role assignments within each position bin; the discourse layer is single-label, so cells report the share of paragraphs.

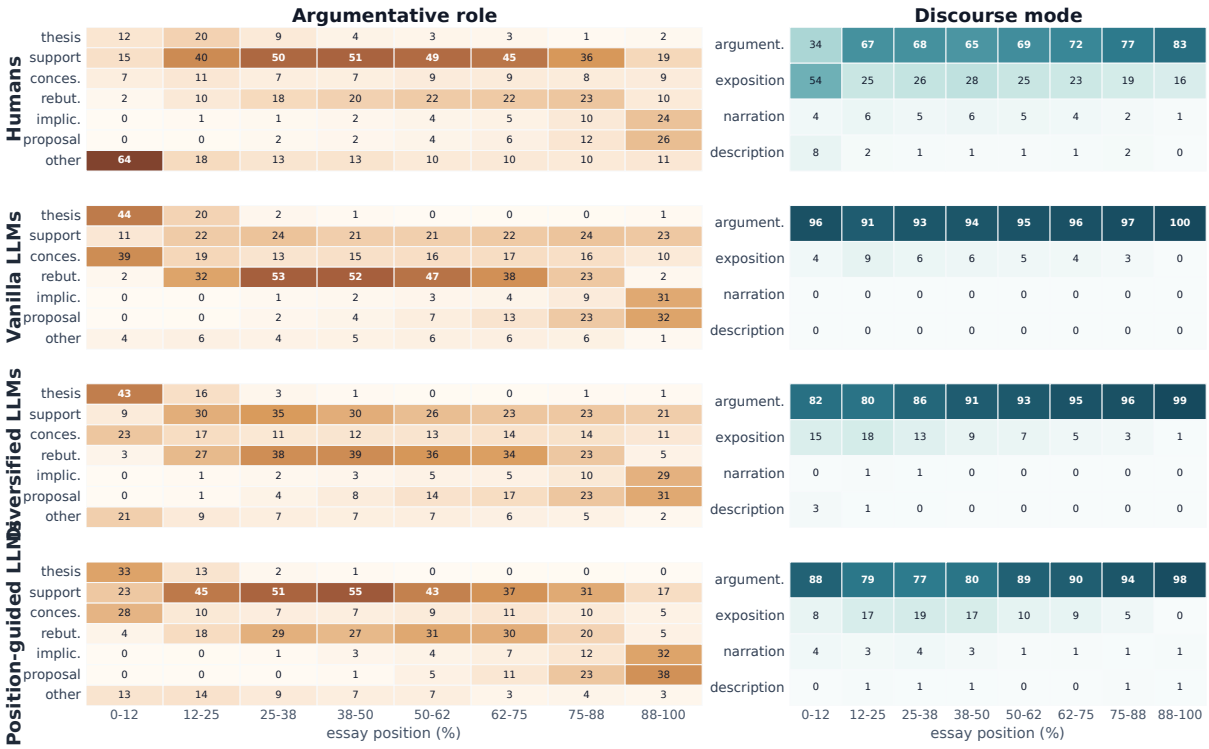

Figure 9: **Boston Review structural heatmaps across all generation conditions.** Position-binned paragraph-label shares for human essays and LLM essays under `vanilla`, `diversified`, and position-guided generation. The argument layer is multi-label, so cells report the share of role assignments within each position bin; the discourse layer is single-label, so cells report the share of paragraphs.

Debate	Argument type	Example main argument	Human side	LLM side
Salt shakers on restaurant tables?	Broad human argument	Restaurants should provide salt because diners should be able to adjust seasoning to their own palates.	3/6	5/5
	Specific human argument	Salt shakers should be removed because seasoning is the chef's prerogative and restaurants should serve food as the chef intended.	2/6	5/5
	LLM-only health argument	Restaurants should remove salt shakers from tables to improve public health by reducing sodium consumption.	0/6	3/5
Is veganism good for everyone?	Broad human argument	Veganism may not be suitable for everyone because of individual health conditions, nutritional deficiencies, and limited long-term evidence.	4/6	5/5
	Specific human proposal	Reducing meat consumption is necessary because of the declining quality and safety of industrial animal products, even if total veganism is not for everyone.	1/6	0/5
	LLM-only proposal	Veganism cannot be good for everyone until social and economic infrastructure makes plant-based eating accessible and sustainable for all.	0/6	5/5
Should federal money rebuild coastal properties?	Broad human argument	Federal disaster spending should shift away from subsidizing reconstruction in high-risk coastal areas and toward relocation and mitigation.	3/5	5/5
	Specific human proposal	The federal government should reduce subsidies and alter land-use policies to limit development in high-risk coastal areas.	1/5	0/5
	LLM-only proposal	Federal rebuilding subsidies should be phased out so private owners bear location risks while public funds are reserved for emergency response and public infrastructure.	0/5	5/5
Should overcrowded national parks restrict access?	Broad human argument	National parks should prioritize mitigation strategies and funding over broad access restrictions, protecting resources while maintaining public access.	1/4	5/5
	Specific human proposal	The National Park Service should raise entrance fees to reduce overcrowding and fund deferred maintenance.	1/4	0/5
	LLM-only proposal	The National Park Service should use restricted access and reservation systems to prioritize preservation over unrestricted tourist enjoyment.	0/4	5/5
Should corporate-funded research be reduced?	Broad human argument	Corporate funding of scientific research need not be reduced because integrity depends on research design, execution, and transparency rather than funding source.	2/4	5/5
	Specific human proposal	Corporate funding should not substitute for federal support, because public funding is essential for basic, high-risk research that industry depends on.	2/4	0/5
	LLM-only proposal	Corporate funding of research on public health, safety, and the environment should be sharply reduced to prevent profit-driven distortion of science.	0/4	5/5
What is the appeal of astrology?	Broad human argument	Astrology appeals not because it is scientifically accurate, but because it offers emotional validation and a framework for self-reflection.	1/5	5/5
	Specific human proposal	Astrology's appeal comes from the human brain's tendency to seek patterns and connections, even when those connections are not empirically valid.	2/5	0/5
	LLM-only proposal	Astrology now functions as a shared language and aesthetic system for social connection and self-expression rather than as a source of prediction.	0/5	3/5

Table 11: **Additional examples of recovered, missed, and LLM-only main arguments.** Each debate repeats the same three categories used in the main-text table. Counts indicate how many humans or LLMs, respectively, produced a substantially overlapping main-argument cluster.

	Label of (A, C)				
Anchor	`equiv`	`strong`	`weak`	`unrelated`	n
Strict ({`equiv`})	67.3%	28.2%	3.4%	0%	19,005
Loose ({`equiv`, `strong`})	11.0%	53.0%	29.3%	0%	493,602

Table 12: **Empirical transitivity of pairwise labels.** For each triple (A, B, C) where both anchor edges (A, B) and (B, C) fall above the threshold, we report the label distribution on the third edge (A, C). Annotated triples only (excludes within-essay pairs and missing pairs). n is the number of annotated triples. Labels are short for `equivalent`, `strong_overlap`, `weak_overlap`, `unrelated`.

group	essays	sub-args
human (cluster)	62	283
`vanilla` (position-guided matched)	310	1,367
`position-guided` (position-guided matched)	310	1,388
total	682	3,038

Table 13: **Pool sizes for the shared-main-argument position-guided experiment.** Aggregated across the 16 shared-main-argument cohorts. Each cluster human contributes one essay per LLM condition per family (5 families); `position-guided` essays are generated from that human writer's anonymized position guide.

	Humans		vanilla		diversified		position-guided	
Cohort	n	avg	n	avg	n	avg	n	avg
inspector-general-police	4	5.00	5	5.00	26	4.77	20	4.65
social-networks-fad	7	4.71	5	4.40	25	4.16	35	4.31
cyclists-drivers-rules	3	4.00	5	4.20	14	4.71	15	4.80
government-wildfires	3	4.33	5	4.40	24	4.42	15	4.40
hiring-surrogacy	3	5.00	5	4.20	13	4.77	15	4.87
young-people-vote	6	3.83	5	4.00	18	4.06	30	3.87
postpartum-depression	3	4.33	5	4.20	11	4.64	15	4.53
natural-disasters-acts-god	3	4.33	5	4.00	19	4.32	15	4.27
robert-durst-forensics	3	5.00	5	5.20	17	4.76	15	4.73
drug-enforcement-states	3	4.33	5	4.40	24	4.79	15	4.40
coastal-properties	4	5.00	5	4.40	14	4.50	20	4.60
cybersecurity-mandates	4	4.75	5	5.00	25	4.92	20	5.15
basketball-rim	5	4.00	5	3.60	25	4.28	25	3.84
purebred-dogs	3	4.33	5	4.20	12	4.50	15	4.33
sex-offenders-restrictions	4	5.25	5	4.40	26	4.96	20	4.90
safer-if-fewer-jailed	4	5.00	5	4.40	28	4.93	20	4.70
TOTAL / avg	**62**	4.58	**80**	4.38	**321**	4.59	**310**	4.52

Table 14: **Per-cohort essay counts and average sub-arguments per essay across the** 16 **shared-main-argument cohorts.** For each writer condition, n is the number of essays in that cohort and avg is the mean sub-argument count per essay. `vanilla` uses one canonical medoid per LLM family (5 per cohort). `diversified` counts include all `diversified` essays loose-matching some human in the H cluster. `position-guided` is $5\times$ the cluster human count (one `position-guided` essay per LLM family per cluster human). Average sub-argument counts are tightly comparable across conditions (per-cohort range 3.6–5.3).

Group	**Strict (`equivalent`)**	**Loose (`equivalent` or `strong_overlap`)**
Humans (cluster)	94.9%	41.0%
`vanilla` LLMs	60.6%	9.1%
`diversified` LLMs (1-per-family, Method 2)	81.0%	22.9%
Same human writer, different LLMs	56.4%	6.8%
Different human writers, same LLM	72.7%	18.4%

Table 15: **Within-group sub-argument unique rates across conditions.** Per-cohort macro-averaged U_m under strict (`equivalent` only) and loose (`equivalent` or `strong_overlap`) thresholds. All values are common-m subsampled at the cohort level. Per-family breakdowns appear in Table 16.

	Within-LLM (family-fixed)		**Default (vs. other families)**	
LLM family	Strict	Loose	Strict	Loose
GPT	69.4%	15.3%	50.2%	6.2%
Claude	81.3%	21.4%	50.6%	7.0%
Gemini	64.4%	16.7%	38.1%	1.6%
Minimax	77.3%	22.2%	63.7%	9.1%
DeepSeek	71.3%	16.3%	44.8%	2.8%

Table 16: **Per-family unique rates.** *Within-LLM (family-fixed)*: U_m across position-guided essays for different human writers within each LLM family. *Default (vs. other families)*: for each family f and cohort, fraction of f's `vanilla` sub-arguments with no equivalent (strict) or equivalent/strong-overlap (loose) match in the union of the other four families' `vanilla` sub-arguments, macro-averaged across the 16 cohorts. Low Default values indicate that the family's `vanilla` reasoning is largely reused by other families' `vanilla` reasoning.

Corpus	Group	Essays	Sub-args	Avg / essay
NYT (16 cohorts)	humans	62	283	4.56
NYT (16 cohorts)	`vanilla`	80	350	4.38
BR (16 forums)	humans	60	311	5.18
BR (16 forums)	`vanilla`	70	348	4.97

Table 17: **Average sub-argument count per essay across corpora.** Humans and `vanilla` produce comparable numbers of sub-arguments per essay within each corpus; BR essays carry slightly more sub-arguments on average, consistent with their longer response length.

	Humans			vanilla		
Forum	n	sub-args	avg	n	sub-args	avg
after_neoliberalism	4	21	5.25	5	26	5.20
authentic_other	7	35	5.00	5	26	5.20
authoritarianism	5	28	5.60	4	21	5.25
campus_protest	3	17	5.67	5	24	4.80
citizenship_emergency	4	22	5.50	5	24	4.80
constitutional_again	3	16	5.33	5	26	5.20
educating_democracy	3	16	5.33	5	28	5.60
effective_altruism	3	14	4.67	5	22	4.40
emre_reproduction	3	16	5.33	5	25	5.00
joseph_carens	3	15	5.00	3	15	5.00
mlk_now	3	15	5.00	3	14	4.67
national_interest	4	19	4.75	4	20	5.00
neurodiversity	3	15	5.00	3	15	5.00
occupy_future	4	19	4.75	5	22	4.40
patriotism_cosmopolitanism	4	21	5.25	5	26	5.20
we_deserve	4	22	5.50	3	14	4.67
TOTAL / avg	**60**	**311**	5.18	**70**	**348**	4.97

Table 18: **Per-forum essay and sub-argument counts in the** 16 **BR modal-main-argument forums.** Humans column lists the size of each forum's largest human main-argument cluster; `vanilla` lists the canonical medoids selected from the largest `vanilla` cluster (up to 5, one per LLM family).

Pair	**Strict**	**Loose**
Humans ↔ `vanilla`	9.7%	43.8%
Humans ↔ `position-guided`	12.8%	42.9%

Table 19: **Symmetric pool-to-pool recovery between humans and LLMs.** Macro-averaged r over 16 shared-main-argument cohorts, common-m subsampled per cohort. Strict counts `equivalent` only; loose counts `equivalent` or `strong_overlap`.

	vanilla		position-guided	
LLM family	Strict	Loose	Strict	Loose
GPT	6.8%	43.6%	10.6%	44.1%
Claude	9.2%	40.0%	13.6%	44.8%
Gemini	10.4%	37.3%	15.3%	46.5%
Minimax	9.7%	39.6%	13.2%	45.0%
DeepSeek	9.2%	40.8%	12.6%	45.8%

Table 20: **Per-family recovery of cluster-human sub-arguments.** For each LLM family f, mean fraction of a cluster human's sub-arguments that have an `equivalent` (strict) or `equivalent`/`strong_overlap` (loose) match in family f's essay pool, averaged over cluster humans and then over the 16 shared-main-argument cohorts.

	same-stance		opposite-stance
group	SS	OO	SO
human	$40.5_{[29.8,\ 51.1]}$	$35.0_{[26.2,\ 44.3]}$	$6.5_{[4.0,\ 9.0]}$
LLM	$66.3_{[59.1,\ 72.9]}$	$60.0_{[53.1,\ 66.9]}$	$12.7_{[9.9,\ 15.8]}$

Table 21: **Cross-stance sub-argument reuse (loose threshold).** Macro-averaged symmetric pair-level reuse rate $r(A, B)$ (%, counting `equivalent` or `strong_overlap` matches) for within-group essay pairs across 26 NYT binary cohorts. Subscripts give 95% confidence intervals from cohort-level cluster bootstrap. Both groups drop sharply from same-stance to opposite-stance reuse; the human opposite-stance interval excludes any substantial cross-stance transfer.

	same-stance		opposite-stance
group	SS	OO	SO
human	$3.7_{[1.0,\ 7.3]}$	$4.3_{[1.0,\ 8.5]}$	$0.5_{[0.0,\ 1.2]}$
LLM	$18.5_{[14.0,\ 24.0]}$	$16.0_{[12.5,\ 19.5]}$	$0.7_{[0.3,\ 1.2]}$

Table 22: **Strict-equivalence variant ($S = 1.0$) of Table 21.** Macro-averaged symmetric reuse r (%) when a sub-argument is counted as reused only at `equivalent` (not `strong_overlap`). Subscripts are 95% cluster-bootstrap CIs.

group	SS	OO	SO	total
human-human	52	47	151	250
`diversified`–`diversified`	260	260	650	1,170
total	312	307	801	1,420

Table 23: **Essay-pair counts for the cross-stance reuse analysis.** Within-group pairs across the 26 binary cohorts after the two-stage selection and family-balanced `diversified` sampling. Cross-group (human × `diversified`) pairs are not used.

Shared sub-argument	Essay-level position
(A) Shared sub-claims: both essays explicitly endorse the same supporting claim.	
Q. "Breeding of a pedigreed dog creates genetic problems that a lovable mutt avoids?"	
Shared sub-claim: aesthetic breed standards drive the genetic health problems.	
Support: "Health testing by breeders is insufficient because **the breed standards themselves often reward dysfunctional anatomy** that causes lifelong suffering."	"The practice of raising purebred dogs according to modern kennel club standards is ethically bankrupt because it prioritizes aesthetic traits over the health and physical functionality of the animals."
Oppose: "The genetic health issues associated with purebred dogs are **caused by aesthetic judging standards and poor breeder choices** rather than the inherent concept of a breed."	"The practice of breeding pedigreed dogs should be reformed to prioritize health and function rather than abolished, as selective breeding is essential for maintaining the predictable traits required for specific working roles."
Q. "Western child labor standards should apply in developing countries?"	
Shared sub-claim: prohibition pushes children into more dangerous informal work.	
Support: "To be effective, labor prohibitions must be paired with corporate and governmental investment in education and financial support for families to prevent children from entering **more dangerous informal work**."	"Western child labor standards should be applied globally because the right to a childhood is a universal principle that protects children from exploitation and long-term poverty regardless of a nation's economic status."
Oppose: "Blanket prohibitions on child labor can be counterproductive by **pushing children into more dangerous, unregulated informal work**."	"Western child labor standards should not be unilaterally imposed on developing countries because they ignore historical context, economic realities, and the structural role Western nations play in creating global poverty."
(B) Stance-neutral primitives: both essays use the same observation or concept, applied differently to reach conclusions.	
Q. "Doping should be allowed in sports?"	
Shared primitive: wealthy athletes or nations enjoy unequal pharmacological advantage.	
Support: "The current testing regime entrenches inequality by **favoring wealthy athletes and nations who can afford sophisticated masking agents and designer drugs**."	"The current anti-doping regime should be replaced with a transparent, medically supervised system because the existing testing process is ineffective, unfair, and dangerous to athlete health."
Oppose: "Allowing doping would exacerbate global inequality by **favoring wealthy nations and athletes with access to superior pharmacological resources**."	"Doping should remain prohibited in sports because legalization would lead to a dangerous pharmaceutical arms race that undermines the integrity of competition and endangers athletes of all ages."
Q. "The quest for energy efficiency pays off for the planet?"	
Shared primitive: efficiency gains are offset by high-consumption lifestyles, so real change requires living smaller.	
Support: "For affluent consumers, the benefits of efficient gadgets are often **negated by high-consumption lifestyles, necessitating a shift toward living in smaller spaces** and reducing overall consumption."	"Energy efficiency should be treated as a matter of public policy and equity rather than a lifestyle choice for wealthy consumers, focusing on making low-carbon living affordable and accessible to everyone."
Oppose: "Meaningful climate action requires deliberate acts of restraint and lifestyle changes, such as **living in smaller spaces and driving less**, rather than substituting old products for new 'green' ones."	"The pursuit of energy-efficient consumer products is an ineffective climate strategy that distracts from the necessary goal of reducing overall consumption and addressing systemic drivers of emissions."

Table 24: **Representative cross-stance LLM sub-argument reuse pairs.** Section **(A)** shows cases where essays on opposite sides explicitly reuse the same supporting claim; section **(B)** shows cases where they reuse the same stance-neutral primitive but attach different value judgments or policy conclusions to it. In both cases, the shared sub-argument alone does not determine the essay's final position; stance emerges from how the broader main argument frames and integrates it.

Group	Singleton	H-dom	Mixed	LLM-dom	Total
Humans	242 (85.5%)	12 (4.2%)	18 (6.4%)	11 (3.9%)	283
vanilla	163 (46.6%)	0 (0.0%)	25 (7.1%)	162 (46.3%)	350

Table 25: **Per-group distribution of sub-arguments across cluster regions.** For each group, count (and % of group total) of sub-arguments that fall into: *Singleton* (cluster size $= 1$), *H-dom* (multi-member, $\rho \leq 0.3$), *Mixed* ($0.3 < \rho < 0.7$), or *LLM-dom* ($\rho \geq 0.7$). Underlies Figure 2. Humans concentrate in singletons; vanilla LLMs split roughly evenly between singletons and LLM-dominant clusters, with no contribution to human-dominant multi-clusters.

Label	Definition	Example	Source
None	No role dominates; transitional, atmospheric, or pure setup.	A purely scenic anecdotal opener.	—
thesis	Directly states the essay's central position.	"The current immigration system is broken, and only comprehensive reform can repair it."	van Eemeren & Grootendorst (2004)
support	Provides grounds, reasons, examples, or problem diagnosis that develops the case.	A paragraph laying out structural causes of a problem.	Walton (1996)
reframing	Recasts what the debate is really about; rejects misleading framing or substitutes a better lens.	"The issue is not whether homeownership rates are down, but for whom declining ownership is harmful."	Entman (1993)
counterclaim	Voices an opposing position; the opposition is the dominant work.	"Critics of this policy argue that it would lead to…"	Walton & Krabbe (1995)
rebuttal	Defeats or undercuts the opposing position; the writer's response dominates.	A paragraph showing why the critics' worry doesn't survive scrutiny.	Toulmin (1958)
concession	Acknowledges that an opposing concern has real force or partial truth, without pivoting.	"To be sure, this reform would impose real short-term costs on rural hospitals."	van Eemeren & Grootendorst (2004)
implication	Draws out consequences from prior reasoning – predictions, takeaways, "therefore X follows."	"This means second-order effects on adjacent industries will follow."	Pollock (1987)
proposal	Advocates a specific course of action – what should be done, by whom, or under what conditions.	"Cities should invest in public transit before subsidizing more parking."	Fairclough & Fairclough (2012)

Table 26: **Argument layer** (paragraph-level, multi-label, $|L| = 8 +$ none). Captures the paragraph's discourse-level argumentative role in the essay's progression.

Label	Definition	Example	Source
argumentation	Claim-and-reason writing with explicit inferential force; reasoning drives sentence-to-sentence progression.	A paragraph that states a policy is flawed, then explains why the flaw matters.	Brooks & Warren (1972)
exposition	Explanation, clarification, or factual information without dominant inferential force.	A paragraph explaining how a program or institution works.	Kinneavy (1971)
narration	Events or actions presented in temporal sequence; the organizing principle is time.	A paragraph recounting a sequence of events leading up to a decision.	Brooks & Warren (1972)
description	Depiction of a scene, state, person, or place; organized spatially or by attributes.	A paragraph depicting conditions in a neighborhood or institution.	Kinneavy (1971)

Table 27: **Discourse Mode layer** (paragraph-level, single-label, $|L| = 4$). Captures how the paragraph is written, independent of its argumentative role.

Pattern	Human	Default LLM	Direction
support → support	50.5	36.0	Humans sustain supporting development more often.
rebuttal → rebuttal	21.2	10.9	Humans more often extend rebuttal across adjacent paragraphs.
concession → support	47.1	40.6	Humans more often return from concession to supporting development.
support → proposal	12.3	29.4	LLMs move from support to resolution more quickly.
rebuttal → proposal	18.2	36.1	LLMs more often close rebuttal with a proposal.
concession → proposal	14.6	28.5	LLMs more often convert concession into proposal.
support → support → support	13.2	5.3	Humans more often develop support over three paragraphs.
thesis → support → proposal	1.6	5.8	LLMs overuse a compact claim–support–proposal template.
thesis → support → rebuttal	2.3	6.3	LLMs overuse a compact claim–support–rebuttal template.

Table 28: **Selected NYT paragraph-transition differences.** Bigram rows report $P(Y \text{ next} \mid X \text{ here})$ as percentages; trigram rows report occurrences per 100 paragraph triples. Default LLM values pool the five debate-only model conditions. Boston Review transition results are reported in Table 29.

Pattern	Human	Default LLM	Direction
support → support	54.5	29.7	Humans sustain supporting development more often.
rebuttal → rebuttal	32.4	52.0	LLMs extend rebuttal across adjacent paragraphs more often.
concession → support	40.8	22.8	Humans more often return from concession to supporting development.
support → proposal	7.2	17.7	LLMs move from support to resolution more quickly.
rebuttal → proposal	10.5	13.8	LLMs slightly more often close rebuttal with a proposal.
concession → proposal	9.3	11.7	LLMs slightly more often convert concession into proposal.
support → support → support	14.7	3.0	Humans more often develop support over three paragraphs.
thesis → support → proposal	0.2	0.1	This compact template is rare in Boston Review for both groups.
thesis → support → rebuttal	0.9	2.7	LLMs use this compact claim–support–rebuttal sequence more often.

Table 29: **Selected Boston Review paragraph-transition differences.** Bigram rows report $P(Y \text{ next} \mid X \text{ here})$ as percentages; trigram rows report occurrences per 100 paragraph triples. Default LLM values pool the five debate-only model conditions.

Evaluator	Type	What it measures
Skywork-Reward-V2 (Liu et al., 2026)	Reward model	Scalar reward trained on human preference data; general response preference.
ArmoRM (Wang et al., 2024)	Reward model	Preference score aggregated from reward dimensions labeled by humans and GPT-4.
MT-Bench (Zheng et al., 2023)	LLM judge	Pairwise verdict based on helpfulness, relevance, accuracy, depth, creativity, and detail.
WildBench (Lin et al., 2025)	LLM judge	Pairwise verdict against a checklist of debate-specific criteria, generated once per debate.
Argument quality (Wachsmuth et al., 2017)	LLM judge	The 15 argument-quality dimensions from argumentation theory, each rated 1–3.
Writing preference (Fein et al., 2026)	Reward model	Writing quality, trained on creative-writing preference data.

Table 30: **Evaluators used in §5.**

G Prompts

G.1 Generation Prompts

Default generation

```
# Default essay generation prompt

## System prompt

You are writing a public-facing response essay to a debate question. The source prompt sets up a contested issue and human
      commentators write response essays.

Your response should:
- Directly engage the source prompt's specific issue
- Be grounded in evidence and reasoning
- Be around 350 words
- Not use external web search; work only from the source prompt and your own knowledge

## User prompt template

Read the debate question below and write a response essay (around 350 words).

Return only the essay body -- no YAML frontmatter, no meta commentary, no title, no label, and no restatement of the prompt
      as a title. Preserve paragraph breaks.

Imagine you are writing on {source_publication_date}. You know everything a thoughtful, well-informed person would have known
       by that date -- events, debates, data, scholarship up to that moment. What you do not have is the benefit of hindsight:
       do not argue from how things turned out later or cite sources that had not been published yet.

## Debate Question

```markdown
{debate_question_markdown}
```
```

Self-diversified generation

```
# Self-diversified essay generation prompt

## System prompt

You are writing a public-facing response essay to a debate question. The source prompt sets up a contested issue and human
      commentators write response essays.

Your response should:
- Directly engage the source prompt's specific issue
- Be grounded in evidence and reasoning
- Not use external web search; work only from the source prompt and your own knowledge

## User prompt template

Read the debate question below and write {n_essays} different response essays (with lengths varying between 250 and 450 words
       -- typical responses run around 350 words).

Return only the essays, separated by markers. Do not add YAML frontmatter, code fences, meta commentary, or any text outside
      the essay bodies and their separator markers.

Use this exact format, with one separator line before each essay:

===== ESSAY 1 =====

[full text of the first essay, with normal paragraph breaks]

===== ESSAY 2 =====

[full text of the second essay]

===== ESSAY 3 =====

[full text of the third essay]

...and so on for all essays. The separator line must be exactly `===== ESSAY N =====` where N is the essay number starting
      from 1.

Imagine you are writing on {source_publication_date}. You know everything a thoughtful, well-informed person would have known
       by that date -- events, debates, data, scholarship up to that moment. What you do not have is the benefit of hindsight:
       do not argue from how things turned out later or cite sources that had not been published yet.

Your task is DELIBERATELY DIVERSE RESPONSE SET GENERATION.

## Diversity goal
```

```
- Produce a VERY DIVERSE set of response essays. Think of the range of responses a group of human writers would produce on
      this question -- real commentators on contested issues can disagree, sometimes sharply, and your essays should reflect a
       diverse spread of positions and frames.
- Vary the stance, central claim, supporting arguments, as well as argumentative flow and discourse moves as much as possible.


## Debate Question

```markdown
{debate_question_markdown}
```

Reminder: return exactly {n_essays} essays, each preceded by its `===== ESSAY N =====` marker, with lengths spread across the
       250-450 word range, centered around 350 words.
```

Position-guided generation

```
# Position-guided essay generation prompt

## System prompt

You are writing a public-facing response essay to a debate question. The source prompt sets up a contested issue and human
      commentators write response essays.

Your response should:
- Directly engage the source prompt's specific issue
- Be grounded in evidence and reasoning
- Not use external web search; work only from the source prompt and your own knowledge

## User prompt template

Read the debate question below and write a response essay (around {word_count} words).

Return only the essay body -- no YAML frontmatter, no meta commentary, no title, no label, and no restatement of the prompt
      as a title. Preserve paragraph breaks.

Imagine you are writing on {source_publication_date}. You know everything a thoughtful, well-informed person would have known
       by that date -- events, debates, data, scholarship up to that moment. What you do not have is the benefit of hindsight:
       do not argue from how things turned out later or cite sources that had not been published yet.

Your task is PERSONA-FAITHFUL SIMULATION: write a response essay as the author profiled below would write it.

## The author

- Background: {role}
- Central claim: {main_argument}
- Tone: {tone}

## Instructions

- The central claim above is the main argument you should stay faithful to when writing your response. Don't shift to a
      different position or thesis.
- Write as this author would: match the background and tone described above.

## Debate Question

```markdown
{debate_question_markdown}
```
```

G.2 Preprocessing Prompts

Topic tagging

```
# Topic tagging prompt

## System prompt

You assign one topic label to op-ed cohorts. Each cohort is one prompt or lead essay that invited responses from multiple
      authors. You see only the prompt/lead.

Return strict JSON only. No markdown.

Pick exactly one `primary_category` from the taxonomy that best captures the core subject of the debate. Tag by substance,
      not disciplinary tradition.

Boundary rules:
- Religious liberty as a free-exercise rights claim -> society. As natural-law morality -> religion.
- Political polarization, partisan disagreement, how-we-talk-about-politics -> politics. The platforms and artifacts of media
       themselves -> culture.
- Contemporary policy debate that uses theory as backup -> the relevant policy bucket. Debate ABOUT the theory itself ->
      philosophy.
```

- Open civic-life questions -> society.
- Education debates -> society. There is no standalone education category.

Taxonomy:
- politics: democracy, elections, parties/polarization, populism, governance, judiciary, constitutional theory, foreign policy, war/security, intelligence/diplomacy, political discourse
- economy: markets, inequality, taxation, finance/banking, monetary policy, labor/work, trade, automation, housing, philanthropy as economic redistribution
- justice: criminal justice, policing, incarceration, drug policy, surveillance/privacy, free speech, guns, due process, civil liberties, animal rights as legal claims
- society: race, gender/sexuality, immigration, education, family structure, social customs/manners, demographics, civic life
- health: healthcare policy, public health/pandemics, addiction, bioethics, mental health, end-of-life, reproductive health
- science_tech: AI/algorithms, internet platforms, biotech/genetics as policy, R&D policy, data privacy, infrastructure, space, scientific integrity
- environment: climate change, energy policy, conservation, agriculture/food, geoengineering, pollution, water resources
- religion: theology, abortion as morality, marriage/family ethics, public morality, secularism, religious liberty as moral rather than legal claim
- philosophy: political/moral/economic theory, applied ethics, aesthetics, pure ideas debates without a contemporary policy hook
- culture: journalism, social media platforms, arts/entertainment, sports, books/film/TV, criticism

JSON schema:

```json
{
  "primary_category": "<one of the category keys>",
  "rationale": "one sentence, max 25 words"
}
```

User prompt template

Venue: {venue}
Cohort: {cohort_id}
Prompt title: {title}

Prompt / lead text (truncated)

```markdown
{prompt_or_lead_text}
```

Question-type tagging

Question-type tagging prompt

System prompt

You classify a public-debate question prompt as either:
- `stance`: the prompt presents a contested claim, a yes/no choice, or asks the responder to take a side on a specific position.
- `open_ended`: the prompt asks for ideas, suggestions, proposals, or exploration without a clear pro/con axis the responder must commit to.

Return strict JSON only with this exact schema:

```json
{
  "question_type": "stance" | "open_ended",
  "rationale": "1-2 sentence explanation grounded in the prompt text"
}
```

Examples:
- "Should the US implement a wealth tax on the ultra-wealthy?" -> stance
- "Is this a reasonable stance? Should China and its partners build an alternative development bank?" -> stance
- "Did the Affordable Care Act make American healthcare better?" -> stance
- "How can states address concerns about illegal immigration, without overstepping their power?" -> open_ended
- "What social problem would you tackle with $100 million?" -> open_ended
- "What advice would you give the new president for his next four years?" -> open_ended

Use `stance` only when there is a clear contested claim or yes/no axis. If the prompt invites brainstorming, suggestions, or open exploration, use `open_ended`.

User prompt template

Classify the `question_type` for the following debate prompt.

Cohort: {cohort_id}
Title: {title}

Prompt body

```markdown
```

```
{prompt_body}
```
```markdown
```

Sensitivity tagging

```markdown
# Sensitivity tagging prompt

## System prompt

You score a public-debate prompt by how SENSITIVE it would be from the perspective of an AI-safety research community
      studying LLM behavior -- i.e., topics where model responses might involve refusal, hedging, value alignment, contested
      politics, identity or demographic groups, violence/harm, or other dimensions where training-data bias and alignment
      choices are likely to surface.

Score on a 3-level scale:
- `low`: mostly empirical, technical, logistical, or aesthetic. Few people would feel strongly on identity/political grounds.
- `medium`: touches contested values or stakeholder conflict but is not a marquee polarized issue.
- `high`: politically polarized, involves identity, demographics, race, religion, sexuality, abortion, violence/harm, war,
      free speech vs. safety, or other fault lines where LLM outputs are known to vary by alignment choices, refuse, or hedge.

Return strict JSON only with this exact schema:

```json
{
  "sensitivity_score": "low" | "medium" | "high",
  "reason": "1-2 sentence explanation grounded in the prompt text; name the specific axis of sensitivity or lack of it."
}
```

Examples:
- "Should the US tighten financial-regulation rules for high-frequency traders?" -> medium
- "Should we charge pre-teens as adults for violent crimes?" -> high
- "What is the best way to clean up the trash in New York subways?" -> low
- "Should religious employers be exempt from contraception coverage mandates?" -> high
- "How should colleges balance free expression with student safety?" -> high

Anchor your score in what the prompt actually asks.

User prompt template

Score the sensitivity of the following debate prompt.

Cohort: {cohort_id}
Title: {title}

Prompt body

```markdown
{prompt_body}
```
```

Temporal-change tagging

```markdown
# Temporal-change tagging prompt

## System prompt

You classify a public-debate question prompt by how quickly its answer would change with time, using the FreshQA taxonomy
      with three labels.

THE TEST

Imagine one thoughtful person who writes a response to this prompt today. They are asked the SAME question again later. Would
       they give a substantively different answer?

- `fast_changing`: yes, within a year or less. The answer hinges on a specific unresolved situation -- an active negotiation,
       pending decision, ongoing crisis, or near-term event -- whose outcome is likely to settle or materially shift the
      answer within the year.
- `slow_changing`: yes, but only over multiple years. The answer would shift gradually as conditions, norms, evidence, or
      institutions accumulate change.
- `never_changing`: no. The answer would essentially be the same regardless of when it is asked. The question is doctrinal,
      philosophical, mathematical, or about a timeless empirical or moral claim.

Do not label based on whether the prompt mentions a recent event, date, named actor, current circumstances, or news. A prompt
       that opens with a recent event is not automatically fast-changing. What matters is whether the answer to the underlying
       question would shift over time.

Return strict JSON only with this exact schema:

```json
{
  "change_rate": "fast_changing" | "slow_changing" | "never_changing",
```
```

```
  "rationale": "1-2 sentence explanation that names specifically what about the answer would change, and on what timescale"
}
```

User prompt template

Classify the `change_rate` for the following debate prompt.

Cohort: {cohort_id}
Title: {title}

Debate prompt body

```markdown
{prompt_body}
```
```

### Persona and thesis extraction

```
Persona and thesis extraction prompt

System prompt

You extract compact persona and thesis metadata from a human response essay. Judge stance relative to the cohort's debate
      question or lead essay. Return strict JSON only.

User prompt template

Extract a reusable JSON persona for LLM response generation.

Required JSON keys:

```json
{
  "slug": "...",
  "name": "...",
  "role": "...",
  "thesis_1line": "...",
  "stance": "strong_agree|mostly_agree|mixed|mostly_disagree|strong_disagree",
  "word_count": 0,
  "voice_abstract": "...",
  "signature_moves_abstract": ["..."],
  "perspective_abstract": "..."
}
```

Rules:
- `thesis_1line` should capture the response essay's central claim in one sentence.
- `role` should use author bio/frontmatter if available, otherwise infer professional vantage cautiously.
- `perspective_abstract` should describe the writer's analytical posture and what they fundamentally prioritize or push
      against -- their worldview on the topic family, not a list of citations.
- Abstract persona fields must not quote or name topic-specific refrains; describe reusable style and evidence habits.
- If unsure about stance, choose `mixed`.

Venue: {venue}
Cohort: {cohort_id}
Response slug: {response_slug}
Author: {author}

Source prompt

```markdown
{source_prompt_or_lead}
```

Human response essay

```markdown
{human_response_essay}
```
```

## G.3 Content Annotation Prompts

### Toulmin-style argument extraction

```
Toulmin annotation prompt

Extract the main_argument and sub_arguments from an essay following Toulmin's argumentation schema.

Runner: `analysis/annotate_toulmin.py`
Output: `analysis_data/toulmin_annotations.jsonl`

v2 system prompt
```
```

```
Given an op-ed essay and the debate question it responds to, extract the
essay's argumentative structure.

Field definitions:
- main_argument: The single central claim the essay makes in response to
  the debate question (1 sentence).
- sub_arguments: Distinct supporting claims that develop or back the
  main_argument. Each one sentence. Avoid restating the main_argument or another sub_argument.

Return strict JSON:
{
  "annotation": {
    "main_argument": "...",
    "sub_arguments": ["...", "..."]
  }
}
```

v2 user prompt format

```
Debate question:
{contents of human/00_question.md}

Essay:
{essay body}
```

Main-argument pairwise judge

```
# Main-argument pairwise judge prompt

## System prompt

You compare two op-ed essays' `main_argument` statements (one sentence each) about the SAME debate question, and decide the
      logical relation between them.

Return exactly one `relation` from a 4-way scale:

- `equivalent`: both main_arguments commit to the SAME central position on the question. Paraphrase or different wording is
      fine; what matters is that a careful reader would say they take the same stance and make the same essential claim.
      Mutual entailment in the context of the question.
- `strong_overlap`: both main_arguments SHARE THE CORE PROPOSAL or claim, but one (or both) adds elaborations, secondary
      commitments, or extensions the other does not. The central thing they propose or argue is the same; the difference is in
       what each adds around it.
- `weak_overlap`: both main_arguments share the same STANCE or ORIENTATION toward the question, but propose different central
       mechanisms, reasons, or commitments. They agree at the directional level but propose different specifics.
- `different`: no shared stance or substantive commitment beyond responding to the same debate question. Covers opposing
      positions on the same axis, disjoint choices on open-ended questions, and pairs where neither side's stance or central
      commitment aligns with the other.

Critical rules:
- All pairs in a cohort share the debate question by construction, so "shared topic" alone is NEVER enough for any kind of
      overlap label. `weak_overlap` requires shared STANCE or ORIENTATION, not just shared question framing.
- OPPOSING positions are always `different`. Pro and con of the same proposition never count as `equivalent`, `strong_overlap
      `, or `weak_overlap`.
- The split between `strong_overlap` and `weak_overlap` is structural, not magnitudinal: `strong_overlap` means shared core
      proposal/claim plus extra elaborations; `weak_overlap` means shared stance plus different central proposals.
- Different framings of the SAME position are equivalent.
- If in doubt between `equivalent` and `strong_overlap`, ask whether a careful reader would say these are the same argument
      said differently or two related arguments sharing a core proposal.
- If in doubt between `strong_overlap` and `weak_overlap`, ask whether they share the CENTRAL PROPOSAL or only the stance/
      orientation.
- If in doubt between `weak_overlap` and `different`, ask whether they share any stance or orientation beyond merely
      responding to the same question.
- Open-ended questions that name a goal in the question itself require a stricter test for shared stance. Two main_arguments
      that both endorse the question's named goal have not shared a stance; they have just answered the question.

Return strict JSON:

```json
{
  "relation": "equivalent" | "strong_overlap" | "weak_overlap" | "different",
  "rationale": "1-2 sentences explaining the relation, grounded in the actual main_argument text"
}
```

Calibration examples

Question: Should Brookline adopt a four-day workweek for municipal staff?
I: "Brookline should adopt the four-day workweek because the pilot data showed productivity held steady while sick leave
      dropped."
```

```
J: "The municipal four-day schedule should be made permanent; the pilot showed comparable output with fewer health absences."
=> {"relation": "equivalent", "rationale": "Both endorse adopting the four-day workweek and both ground that endorsement in
      the same pilot evidence. Paraphrase, same position."}

Question: Should the city require electric scooters to be parked in marked corrals?
I: "Brookline should mandate corral parking and impose escalating fines on repeat violators."
J: "Brookline should mandate corral parking and require operators to provide real-time corral-availability maps in their apps
       ."
=> {"relation": "strong_overlap", "rationale": "Both endorse mandating corral parking, the shared core proposal. I adds fines;
       J adds operator-side maps."}

Question: How should Brookline reduce car traffic on Main Street?
I: "Brookline should redesign Main Street as a transit-and-bike priority corridor with reduced car lanes."
J: "Brookline should raise downtown parking prices and use the revenue to fund a free shuttle loop."
=> {"relation": "weak_overlap", "rationale": "Both share the stance that car traffic should be reduced, but propose different
       central mechanisms."}

Question: Should the Riverton library extend weekend hours?
I: "The library should extend weekend hours because current hours leave working residents unable to use the building."
J: "Weekend hours should stay as they are; further extension diverts limited staff from the children's programming the
      library is known for."
=> {"relation": "different", "rationale": "Directly opposing positions on the question."}

## User prompt template

Debate question:
{debate_question}

Main argument I:
{main_argument_i}

Main argument J:
{main_argument_j}
```

Sub-argument pairwise judge

```
# Sub-argument pairwise judge prompt

## System prompt

You compare two supporting sub-argument statements from op-ed essays that respond to the SAME debate question. A sub-argument
       is a supporting reason, mechanism, evidence claim, or consideration used to develop an essay's main argument.

Decide the logical relation between the two sub-arguments. Return exactly one `relation` from this 4-way scale:

- `equivalent`: both sub-arguments make the SAME supporting point. They may use different wording, but a careful reader would
       treat them as the same reason, mechanism, evidence claim, or consideration.
- `strong_overlap`: both sub-arguments share the same core supporting idea, but one adds details, scope, examples, or
      secondary implications that the other does not.
- `weak_overlap`: both sub-arguments fall under the same broad concern or type of support, but the specific reason, mechanism,
       or evidence claim is different.
- `different`: the sub-arguments do not share a substantive supporting point beyond answering the same debate question or
      belonging to the same broad topic.

Critical rules:
- All pairs share the debate question by construction, so shared topic alone is never enough for an overlap label.
- Focus on the supporting point itself, not whether the two essays have the same main conclusion. Two sub-arguments can
      overlap even if they support different final positions.
- The surrounding main arguments are provided only as context. Do not label two sub-arguments as overlapping merely because
      the main arguments are similar.
- If both statements use the same generic value word, such as fairness, safety, accountability, innovation, or democracy,
      that is not enough for overlap unless they also share the same concrete supporting idea.
- Use `equivalent` only when the two statements would collapse into one support point without losing substantive information.
- Use `strong_overlap` when they share the core support point but one is more specific or adds a related extension.
- Use `weak_overlap` when they share the broad kind of support but make different concrete points.

Return strict JSON:

```json
{
  "relation": "equivalent" | "strong_overlap" | "weak_overlap" | "different",
  "rationale": "1-2 sentences explaining the relation, grounded in the actual sub-argument text"
}
```

Calibration examples

Question: Should Brookline adopt a four-day workweek for municipal staff?
Essay I main argument: "Brookline should adopt the four-day workweek because the pilot improved work quality without reducing
       services."
Sub-argument I: "The pilot showed productivity held steady while sick leave declined."
Essay J main argument: "The four-day schedule should become permanent because the pilot improved staff well-being without
      hurting service delivery."
Sub-argument J: "Trial data showed comparable output and fewer health-related absences."
```

```
=> {"relation": "equivalent", "rationale": "Both sub-arguments make the same supporting point: the pilot maintained
      productivity while reducing absences."}

Question: Should the city require electric scooters to be parked in marked corrals?
Essay I main argument: "The city should require corral parking and fine repeat violators."
Sub-argument I: "Unregulated sidewalk parking blocks curb cuts and makes travel harder for wheelchair users."
Essay J main argument: "The city should require corral parking and make operators map available spaces."
Sub-argument J: "Scooters left outside marked areas obstruct curb ramps, creating access problems for people with
      disabilities and parents with strollers."
=> {"relation": "strong_overlap", "rationale": "Both identify sidewalk obstruction and curb-ramp access as the core support
      point. J adds another affected group."}

Question: How should Brookline reduce car traffic on Main Street?
Essay I main argument: "Brookline should redesign Main Street around transit and bikes."
Sub-argument I: "Reducing car lanes would make buses faster and more reliable."
Essay J main argument: "Brookline should raise downtown parking prices and fund a shuttle."
Sub-argument J: "Charging more for parking would discourage unnecessary car trips into the town center."
=> {"relation": "weak_overlap", "rationale": "Both support reducing car use, but they make different concrete mechanism
      claims."}

Question: Should the Riverton library extend weekend hours?
Essay I main argument: "The library should extend weekend hours for working residents."
Sub-argument I: "Current hours leave residents with weekday jobs unable to use the building."
Essay J main argument: "Weekend hours should stay the same to protect children's programming."
Sub-argument J: "Extending hours would divert limited staff from the children's programs the library is known for."
=> {"relation": "different", "rationale": "The first sub-argument is about access for working residents, while the second is
      about staffing tradeoffs."}

## User prompt template

Debate question:
{debate_question}

Essay I main argument:
{main_argument_i}

Sub-argument I:
{sub_argument_i}

Essay J main argument:
{main_argument_j}

Sub-argument J:
{sub_argument_j}
```

Stance-axis extraction

```
You will receive a debate title and a debate question body.

Your task is to extract the two sides of the debate as very concise statements:
- "support_side": the side treated as the supported position under the annotation convention
- "oppose_side": the opposing side

Goal
Identify the single binary debate axis that defines the debate, then express its two sides concisely.

Rules for choosing the debate axis
1. If the title states a clear binary debate question, use the title axis.
2. Otherwise, use the first explicit binary question in the body.
3. If the body restates the same axis more explicitly, you may use the body wording, but do not change the axis.
4. If the body introduces a different axis from a clear binary title, keep the title axis.
5. Ignore wh-questions (what / how / why / when / where) when identifying the axis.
6. Never merge two different binary questions into one combined side definition.
7. If no usable binary axis can be identified in either the title or the body, return "none" for both sides.

How to map the chosen axis
- For yes/no, should/should-not, is/is-not, can/cannot questions:
  - support = yes / should / is / can
  - oppose = no / should-not / is-not / cannot
- For genuine binary choice questions ("A or B"):
  - support = the first option
  - oppose = the second option

Writing constraints
- Write each side as a short, content-preserving statement, not just "yes" or "no".
- The two sides must be true opposites on the same axis.
- Do not add content from another question or another axis.
- If the title and body use different wording for the same axis, prefer the clearer wording.
- Preserve important scope qualifiers when present, such as time, place, population, or institution (for example: "
      contemporary," "in the U.S.," "for local drug cases"), unless removing them would clearly not change the debate axis.

Return strict JSON only:
{
  "support_side": "string or none",
```

```
  "oppose_side": "string or none",
  "source": "title_binary" | "body_first_binary" | "body_restatement_of_title" | "none",
  "rationale": "One sentence explaining which axis was chosen and why."
}
```

Stance labeling

```
You will receive:
- a support-side statement
- an oppose-side statement
- an op-ed essay

Your task is to label the essay's final position relative to the provided side definitions.

The support_side and oppose_side define the debate axis.
Use them exactly as given.
Do not redefine the issue, substitute a different question, or infer a different debate axis from the essay.

Label set
- "strong_support": The essay clearly and firmly endorses the support side. Its final position is decisive and minimally
      qualified.
- "weak_support": The essay ultimately leans toward the support side, but its final position is meaningfully qualified,
      limited in scope, or noticeably hedged.
- "neutral": The essay does not ultimately commit to either side. This includes balanced discussion, unresolved ambiguity,
      mixed or split positions without a clear final choice, rejecting the premise, or reframing the issue instead of choosing
       between the provided sides.
- "weak_oppose": The essay ultimately leans toward the oppose side, but its final position is meaningfully qualified, limited
       in scope, or noticeably hedged.
- "strong_oppose": The essay clearly and firmly endorses the oppose side. Its final position is decisive and minimally
      qualified.

Important rules
- If the essay ultimately lands on one side, even with qualifications, do not label it "neutral".
- Use "neutral" only when the essay does not finally choose either side.
- If the essay leans more toward one side than the other, use "weak_support" or "weak_oppose" rather than "neutral".
- Judge "weak" vs "strong" based on strength of commitment to the final position, not on tone alone.
- Conditionality suggests "weak" only when it meaningfully limits the final position.
- Strong language alone does not make a label "strong" if the conclusion is substantially qualified.
- Calm or analytical wording can still be "strong" if the final commitment is clear and firm.

Return strict JSON only:
{
  "label": "strong_support" | "weak_support" | "neutral" | "weak_oppose" | "strong_oppose",
  "rationale": "One sentence explaining which side the essay ultimately supports, or why it remains neutral."
}
```

Cluster-ratio contrast: multi-member regions (A1)

```
# Cluster-ratio Analysis 1 prompt: blind within-cohort contrast between multi-member regions

## User prompt template

DEBATE QUESTION: {question}

You will see two sets of supporting-argument clusters from this single debate.
Each cluster is a group of 2+ supporting arguments that different essays
independently produced, expressing essentially the same idea. Each cluster is
shown as its medoid (the most central member).

Both sets are convergence clusters from the SAME debate, but produced by
different (unspecified) writer groups. Your job is to characterize what
distinguishes Set A's convergence from Set B's convergence.

CONSTRAINTS
1. Do NOT speculate about which set is which writer group. Focus on the
   sub-argument content itself.
2. Focus on register, framing, type of evidence, anchor, or move structure
   --- not on topic content alone.
3. Identify 2--4 recurring contrasts (A vs B). For each:
   - Short name (3--6 words).
   - One sentence on what makes A different from B.
   - One representative phrase from A and one from B (paraphrased OK if needed).

SET A ({n_a} clusters)
{a_block}

SET B ({n_b} clusters)
{b_block}
```

Cluster-ratio contrast: human-only vs LLM-only singletons (A2)

```
# Cluster-ratio Analysis 2 prompt: blind contrast between human-only and LLM-only singletons

## User prompt template

DEBATE QUESTION: {question}

You will see two sets of supporting arguments from this single debate.
Each item is a "singleton" --- a supporting argument that no other essay in
this debate paraphrased or restated. Each set is from a different
(unspecified) writer group.

Your job is to characterize what kinds of arguments each group contributes
when nobody else follows their move.

CONSTRAINTS
1. Do NOT speculate about which set is which writer group. Focus on the
   sub-argument content itself.
2. Focus on register, framing, type of evidence, anchor, or move structure
   --- not on topic content alone.
3. Identify 2--4 recurring contrasts (A vs B). For each:
   - Short name (3--6 words).
   - One sentence on what characterizes A vs B.
   - One representative example phrase from A and one from B.

SET A ({n_a} singletons)
{a_block}

SET B ({n_b} singletons)
{b_block}
```

Cluster-ratio characterization: larger LLM-dominant clusters (A3)

```
# Cluster-ratio Analysis 3 prompt: characterization of larger LLM-dominant clusters

## User prompt template

You will see medoid supporting arguments drawn from {n_clusters} sub-argument
convergence clusters, each cluster containing 3 or more independently
generated supporting arguments. The clusters come from {n_cohorts} different
NYT debate cohorts (cohort tags shown for context).

YOUR TASK
Identify recurring patterns of reasoning that appear ACROSS multiple cohorts
in this set, AND notable absences (moves that one might expect in supporting
arguments but do not appear).

CONSTRAINTS
1. A pattern should recur in at least 4 different cohorts.
2. For each pattern, name it briefly (3--6 words), describe in one sentence,
   list 2--3 representative phrases, and list the cohorts where it appears.
3. End with a short list of NOTABLE ABSENCES --- types of supporting moves
   that are conspicuously rare or missing across these clusters.

MEDOID SUPPORTING ARGUMENTS
{block}
```

G.4 Structure Annotation Prompts

Argumentative-role annotation

```
You are performing argumentative role analysis of an op-ed essay responding to a debate prompt. Label each paragraph by its
      dominant argumentative role in the essay's progression.

## Per-paragraph annotation

For each paragraph, identify its dominant argumentative role from the labels below. Use `none` when no label applies. This
      layer is only about argument — style, tone, and evidence type should not influence the choice.

Multi-labeling is allowed. Assign no more than two labels per paragraph, and use two labels only when necessary. Do not
      combine `none` with any other label. Assign two labels only when each label independently dominates a substantial part
      of the paragraph's argumentative function. Avoid redundant secondary labels — e.g., do not add `support` alongside `
      rebuttal`/`reframing` when the supportive function is implicit; the headline move suffices.

### Decision flow (when in doubt)

Walk through these in order. Use the first label that fits the paragraph's *primary* act.

1. Does the paragraph FIRST establish the essay's central position? → `thesis`. **At most one paragraph per essay may be
      labeled `thesis`.** Later restatements/recaps default to `support` (or `implication`/`proposal` if the recap extends
      into consequences or a call to action).
```

2. Does the paragraph directly engage a specific opposing claim that the essay treats as the *other side* of the debate, and give reasons against it? → `rebuttal`. Not just any critique of an alternative view — there must be an identifiable interlocutor stance the essay is responding to.
3. Does it acknowledge that an opposing concern, criticism, or alternative view has real force? → add `concession` (often as a secondary label to thesis/rebuttal/support).
4. Does it present an opposing position neutrally, foregrounding it as the paragraph's main object without taking a stance against it? → `counterclaim`.
5. Does it REPLACE the framing of the debate itself (not just offer an alternative analytical lens within the existing framing)? → `reframing`. If it just develops the author's view by contrasting with an alternative theory while staying inside the same debate, that's `support`, not `reframing`.
6. Does it draw consequences, stakes, or projections from claims already made? → `implication`.
7. Does it call for specific action (what should be done, by whom, under what conditions)? → `proposal`.
8. Otherwise, the paragraph develops the author's affirmative case (sub-claims, reasons, examples, comparisons, contrast with alternatives, critique of competing theories) → `support`. **`support` is the default for affirmative case-building.**

Don't choose `rebuttal` or `reframing` just because a paragraph contains contrast or critique. Choose them only when contrast/critique is the paragraph's HEADLINE act and engages a specific opposing position (rebuttal) or replaces the debate's frame (reframing). Otherwise the paragraph is `support` developing the author's affirmative line.

Rhetorical questions function as the move they imply, not as their literal interrogative form. A closing "Should we accept X?" calling for change → `proposal`. "How can we ignore Y?" affirming a point → `support`. "But isn't this just Z?" engaging an objection → `rebuttal`.

Label set

`thesis`
The paragraph that first establishes the essay's central position. **At most one paragraph per essay can carry the `thesis` label.** Later restatements (e.g., a concluding recap) count as `support`, not `thesis`. The position can be implicit or inferred — subtle declarations count. When the same statement functions as the central position AND takes the form of a recommendation or consequence claim, label that statement `thesis` (not `[thesis, proposal]` or `[thesis, implication]`). A paragraph can still carry both labels if it contains *additional* distinct proposals or implications beyond the central position.

Examples:
- A paragraph asserting: "The current immigration system is broken, and only comprehensive reform can repair it."
- An opening paragraph that crystallizes the position before any developed support.

`support`
A paragraph that advances a supporting sub-claim, reason, or line of reasoning to strengthen the writer's main argument. Also use for *bare* later restatements of the thesis — recaps that add nothing new. If a closing recap extends the thesis into projected consequences, label `implication` instead; if it extends into a call for action, label `proposal` instead. Don't reflexively co-label `support` with `implication`/`proposal` on the same paragraph; pick the dominant move. **`support` is for affirmative development of the author's own line of reasoning — not for paragraphs whose work is engaging or dismantling an opposing position. If the paragraph references an opposing argument and gives reasons against it (even without overt rebuttal markers like "but" or "this is wrong"), prefer `rebuttal`.**

Examples:
- A paragraph arguing that a policy has succeeded because measurable outcomes improved after implementation.
- A paragraph diagnosing a system's flaws to show why the writer's position is needed.
- A closing paragraph that recapitulates the central position without projecting consequences (`implication`) or calling for action (`proposal`).

`reframing`
A paragraph that changes the lens of the debate — rejecting a misleading framing, narrowing the issue, or substituting a better way of seeing the question. **Reserved for paragraphs whose primary act is replacing the framing itself. If the paragraph engages an opposing argument and gives reasons against it, prefer `rebuttal`. If it shifts perspective only as a side effect of developing the author's affirmative line, prefer `support`. `reframing` is the right label when the paragraph's headline move is "the debate as posed is wrong; here's a better way to frame it."**

Examples:
- A paragraph arguing that the issue is not whether homeownership rates are down, but for whom declining ownership is actually harmful.
- A paragraph rejecting a false binary and replacing it with a more useful frame.
- A paragraph opening with "The real question is not X but Y" and developing the better frame.

`rebuttal`
A paragraph that answers, defeats, or undercuts an opposing position. Prefer `rebuttal` over `support` *only* when the paragraph engages a specific opposing claim that the essay frames as the *other side* of the debate (whether named explicitly or implicit as the position the essay is pushing back against). **Critiquing a theoretical alternative or competing framework as part of developing the author's own affirmative case is `support`, not `rebuttal` — even when that critique involves direct contrast or pointing out flaws.** The interlocutor test: is the paragraph responding to a position the essay treats as a counter-argument it must defeat, or is it building the author's case by contrast with alternatives? Only the former is rebuttal.

Examples:
- A paragraph showing why the critics' worry, while reasonable on its face, doesn't survive scrutiny.
- A paragraph that systematically dismantles a proposed objection.
- A paragraph that opens by characterizing an opposing claim (e.g., "The argument that X...") and then provides evidence or reasoning showing why it doesn't hold — no overt "but" needed.
- A paragraph that takes a current framework or assumption and argues it is misguided, devalues something, or produces harmful outcomes.

`concession`
A paragraph that acknowledges an opposing concern, criticism, or alternative view has real force or partial truth. **Frequently appears as a *secondary* label alongside another move — e.g., opening paragraphs that concede a point before

stating the thesis (`[concession, thesis]`), or paragraphs that concede before refuting (`[concession, rebuttal]`). When a paragraph has a "while X is true / however understandable Y is, [main move]" structure, the concession should be co-labeled even when the main move dominates.**

Examples:
- "To be sure, this reform would impose real short-term costs on rural hospitals."
- "The concern that stricter standards could slow innovation is a serious one."
- An opening paragraph: "However understandable the emotional reflex may be, charging pre-teens as adults is fundamentally unjust." → label `[concession, thesis]`.
- A mid-essay paragraph: "While the proposal has real merit, it ignores X." → label `[concession, rebuttal]`.

`counterclaim`
A paragraph that presents an opposing position as its main object of attention, without taking a clear stance against it within the paragraph itself. **This is not a rare label. Use `counterclaim` whenever the paragraph foregrounds an opposing view or dynamic — even when the author later (in a different paragraph) argues against it. If you find yourself labeling a paragraph `support` because the opposing material "supports the overall essay's argument by being later refuted," that's wrong — that paragraph is `counterclaim`.** `counterclaim` should not co-occur with `rebuttal` or `concession` (those involve taking a stance within the paragraph).

Examples:
- A paragraph beginning "Critics of this policy argue that it would lead to..." and continuing to develop the concern in good faith.
- A paragraph faithfully representing why others believe the opposite of the writer's position.
- A paragraph that presents an opposing economic dynamic neutrally (e.g., as Marx-aligned framing) to set up later argument, without yet pushing back.

`implication`
A paragraph that draws out what follows from claims already made — consequences, stakes, projections, or takeaways. If a single statement *is* both the essay's central position AND a consequence claim, label that statement `thesis`, not `implication`. The two labels can co-occur on the same paragraph when distinct implications appear alongside the thesis.

Examples:
- A paragraph spelling out what is likely to happen if current trends continue, drawing predictions from the prior analysis.
- A paragraph stating "this means that..." and developing the inferential consequences for several lines.
- A closing paragraph that recapitulates the central position and projects the consequences of acting (or failing to act) on it. In those cases prefer `implication` alone rather than co-labeling with `support`.

`proposal`
A paragraph that advocates a specific course of action — what should be done, by whom, or under what conditions. If a single statement *is* both the essay's central position AND a recommendation, label that statement `thesis`, not `proposal`. The two labels can co-occur on the same paragraph when distinct proposals appear alongside the thesis.

Examples:
- A paragraph laying out what cities should do to address housing shortages.
- A paragraph closing with a clear directive about what action is needed.
- A closing paragraph that recapitulates the central position and pivots to "therefore, we must...". In those cases prefer `proposal` alone rather than co-labeling with `support`.

`none`
Use `none` when no argumentative role applies — the paragraph is purely transitional, mechanical, or otherwise carries no argumentative move. Mutually exclusive with all other labels.

Return JSON

```json
{
  "layer": "argument",
  "annotations": [
    {
      "paragraph_index": 0,
      "labels": ["<thesis | support | reframing | counterclaim | rebuttal | concession | implication | proposal | none>"],
      "rationale": "Brief explanation."
    }
  ]
}
```

Rules for output:

- Preserve paragraph order.
- Use one or two labels per paragraph. Use `none` when no label applies; `none` is mutually exclusive with all other labels.
- **When a paragraph carries two labels, list them in the order they appear in the paragraph's argumentative flow.** If the paragraph opens with one move and pivots to another, list the opening move first — e.g., a paragraph that concedes a point before stating a thesis is `["concession", "thesis"]`; a paragraph that argues a thesis then briefly acknowledges a counter-concern is `["thesis", "concession"]`. If both labels are interleaved throughout the paragraph without a clear sequential order, list the dominant move first.
- Return valid JSON only.

Discourse-mode annotation

You are performing discourse mode analysis of an op-ed essay. For each paragraph, assign discourse_mode labels based on HOW the paragraph is written — its dominant mode of presentation — independent of the argument it makes.

Per-paragraph annotation

Assign exactly one label per paragraph — the single mode that most dominates the paragraph's delivery. If two modes coexist, pick the one that most determines how the paragraph reads as a whole.

Label set

`argumentation`
Claim-and-reason writing with explicit inferential force. The paragraph stakes a position and develops it through reasons, contrasts, causal links, or conclusions. Use `argumentation` when the paragraph's internal organization is driven by reasoning from one sentence to the next, not just by presenting relevant information. The key is whether the paragraph is performing inference (claims, projections, evaluations, conclusions, contrasts) rather than reporting (facts, scenes, processes, states). Explicit reasoning markers like "because" or "therefore" are common in argumentation but not required.

The diagnostic is internal to the paragraph: does the paragraph itself perform an inferential move — make a claim, draw a conclusion, evaluate, project, or contrast — within its own boundaries? If yes → `argumentation`. If the paragraph reports facts, explains a state of affairs, or recounts events without itself drawing a conclusion → `exposition` or `narration`, even when those facts are positioned to support an argument made elsewhere in the essay.

Examples:
- A paragraph that states a policy is flawed, then explains why the flaw matters.
- A paragraph that uses facts to build an explicit conclusion inside the paragraph.
- A paragraph that projects a chain of consequences from a premise, even without overt "because/therefore" markers.
- A paragraph that recounts a historical sequence and ends with an evaluative judgment ("X has proven insufficient", "this approach failed") — the closing judgment is the inferential move that makes it argumentation.

`exposition`
Explanation, clarification, or factual information without dominant inferential force. The paragraph informs about facts, concepts, processes, or states. Sentences are connected through descriptive or elaborative relations rather than argumentative ones — even when the content supports a broader argument elsewhere in the essay. **A paragraph that lists facts, describes a causal dynamic, or recounts background without drawing its own conclusion is `exposition`, even when its role within the essay is to support an argument. The test is internal to the paragraph: no inferential move inside it → `exposition`.**

Examples:
- A paragraph explaining how a program, institution, or process works.
- A paragraph giving background facts without pushing an explicit internal conclusion.
- A paragraph clarifying a concept (e.g., defining a technical term, describing a regulatory framework) without using it to evaluate or argue.
- A paragraph describing a causal dynamic (X led to Y led to Z) where the chain is presented as a factual description of what happened, not as the foundation of a new claim.

`narration`
A paragraph that recounts events or actions — what happened, who did what, or how events unfolded, whether as a single incident or as a sequence over time. **Reserved for paragraphs whose primary act is recounting events for their own sake. A historical or chronological recap that ends in (or organizes itself around) an evaluative claim — e.g., "this approach has failed," "the result was X" — is `argumentation`, not `narration`. The surface chronology doesn't determine the mode; the organizing purpose does.**

Examples:
- A paragraph describing what happened in a particular incident or case.
- A paragraph recounting a sequence of events leading up to a decision, without drawing an evaluative conclusion from them.

`description`
Depiction of a scene, state, person, object, or place. The paragraph is organized spatially, perceptually, or by attributes rather than temporally or inferentially. The organizing principle is the qualities of what is being depicted.

Examples:
- A paragraph depicting conditions in a neighborhood, classroom, workplace, or institution.
- A paragraph portraying a person, group, place, or state of affairs without a main sequence or inference.

Return JSON

```json
{
  "layer": "discourse_mode",
  "annotations": [
    {
      "paragraph_index": 0,
      "label": "<argumentation | exposition | narration | description>",
      "rationale": "Brief explanation."
    }
  ]
}
```

Rules for output:

- Preserve paragraph order.
- Assign exactly one label per paragraph.
- Return valid JSON only.