

The CreateDebate Arguments Dataset: A Large-Scale Naturalistic Corpus for Computational Argumentation

Antonios Charalampous and Constantinos Djouvas

Department of Communication and Internet Studies

Cyprus University of Technology

ag.charalampous@edu.cut.ac.cy, constantinos.djouvas@cut.ac.cy

Abstract

We introduce the CreateDebate Arguments Dataset, a large-scale, unfiltered collection of 383,220 arguments from 7,435 of the most active for/against debates on the CreateDebate online debate platform, spanning 16 years of activity. Unlike existing computational argumentation resources, the dataset combines large scale with hierarchical debate-tree structure, author-assigned stance labels, community vote scores, and relational tags that explicitly mark when one argument disputes another. Because the full dataset is unfiltered, we release it under a gated access agreement and additionally release an openly available, manually curated subset of 84,872 arguments from 578 debates. To demonstrate the dataset’s utility, we define two counter-argument retrieval benchmarks on the curated subset: a *targeted* setting that uses the dispute tag as author-verified ground truth, and a *general* setting in which any opposing-side argument is a valid counter. Evaluating a BM25 baseline, a frozen bi-encoder, and a fine-tuned bi-encoder, we find consistent but bounded gains from fine-tuning on disputed pairs, a negative effect of debate-level context on within-debate retrieval, and limited cross-task transfer between the targeted and general settings. A qualitative analysis surfaces failure modes specific to naturally occurring debate. The dataset and models are released to the community.

1 Introduction

The rise of online debate platforms, social media, and digital news has made argumentation an increasingly central phenomenon in public discourse. As a result, the automatic analysis of argumentative content, broadly referred to as *computational argumentation*, has attracted considerable attention in the natural language processing community (Stab and Gurevych, 2014; Lippi and Torroni, 2016). Research in this area spans a wide range of tasks, from

identifying argumentative components within documents to assessing argument quality and detecting stance.

Two broad paradigms have emerged. *Discourse-level argument mining* focuses on identifying the internal argumentative structure of a document and recovering the inferential relations between its parts (Stab and Gurevych, 2014). *Information-seeking argument mining* takes a wider view: rather than analyzing a single document in depth, it aims to identify self-contained argumentative statements relevant to a given topic or query from large, heterogeneous text collections (Trautmann et al., 2020). In an information landscape saturated with one-sided content, there is a growing need for tools that surface diverse argumentative perspectives on controversial topics on demand (Levy et al., 2018; Ajjour et al., 2019). Existing resources each make different trade-offs between scale, structural richness, naturalness, and metadata granularity. We position our dataset as an additional point in this trade-off space: a large-scale collection of naturally occurring debate threads with rich metadata supporting a wide range of computational argumentation tasks.

We introduce the **CreateDebate Arguments Dataset**, a large-scale collection of **383,220 arguments** across **7,435 debates** collected from CreateDebate, an online platform where users post arguments, reply to one another, and vote on argument quality. Arguments are organized in hierarchical debate trees and carry author-assigned stance labels, community vote scores, and relational tags that explicitly mark when one argument disputes another. This combination of scale, structural richness, and naturally occurring metadata makes the dataset suitable for stance detection, argument quality assessment, discourse structure analysis, and counter-argument retrieval. The collection targets the most active debates of a single format on the platform rather than its entire content (Section 3).

Active since 2008, CreateDebate combined an open debate community with a classroom mode in which instructors could set up private debate communities for their students, and it was promoted as a venue for practising persuasive writing; particularly in its early years, part of its user base therefore plausibly consisted of students and educators. Its debates have supported research on stance and reason classification (Hasan and Ng, 2014), dialogic argumentation (Abbott et al., 2016), and explainable stance detection (Saha et al., 2024). Because the platform is no longer accessible, a large collection of its debates also has documentary value: spanning 16 years, it makes it possible to study how participation, topics, and argumentative register evolved on a single platform, including shifts such as the growth of trolling and low-effort contributions.

Because the dataset is unfiltered, it necessarily contains debates on extremely sensitive and potentially disturbing topics, alongside the large majority of debates that are not. We therefore release the full dataset under a gated access agreement restricting its use to research purposes, and additionally release a smaller, manually curated subset of **84,872 arguments** across **578 debates**, selected for topical coherence and quality. Details of the access procedure and curation methodology appear in Section 3.

To demonstrate the dataset’s utility, we focus on *counter-argument retrieval*: given an argument advancing a position, retrieve from a corpus the arguments that most effectively oppose it (Wachsmuth et al., 2018; Orbach et al., 2020; Jo et al., 2021). We conduct this demonstration on the curated subset, since the task does not benefit methodologically from the most sensitive content. The dataset’s *disputed* tag provides a naturally occurring, author-verified supervision signal directly suited to this task. We introduce two retrieval benchmarks: a *targeted* setting that uses the disputed tag as strict, author-verified ground truth, and a *general* setting in which any opposing-side argument is a valid counter. We evaluate three systems: a BM25 lexical baseline, a frozen pre-trained bi-encoder, and a fine-tuned bi-encoder trained with contrastive learning, and report findings on the role of debate context, hard-negative mining, and the relationship between the two settings.

Our contributions are:

- We release a large, unfiltered collection of

383,220 arguments from 7,435 debates under a gated access agreement, with stance labels, relational tags, community vote scores, and hierarchical tree structure¹.

- We release a manually curated subset of 84,872 arguments from 578 debates, suitable for open downstream use.²
- We demonstrate the dataset on two counter-argument retrieval tasks and release the resulting models³.

2 Related Work

2.1 Computational Argumentation Datasets

The development of computational argumentation has been closely tied to the availability of dedicated corpora. The Internet Argument Corpus 2.0 (Abbott et al., 2016) aggregates posts from three debate platforms (4forums.com, Convinceme.net, and CreateDebate.com) with annotations for stance, agreement, and dialogue structure. Its CreateDebate subset comprises 2,958 posts and lacks the fine-grained relational metadata, such as explicit dispute relations and hierarchical tree structure, that characterizes platform-native debate threads. CreateDebate has also been a recurring source for stance research. Hasan and Ng (2014) annotate ideological debate posts from four domains with the reasons behind each author’s stance and jointly model stance and reason classification, while Saha et al. (2024) crawl CreateDebate, together with Room For Debate, to build claim–article benchmarks for explainable stance detection (about 66,000 samples across the two sources). These resources are organized around individual posts or claim–document pairs for stance classification, rather than around the reply structure and relational tags between arguments. Our dataset draws substantially more content from the same platform while preserving its full structural richness.

The args.me corpus (Ajjour et al., 2019) harvests 387,606 arguments from four debate portals using distant supervision over explicit pro/con structure, but its arguments are drawn from pre-structured schemas with predefined stances, limiting naturalness and relational diversity. The accompanying

¹<https://huggingface.co/datasets/azza1625/createdebate-arguments-full>

²<https://huggingface.co/datasets/azza1625/createdebate-arguments>

³<https://huggingface.co/azza1625/counter-argument-retrieval>

args.me search engine (Wachsmuth et al., 2017b) and TARGER (Chernodub et al., 2019) demonstrate the value of such corpora for argument retrieval over web-scale collections. Our dataset complements these efforts with a resource grounded in naturally occurring debate interactions and richer relational structure.

IBM’s Project Debater program produced several influential resources (Levy et al., 2018; Slonim et al., 2021). Most relevant is the dataset of Orbach et al. (2020), comprising over 3,600 professionally recorded debate speeches annotated for counter-speech relations; it is restricted to formal speeches by expert debaters. The ArguAna Counterargs corpus (Wachsmuth et al., 2018), derived from idebate.org, provides 6,753 argument-counterargument pairs across 1,069 debates and is a key counter-argument retrieval benchmark, but its scale and reliance on a single structured portal constrain stylistic and topical diversity. Counter-argument generation work has frequently used the ChangeMyView dataset (Jo et al., 2021; Alshomary et al., 2021; Alshomary and Wachsmuth, 2023), which preserves threaded structure but provides no explicit debate sides or dispute tags. A related line studies argument quality (Wachsmuth et al., 2017a; Habernal and Gurevych, 2016; Gretz et al., 2020), relevant to ours given the community vote scores in our dataset.

2.2 Argument Retrieval and Generation

Work beyond classification follows two complementary directions: retrieval of relevant arguments and generation of novel argumentative content, both situated within the information-seeking paradigm (Trautmann et al., 2020). Systems such as ArgumenText (Stab et al., 2018; Si et al., 2023) show that identifying argumentative content from web-scale corpora is feasible and that topic information improves retrieval. On the retrieval side, Wachsmuth et al. (2018) model the similarity and dissimilarity of argument pairs to find the best counterargument, defining eight retrieval tasks of varying complexity over ArguAna. Orbach et al. (2020) frame counter-argumentation as detection (identifying which opposing speech most directly counters a given one), and find that expert humans substantially outperform automatic methods. Jo et al. (2021) retrieve counter-evidence to specific statements using a knowledge-enhanced NLI model, while Le et al. (2018) build retrieval- and generation-based argumentative dialogue agents on

the Internet Argument Corpus. On the generation side, CANDELA (Hua et al., 2019) combines retrieval with a two-step neural generator; Alshomary et al. (2021) attack weak premises to undermine arguments; and Alshomary and Wachsmuth (2023) jointly model conclusion and counter-argument generation with a stance-based ranking component.

Prior datasets occupy different points in a trade-off space defined by scale, structural richness, naturalness, and metadata granularity. The CreateDebate Arguments Dataset complements these resources by offering substantial scale, naturally occurring hierarchical conversational structure, and author-tagged relational metadata between argument pairs within a single platform; to our knowledge, it is the largest public CreateDebate corpus that preserves the full reply tree together with the author-assigned relational tags.

3 The CreateDebate Arguments Dataset

We begin by introducing the CreateDebate platform and the two-phase collection procedure, then describe the curated subset and a descriptive analysis of the dataset’s structural characteristics.

3.1 Platform and Data Collection

The CreateDebate Platform. CreateDebate lets users initiate debates on an unrestricted range of issues. A debate begins with a user-submitted proposition, optionally with a contextual description, and takes one of three forms: open-ended, for/against, or multiple-sides. Discussions unfold as hierarchical trees: arguments receive replies, forming multi-level conversational structures of arbitrary depth. Replies are categorized into three interaction types, *Support*, *Dispute*, and *Clarify*, providing explicit indicators of argumentative relations between posts. The platform also supports up/downvoting of individual arguments, with scores displayed alongside the text, and records debate-level metadata such as total and side-specific scores. These properties make CreateDebate a rich source for argumentative datasets (Levy et al., 2014; Ajjour et al., 2019; Habernal and Gurevych, 2016).

Phase One: Debate Discovery. Collection used automated crawlers, with Selenium driving the dynamic browser interactions required to navigate the platform. The first phase identified the set of debates to extract. We set the topics filter to include all categories, surfacing debates from politics and religion to sports, technology, and entertainment,

and sorted debates in descending order by argument count. We restricted collection to the *for/against* format, which provides a more structured argumentative setting than open-ended or multi-sided debates, and included both open and closed debates. The crawler iterated through the first 100 listing pages, yielding **7,435 unique debates** after deduplication and removal of inaccessible debates. The collection is therefore not an exhaustive crawl of CreateDebate: it covers the most active for/against debates reachable through these listing pages, and excludes open-ended and multi-sided debates as well as less active ones. Because the platform is no longer accessible, the total number of debates it hosted cannot be established, so we do not report a coverage ratio. Throughout the paper, *full dataset* refers to this unfiltered collection, as opposed to the curated subset. Basic metadata (URL, declared winning position when available, and platform-assigned topic) was recorded at this stage.

Phase Two: Argument Extraction. The second phase visited every debate URL to retrieve arguments and metadata. For each argument node we extracted the text, timestamp, user-declared stance, community vote score, depth in the tree, and the relational tag of any reply (*Support*, *Dispute*, or *Clarify*). At the debate level we recorded the title, full description, and aggregate statistics. Extraction yielded **383,220 arguments** across the 7,435 debates. The structure of the resulting debate- and argument-level objects is illustrated in Appendix A (Tables 6 and 7).

3.2 The Curated Subset

Alongside the full dataset we release a manually curated subset of 84,872 arguments across 578 debates, intended for open downstream use without access restrictions.

Selection Methodology. Because CreateDebate is user-generated and minimally moderated, the full dataset includes debates on inappropriate or sensitive topics and near-duplicate debates. The curated subset was produced by manual review at the debate level, carried out by the two authors by examining debate titles. The review had two main goals: filtering out NSFW debates, and making the retained debates as topically varied as possible. In practice, a debate was excluded if (i) its title indicated sexual, graphic, or otherwise NSFW content; (ii) it posed essentially the same proposition

Table 1: Corpus-level statistics of the full dataset and the curated subset. [†]Mean (median). [‡]Mean (maximum).

Statistic	Full	Curated
Debates	7,435	578
Arguments	383,220	84,872
Unique users	26,350	10,869
Vocabulary size	459,726	164,673
Arguments per debate [†]	51.5 (35.0)	146.8 (117.5)
Argument length (words) [†]	77.3 (38.0)	92.6 (45.0)
Tree depth [‡]	4.27 (109)	5.28 (109)
Avg. branching factor	1.28	1.32
Replies (% of arguments)	72.7	73.3
Disputed (% of replies)	70.0	77.5
Balanced debates (%)	56.3	69.8
Date range	Feb 2008 – Mar 2024	

as a debate already selected, which removed near-duplicates and served the variety goal; or (iii) its title did not pose a coherent, debatable proposition. Near-duplicates were identified by manually comparing titles rather than by an automatic similarity threshold. During the same pass, the reviewers assigned topic keywords to each retained debate to organize the subset and support topic-based filtering. Rather than imposing a post-level definition of a valid argument, the subset captures online debate as it actually unfolds, making it more representative of real-world conditions while introducing challenges absent from more controlled corpora.

Additional Annotations. Besides the topic keywords, we computed two structural fields for each curated debate. The *depth* field gives each argument’s position in the tree (0 for top-level), and the *context* field stores the parent argument’s full text for replies. These fields allow the full tree structure of any debate to be reconstructed from the flat representation, and are also present in the full release.

3.3 Dataset Statistics and Analysis

Table 1 summarizes both releases; we first describe the full dataset. It spans 16 years of activity (February 2008–March 2024), comprising 383,220 arguments from 26,350 users across 7,435 debates, with an average of 51.5 arguments per debate (median 35.0) and a vocabulary exceeding 459,000 tokens. Notably, 72.7% of all arguments are replies rather than top-level posts, underscoring the strongly conversational nature of the corpus.

Figure 1 characterizes the full dataset along four dimensions, and Figure 2 (Appendix B) adds two more; panels are labelled (a)–(f) consistently across

both figures.

(a) Argument length. The distribution is strongly right-skewed (mean 77.3 words, median 38.0): 79.8% of arguments contain fewer than 100 words and 92.2% fewer than 200, confirming that contributions are predominantly concise, characteristic of web-based discourse rather than formal debate, and fall comfortably within standard transformer input limits.

(b) Argument depth. The first three depth levels account for 55.1% of the corpus, but the dataset nonetheless reaches a mean depth of 4.27 (90% within depth 11, maximum 109), and the average branching factor of 1.28 indicates debates structured around elongated chains rather than broad, shallow trees.

(c) Arguments per debate. The number of arguments per debate is right-skewed (mean 51.5, median 35.0); 31.6% of debates exceed 50 arguments and 10.1% exceed 100.

(d) Relational tags by depth. The Disputed tag dominates throughout, accounting for 70.0% of all replies overall and rising from 66.1% at depth 1 to 71–74% from depth 3 onwards, indicating that interactions become more confrontational beyond the first reply levels. The Disputed tag reflects the reply author’s intent to challenge rather than a guarantee of argumentative substance: in its natural form, some disputed replies are brief reactions or off-topic remarks rather than substantive counters. This property directly shapes the targeted retrieval benchmark in Section 4, where disputed pairs serve as author-verified ground truth.

(e) Chain length and (f) side balance. Chain length, the number of exchanges from root to a given argument (i.e., depth + 1), is near-uniformly distributed across top-level (27.3%), short (27.8%), medium (20.7%), and deep (24.2%) bands, with over 72% of arguments embedded in exchanges of length two or more (mean 5.27, maximum 110), confirming that the dataset is genuinely dialogic rather than a collection of independent statements. Side balance, the fraction of arguments from the more active side, has mean 0.609 and median 0.588: 56.3% of debates are balanced (ratio 0.5–0.6), 29.2% moderately imbalanced, and 14.5% strongly imbalanced, so most debates are reasonably contested.

Table 2: Computational argumentation tasks supported by the dataset’s fields. † = task demonstrated in this paper. Multiple fields are separated by semicolons.

Task	Field(s)	Description
Stance detection	Side; Body	Predict argument’s side
Argument quality	Points	Rank by vote score
Relation classif.	Tag	Classify reply relation
Counter-arg. retr.†	Tag (Disputed); Side	Retrieve opposing counter
Dialogical mining	context; depth	Model multi-turn exchange
Outcome prediction	Total Score; Side Points	Predict winning side

Comparison with the Curated Subset. Table 1 also contrasts the full dataset and the curated subset. The two are similar in shape but differ in scale and a few properties: the curated subset has larger debates (146.8 vs. 51.5 arguments per debate), a higher disputational rate (77.5% vs. 70.0% of all replies), and tighter side balance (69.8% vs. 56.3% balanced debates), reflecting the selection bias toward larger, more substantive, contested debates. The share of replies and the branching factor are nearly identical across the two releases (73.3% vs. 72.7%; 1.32 vs. 1.28), while curated debates are somewhat deeper (mean depth 5.28 vs. 4.27) and their arguments slightly longer (median 45 vs. 38 words).

3.4 Supported Tasks

The combination of fields in Table 2 supports a range of computational argumentation tasks beyond the counter-argument retrieval benchmarks demonstrated in Section 4. The tasks selected serve to highlight how a few of the dataset’s distinguishing fields such as argument and user vote scores, relational dispute tags, and dialogical structure, each correspond to a different computational argumentation task. Many other tasks (e.g. argument retrieval, evidence detection, conclusion generation) could also be formulated over this data.

4 Counter-Argument Retrieval

We present counter-argument retrieval as a demonstration task, conducted on the curated subset. The curated subset suffices because the task does not benefit methodologically from the most sensitive content, allowing us to release trained models and evaluation data without restrictions, and because it is enriched for larger, more substantive debates

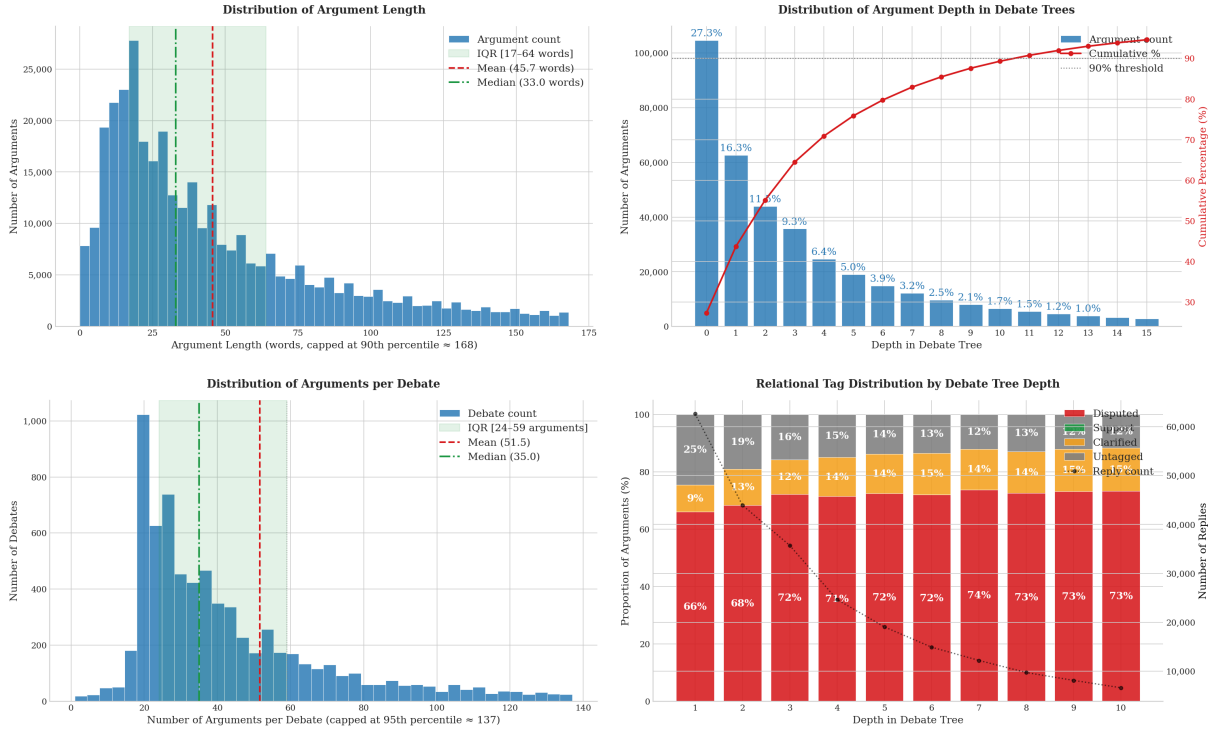


Figure 1: Structural statistics of the full CreateDebate Arguments Dataset: (a) argument length, (b) argument depth in debate trees, (c) arguments per debate, and (d) relational tag proportions by depth. Panels (e) chain length and (f) side balance appear in Figure 2 (Appendix B). In (a) and (c), the x-axis is capped for display, and the mean and median lines in (a) are computed over the displayed range.

that yield candidate pools of sufficient size and topical coherence. Among the tasks in Table 2, we select counter-argument retrieval because it most directly exercises the dataset’s most distinctive field, the author-assigned Disputed tag, which side-label-only corpora cannot provide. The experiments address three questions: (RQ1) does fine-tuning on disputed pairs improve targeted retrieval over lexical and off-the-shelf semantic baselines; (RQ2) does debate-level context help retrieval within a debate; and (RQ3) do targeted and general retrieval share a transferable signal?

4.1 Task Formulation

Given a query argument q advancing a position, counter-argument retrieval retrieves from a candidate pool $\mathcal{C}$ the arguments that most effectively oppose it. The dataset supports two instantiations differing in how ground truth is defined.

Scenario 1 – Targeted Retrieval. This setting uses the *disputed* tag for author-verified ground truth. For each disputed reply we construct an instance whose query is the parent argument, whose positive is the disputed reply itself (which targets the opposing side in 95.1% of cases), and whose

negatives are all other opposing-side arguments within the same debate. The model must identify the specific argument authored as a direct rebuttal to the query from among all opposing-side arguments. After filtering same-side disputed pairs (4.9%), this yields 44,569 query–positive pairs.

Scenario 2 – General Retrieval. This setting treats any opposing-side argument as a valid counter. For each argument, the query is any argument from side A, the positives are all arguments from side B, and the negatives are all other side-A arguments in the same debate. It tests opposition detection at the debate level rather than the claim level. For debates with more than two sides we keep the two dominant sides (dropping 2,466 fringe arguments, 2.9%). After a per-debate cap of 500 pairs, this yields 80,885 query instances.

Data Splits. Both scenarios share a debate-level 75/10/15 train/validation/test split, so all arguments from a debate fall in a single partition, preventing topic-specific leakage, analogous to the train–test overlap documented for open-domain QA by Lewis et al. (2021). Not all debates contribute to both scenarios. Per-split instance counts appear in Appendix C. The average negative pool of roughly

94–120 arguments per query in Scenario 1 makes the task non-trivial, as all candidates are drawn from the same debate context.

4.2 Models

We evaluate three systems of increasing sophistication.

BM25. A lexical baseline (Robertson and Zaragoza, 2009) requiring no training and operating on surface term overlap, a long-standing strong IR baseline (Thakur et al., 2021). We use the `rank_bm25` implementation with default parameters and whitespace tokenization.

Frozen Bi-Encoder. A pre-trained sentence encoder without fine-tuning (Reimers and Gurevych, 2019). Query and candidates are encoded independently and ranked by cosine similarity. We use `all-mpnet-base-v2`, establishing how far off-the-shelf semantic representations suffice without domain adaptation.

Fine-Tuned Bi-Encoder. Our primary model fine-tunes the same backbone with MultipleNegativesRankingLoss (Henderson et al., 2017), a contrastive objective treating other in-batch positives as implicit negatives. We also train a variant with one explicit hard negative sampled from the opposing pool. Training uses batch size 64, learning rate 2×10^{-5} , and 3 epochs with 10% linear warmup; the best checkpoint is selected by MRR@10 on validation. We train separate models per scenario (four variants total) and evaluate each on both test sets, using the `sentence-transformers` library (Reimers and Gurevych, 2019).

4.3 Evaluation Protocol

For **Scenario 1**, each query has a single ground-truth positive, so we report MRR (primary) and Recall@ k for $k \in \{1, 5, 10\}$. For **Scenario 2**, each query has multiple valid positives, so we report Recall@ k for $k \in \{10, 20\}$ at larger k , since the positive pool (mean 83.3) makes low- k recall structurally bounded. Candidate corpora are encoded once per model and scored by dot product between normalized embeddings. For each model we evaluate two input configurations: argument text only, and text prepended with the debate title and description to test whether debate-level context aids retrieval.

Table 3: Scenario 1 (Targeted Retrieval). Best per column in **bold**. Fine-tuned models use argument text only.

Model	Train	MRR	R@1	R@5	R@10
<i>Baselines</i>					
BM25	w/ ctx	0.277	0.182	0.353	0.439
BM25	no ctx	0.362	0.249	0.471	0.562
Frozen	w/ ctx	0.170	0.084	0.224	0.324
Frozen	no ctx	0.381	0.253	0.517	0.631
<i>Fine-tuned on Scenario 1</i>					
Bi-enc.	In-batch	0.444	0.301	0.605	0.711
Bi-enc.	Hard neg.	0.444	0.300	0.603	0.710
<i>Fine-tuned on Scenario 2 (transfer)</i>					
Bi-enc.	In-batch	0.365	0.242	0.490	0.594
Bi-enc.	Hard neg.	0.363	0.240	0.491	0.596

Table 4: Scenario 2 (General Retrieval). All systems lie within 0.002 of each other, so no best system is marked.

Model	Train	R@10	R@20
<i>Baselines</i>			
BM25	—	0.079	0.159
Frozen	—	0.078	0.158
<i>Fine-tuned on S1 (transfer)</i>			
Bi-enc.	In-batch	0.077	0.157
Bi-enc.	Hard neg.	0.077	0.157
<i>Fine-tuned on S2</i>			
Bi-enc.	In-batch	0.077	0.157
Bi-enc.	Hard neg.	0.077	0.157

4.4 Results and Discussion

Tables 3 and 4 report results for all variants.

Effect of Debate Context. Adding debate context consistently degrades Scenario 1 performance: BM25 drops from 0.362 to 0.277 MRR and the frozen encoder from 0.381 to 0.170. Since all candidates come from the same debate, prepending the title introduces shared vocabulary that inflates similarity uniformly, reducing the discriminative signal. At inference time, debate context should therefore not be included when candidates are drawn from the same debate.

Scenario 1 Results. Fine-tuning on disputed pairs improves over the frozen encoder by 6.3 MRR and 8.0 R@10 points (MRR 0.444, R@10 0.711). These gains are more modest than is typical in bi-encoder fine-tuning (Thakur et al., 2021), likely for two reasons: all negatives are drawn from the same debate and are thus already topically similar, so the frozen encoder captures much of the needed discrimination; and the training data mixes substantive rebuttals with non-substantive exchanges, since the disputed tag captures intent rather than

quality. Hard-negative training yields no measurable improvement, as a batch size of 64 already provides 63 diverse implicit negatives per query.

Cross-Scenario Transfer. Training on Scenario 2 and evaluating on Scenario 1 falls below the frozen encoder baseline (MRR 0.363–0.365 vs. 0.381): optimizing for general side-opposition is mildly detrimental to targeted rebuttal retrieval. The two objectives partially compete; distinguishing opposing- from same-side arguments is a different task from identifying a specific rebuttal. Transfer in the opposite direction also provides no benefit.

Scenario 2 Results. Results are uniform across all models (differences below 0.002), with BM25 matching or marginally outperforming all neural models. This points to a fundamental difficulty with the general setting: with a mean positive pool of 83.3 and all candidates sharing the debate context, the textual signal to distinguish valid counters from same-side arguments appears insufficient for any evaluated approach. We regard these results as a lower bound on the difficulty of general counter-argument retrieval on this dataset.

Summary. Fine-tuning on author-verified disputed pairs yields consistent but bounded improvement for targeted retrieval; general retrieval as formulated is difficult for all models, with BM25 remaining competitive; and the two settings are distinct tasks with non-transferable signals, underscoring the value of the disputed tag as a supervision source enabling a qualitatively different setting than side-label-only corpora.

4.5 Qualitative Analysis

Aggregate metrics do not reveal the nature of individual successes and failures. A recurring theme is a tension inherent to the dataset: CreateDebate users produce the full spectrum of argumentative behavior, from substantive rebuttals to brief dismissals and off-topic responses. This naturalistic breadth distinguishes the dataset but also challenges retrieval models optimized for semantic similarity.

Outcome Distribution. Table 5 bins test queries by ground-truth rank. The model retrieves the correct counter at rank 1 for 29.8% of queries and within the top 5 for 60.4%; a further 20.7% are partial failures (rank 6–20) and 18.8% complete failures (rank >20, mean 48.8). The median ground-

Table 5: Test queries by ground-truth rank, with mean lexical (Jaccard) and semantic (cosine) similarity between query and the ground truth (gt) and top-1 retrieved argument (t1).

Bin	N	%	Lexical		Semantic	
			gt	t1	gt	t1
A: Rank 1	1,157	29.8	0.251	0.251	0.672	0.672
B: Rank 2–5	1,187	30.6	0.176	0.224	0.550	0.655
C: Rank 6–20	803	20.7	0.118	0.180	0.416	0.618
D: Rank >20	731	18.8	0.092	0.179	0.298	0.622

truth rank is 3, so near misses are more common than complete failures. The similarity statistics reveal the core failure mode: as retrieval quality degrades (Bins B–D), query–ground-truth semantic similarity drops sharply (to 0.298 in Bin D), yet similarity between query and the model’s top-1 choice stays high (0.655–0.672). The model confidently retrieves something semantically close to the query, just not the author-verified ground truth. Failures thus arise from a mismatch between semantic similarity and the pragmatic logic of the actual rebuttal.

Success and Failure Patterns. The model often succeeds despite minimal lexical overlap: across representative cases (Appendix D, Table 10) it identifies counters addressing the argumentative thrust of the query (lexical overlap 0.063–0.152, semantic similarity 0.776–0.835), confirming it operates on semantic rather than surface patterns. Failures cluster into two types. *Pragmatic and contextual rebuttals* (Table 11) engage with the framing or a specific embedded claim rather than the overall topic; these carry low or indistinguishable semantic similarity, and the single ground-truth assumption underestimates capability where multiple valid counters exist, so Scenario 1 scores are lower bounds. *Non-substantive exchanges* (Table 12) arise because any reply marked Disputed becomes ground truth regardless of substance; failing to retrieve a brief conversational reaction is arguably correct from a retrieval standpoint. These cases follow from how the data was collected, and the conversational noise in training partly explains the modest fine-tuning gains.

Error Patterns by Depth and Length. Performance at depth 0 is lower (MRR 0.334) than at deeper levels (0.416–0.497; Appendix D, Figure 3). Top-level arguments are broad opening claims plausibly countered by many opposing argu-

ments, whereas deeper exchanges are more specific and contextually grounded, narrowing the space of plausible counters. A modest positive trend also holds for query length (MRR 0.436 for 1–20 words, rising to 0.498 for 201–500 words): longer queries provide more signal, while short rhetorical remarks tend to have less substantive disputed replies.

5 Conclusion and Future Work

We introduced the CreateDebate Arguments Dataset, a large-scale, unfiltered collection of 383,220 arguments from 7,435 debates spanning 16 years, distinguished from existing resources by its combination of scale, hierarchical tree structure, author-assigned stance labels, community vote scores, and relational dispute tags. The full dataset is released under gated access reflecting its unfiltered character, accompanied by an openly available curated subset of 84,872 arguments from 578 debates. To demonstrate its utility we introduced two counter-argument retrieval benchmarks and evaluated three systems. Fine-tuning on disputed pairs yields consistent but bounded gains on targeted retrieval; adding debate context degrades within-debate retrieval; and targeted and general retrieval are distinct tasks with non-transferable signals. A qualitative analysis identified three failure modes specific to naturally occurring debate: pragmatic and meta-argumentative rebuttals below the resolution of semantic similarity, ambiguity in the single ground-truth assumption, and non-substantive exchanges carrying the disputed tag.

Several directions follow from these findings. Retrieval is inherently limited to indexed arguments, so generation-based methods (Hua et al., 2019; Alshomary et al., 2021; Alshomary and Wachsmuth, 2023) would allow novel counters. A stance classifier as a post-retrieval verification step (Alshomary and Wachsmuth, 2023) could ensure retrieved arguments genuinely oppose the query. Extending retrieval to cross-debate candidates, supported by the topic keywords, would better reflect real-world conditions. Automatic argument-level quality filtering could separate substantive arguments from conversational noise, and comparing filtered and unfiltered training would quantify the effect of naturalistic noise. Finally, relevance-graded annotation beyond single ground truth would enable fairer evaluation where multiple valid counters exist.

Limitations

The full dataset is not an exhaustive crawl of CreateDebate: it covers the most active for/against debates, so its statistics characterize that population rather than the platform as a whole. The curated subset reflects the judgment of two reviewers working from debate titles; we did not measure inter-annotator agreement, and the subset differs in distribution from the full dataset (Table 1). The counter-argument retrieval demonstration is conducted only on the curated subset, not the full dataset; while the two releases are structurally similar, results may not transfer directly to the full dataset’s broader topical range. Scenario 1 evaluation assumes a single author-verified ground truth per query, which underestimates model capability when several opposing arguments are equally valid counters, so the reported scores should be read as lower bounds. The disputed tag captures a reply author’s intent to challenge rather than argumentative substance, so the training and evaluation data contain a mixture of substantive rebuttals and non-substantive conversational exchanges; we do not isolate the contribution of this noise to the observed performance ceiling, and controlled ablations are left to future work. The general retrieval setting (Scenario 2) is formulated with all candidates drawn from the same debate, and our results establish only a lower bound on its difficulty rather than a characterisation of what is achievable with stronger architectures. We evaluate a single bi-encoder backbone (`all-mpnet-base-v2`) and do not explore cross-encoders, larger models, or generation-based approaches. Candidate pool sizes vary across depth levels, and we do not control for this in the depth-based error analysis. Finally, the dataset is English-only and reflects the user population and topical composition of a single platform, which may limit generalisation to other languages and argumentative settings.

Ethical Considerations

The data was collected from a public online debate platform via automated crawling. Because CreateDebate is user-generated and minimally moderated, the full archive necessarily contains debates on extremely sensitive, offensive, or potentially disturbing topics. To balance research utility against the risk of misuse or harm, we adopt a two-tier release: the full raw archive is released only under a gated access agreement restricting use to research

purposes, while the openly released subset is manually curated to exclude broad NSFW categories and (near-)duplicate debates. We deliberately retain all arguments within each selected curated debate including brief, personal, or off-topic replies, to preserve a representative picture of online debate, rather than imposing a post-level filter that could bias the resource. Argument bodies are released as written and may incidentally reference third parties or public figures; we remove user-identifying fields prior to release, and the platform itself is no longer accessible, reducing re-identification risk. The dataset is intended for research on computational argumentation and should not be used to build systems that manipulate, deceive, or harass, or to generate adversarial or inflammatory content. This concern extends to the released retrieval models, which could in principle be used to automate disagreement at scale (e.g., astroturfing or coordinated discourse manipulation); we release them for research on computational argumentation only. Models trained on naturally occurring debate may reproduce social biases and toxic language present in the source data, and any downstream deployment should incorporate appropriate safeguards.

References

- Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4445–4452.
- Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 48–59. Springer.
- Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827.
- Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967.
- Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.
- Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Asaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.
- Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1214–1223.
- Kazi Saidul Hasan and Vincent Ng. 2014. [Why are you taking this stance? identifying and classifying reasons in ideological debates](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.
- Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.
- Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672.
- Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3074–3094.
- Dieu-Thu Le, Cam-Tu Nguyen, and Kim-Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.
- Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical Papers*, pages 1489–1500.
- Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081.

- Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008.
- Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.
- Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the echo chamber: Detecting countering debate speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7073–7086.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.
- Rudra Ranajee Saha, Laks V. S. Lakshmanan, and Raymond T. Ng. 2024. [Stance detection with explanations](#). *Computational Linguistics*, 50(1):193–235.
- Jiasheng Si, Yingjie Zhu, Xingyu Shi, Deyu Zhou, and Yulan He. 2023. Explainable topic-enhanced argument mining from heterogeneous sources. *arXiv preprint arXiv:2307.12131*.
- Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, and 1 others. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.
- Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, pages 21–25.
- Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.
- Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. 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.
- Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.
- Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.

A Dataset Object Structure

Tables 6 and 7 illustrate the structure and metadata of debate- and argument-level objects.

Table 6: Structure and metadata of a debate object.

Field	Example Value
Title	animal testing
Topic	general
URL	createdebate.com/debate/show/animal_testing
Description	Is it okay to test on animals.
Total Score	195
Total Arguments	83
Total Votes	230
Side: For (Points)	98
Side: Against (Points)	97
Side: For (Tags)	36
Side: Against (Tags)	45

B Additional Dataset Statistics

Figure 2 shows the two structural dimensions of the full dataset discussed in Section 3 but omitted from Figure 1 for readability.

C Data Splits

Tables 8 and 9 report per-split instance counts for the two retrieval scenarios.

Conversational Structure of Debate Arguments

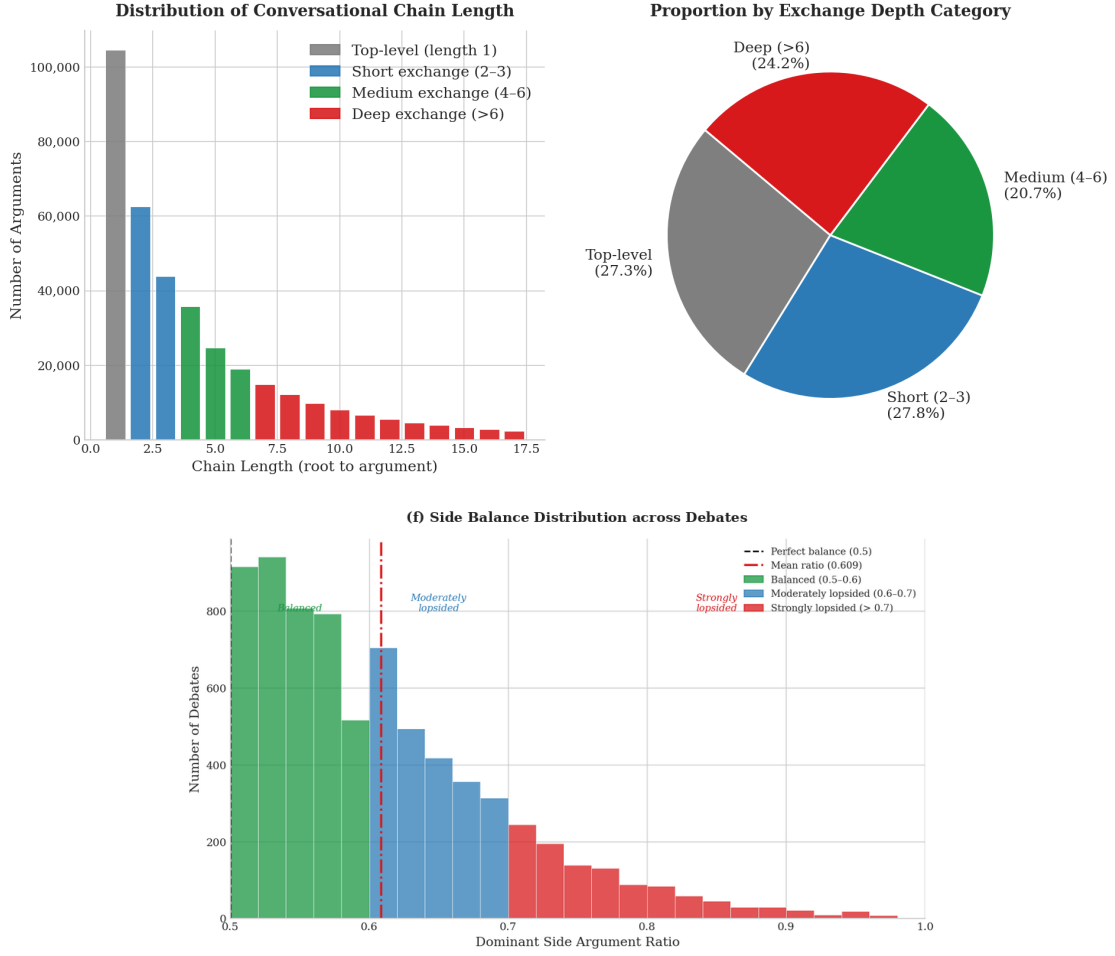


Figure 2: Further structural statistics of the full dataset, continuing the panel labels of Figure 1: (e) conversational chain length and (f) side balance.

Table 7: Structure and metadata of an argument object.

Field	Example Value
Argument ID	arg82671
Parent ID	-1
User Points	6780
Argument Side	For
Argument Points	4
Date	2010-01-15T07:43:03+00:00
Argument Tag	(Empty)
Argument Body	Given the choice between your dog or your mother (or child), who would you choose? I love my cat very much, but I would kill her in an instant to save the life of a loved one. Is animal testing any different?

Table 8: Debate-level data split for Scenario 1 (Targeted Retrieval).

Split	Debates	Queries	Avg. Neg.
Train	456	36,682	119.8
Validation	55	4,009	110.5
Test	56	3,878	93.7
Total	567	44,569	—

cal overlap; Tables 11 and 12 show the two failure types; and Figure 3 shows MRR by query depth and query length.

D Qualitative Examples and Error Analysis

This appendix provides the example tables and depth/length figures referenced in Section 4.5. Table 10 shows successful retrievals despite low lexi-

Table 9: Debate-level data split for Scenario 2 (General Retrieval).

Split	Deb.	Queries	Avg. Pos.	Avg. Neg.
Train	463	65,908	102.6	110.0
Validation	57	7,911	102.9	116.9
Test	57	7,066	83.3	88.7
Total	577	80,885	—	—

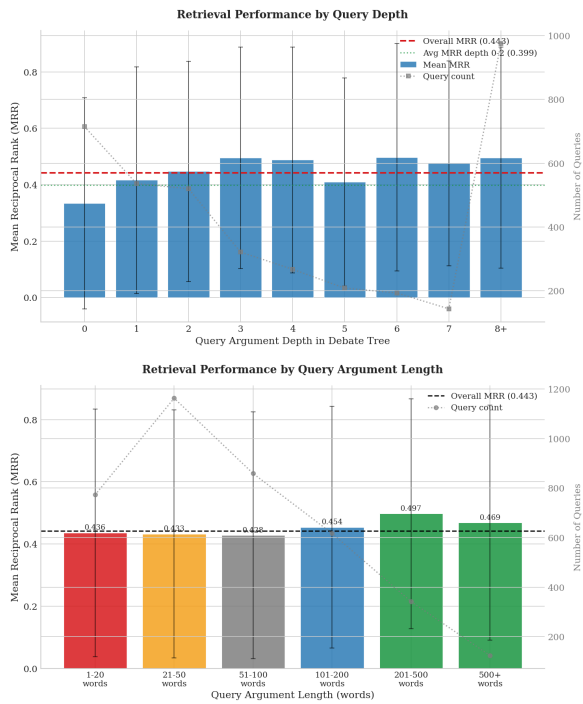


Figure 3: Mean MRR by query depth (top) and by query length in words (bottom) for the fine-tuned bi-encoder on Scenario 1.

Table 10: Success examples: correct counter-argument retrieved at rank 1 despite low lexical overlap.

Query	Retrieved Counter-Argument (Rank 1 = GT)
<i>Debate: Is Wal-Mart Good For America?</i> “Personally I would like to say no... but Walmart holds so many jobs right now for Americans.” (Lex: 0.082 Sem: 0.793)	“It may hold many jobs for Americans, but have you thought about how many jobs the company has taken away from employing companies overseas? Wal-Mart continuously squeezes its U.S. suppliers to cut their prices [...] the U.S. Bureau of Labor Statistics reported an estimated net loss of 170,000 manufacturing jobs in Ohio in 2005.”
<i>Debate: Should the death penalty be used as punishment for specific crimes?</i> “Yes they should, because if they get released they could murder someone again, that’s why they should have the death penalty.” (Lex: 0.063 Sem: 0.776)	“Just keep them in jail forever so they don’t do that again. Don’t keep releasing them again and again, just keep them in jail for life.”
<i>Debate: TV Violence</i> “TV is one of the biggest resources of information. With TV many people can learn the latest news that happened in the world no matter where they are.” (Lex: 0.152 Sem: 0.835)	“There is no argument that TV is useful; the question was ‘is it harmful’. TV offers lots of information, but is that info helpful or harmful? Yes, TV gives us the news, but is the news accurate?”

Table 11: Failure examples: pragmatic and contextual rebuttals the model fails to retrieve.

Query	Ground Truth (GT)	Top-1 Retrieved
<i>Debate: Do gods exist?</i> “[...] They did not feel a higher power, they just had no way to explain their own existence. If you feel a higher power, then I am of the opinion that you should be locked away in a dark, damp place.” (GT Rank: 2 Sem _{q,gt} : 0.619 Sem _{q,t1} : 0.620)	“Countless people have claimed they’ve felt a higher power, yet you say they did not feel a higher power [...] If they want to feel something, they feel it.”	“In the times before communicating with each other could be done easily, people all over the planet independently believed in a higher power [...] People still wonder about that higher power today.”
<i>Debate: Should the government ban the sale of guns?</i> “A guy in China almost did it. He severely injured almost 30 people in a school. Many were close to death.” (GT Rank: 2 Sem _{q,gt} : 0.459 Sem _{q,t1} : 0.459)	“You just made the best anti-gun argument possible against yourself. Guy in China stabs people, no one dies. Guy with gun in CT, 20 die. Get the difference?”	“No it’s the gun. Try killing someone with a finger.”

Table 12: Failure examples: non-substantive ground-truth responses.

Query	Ground Truth (GT)
<i>Debate: Is god true?</i> A 530-word argument concluding with “You wrote a paragraph? Wow, you have no life..” (GT Rank: 97)	“W.T.H is a paragraph? Seriously, from a fellow troll... at least have a basic grasp of the English language.”
<i>Debate: Is healthcare a right?</i> “This debate is about whether healthcare is a right or not. Stay on topic or leave.” (posted at depth 21)	“No thanks.” (GT Rank: 135)