Embedding Models for Stance-Aware Argument Retrieval

Angelo Sparacino[1], Francesca Toni[1] and Adam Dejl[1]

[1]Department of Computing, Imperial College London, UK

Abstract

In computational argumentation, obtaining arguments that explicitly support or attack given claims is a critical precursor to downstream reasoning tasks. When these supporting and attacking arguments are to be retrieved using semantic search methods, they need to be assessed for topic-relevance to the claims of interest as well as for correctness of their (positive or negative) stance towards the claims. In this paper we explore how dense *embedding models* (hereafter, models), powering modern retrieval pipelines, can serve as the basis of semantic search incorporating this dual assessment. We show experimentally that existing models struggle with asymmetric reasoning, exhibiting a strong bias toward topical overlap while ignoring instructional stance. We also show that correcting this bias via contrastive training triggers a new failure mode where models over-correct, over-fixating on polarity keywords (e.g., "supports" or "refutes") at the expense of the semantic topic. We thus introduce diagnostic word-ablation metrics to quantify this phenomenon and propose a data-centric solution. By implementing a balanced argument curriculum alongside LLM-augmented, stance-inverted arguments, we force the (embedding) models to learn deeper directional logic rather than exploiting superficial lexical shortcuts. Our evaluation demonstrates that, for sufficiently powerful models, this approach can alleviate the observed overcorrection, achieving further improvements in stance-aware argument retrieval.

Keywords

argument retrieval, stance-aware retrieval, embedding models, contrastive learning, data augmentation

1. Introduction

The modern digital landscape is associated with an unprecedented proliferation of data. To process this information, the field of computational argumentation relies heavily on automated argument mining pipelines [1]. A foundational prerequisite to these tasks is the accurate retrieval of evidence that explicitly supports or attacks a claim. While Large Language Models (LLMs) are often employed in Retrieval-Augmented Generation (RAG) [2] to ground their reasoning in external evidence, the efficacy of these pipelines is bottlenecked by their retrieval component.

Modern dense retrievers utilise a bi-encoder architecture to map textual sequences into a continuous, high-dimensional vector space [3], quantifying semantic relevance through spatial proximity. However, applying pre-trained embedding models (hereafter, models) to the asymmetric demands of (attacking/supporting) argument retrieval demonstrates that they struggle on this task. Indeed, while instruction-tuned embedding models [4, 5, 6] excel at fetching documents based on general subject matter, their internal mechanisms often default to superficial topical matching. These models heavily rely on semantic overlap while largely ignoring relational stance [7], as illustrated in Figure 1 (left). Here, given an instruction such as "Find statements rejecting" and a claim relating to some topic, models frequently retrieve arguments that perfectly match the topic but directly violate the instruction.

Having evaluated this problem experimentally (Section 4), we initially attempt to correct this behaviour through contrastive tuning, but observe that this may cause the embedding space to undergo a phenomenon we call 'topical collapse'. During optimisation, the model shifts its representational capacity away from the underlying subject matter and over-fixates on explicit polarity keywords.

To address this problem as well as the general limitations of embedding models on the stance-aware argument retrieval task, we investigate several targeted interventions aiming to achieve balance between topic- and stance-sensitivity (Section 5). The primary contributions of this work are as follows:

CMNA'26: 26th Workshop on Computational Models of Natural Argument, September 14–15, 2026, Barcelona

angelojsparacino@outlook.com (A. Sparacino); ft@imperial.ac.uk (F. Toni); adam.dejl18@imperial.ac.uk (A. Dejl)

0009-0005-0126-7836 (A. Sparacino); 0000-0001-8194-1459 (F. Toni); 0009-0006-0274-4160 (A. Dejl)

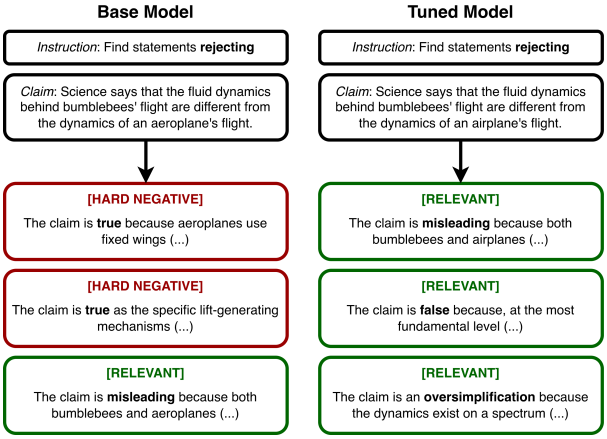

Figure 1: An illustration of the stance-aware argument retrieval problem addressed in this work. The Base Model (left) successfully matches on the semantic topic for all retrieved arguments but ignores the stance instruction ('rejecting'), resulting in the retrieval of two 'hard negatives'. Our Tuned Model (right) successfully overcomes this bias, retrieving topically relevant arguments with the correct opposing stance.

- We conduct an extensive experimental evaluation of embedding models' performance on the stance-aware argument retrieval task, including through novel diagnostic metrics of Relative Instruction/Claim Sensitivity (RIS/RCS) and Directional Impact (DI). The metrics quantify the effects of specific input text portions on the resulting similarity through word ablation, providing deeper insight about model focus and the topical collapse phenomenon.
- We introduce targeted interventions for stance-aware retrieval, aiming to improve the performance of embedding models on this task while mitigating topical collapse. This includes devising a refined fine-tuning curriculum as well as exploring hybrid search approaches.

Overall, our experiments show that while base embedding models perform poorly on stance-aware retrieval, their performance can be substantially improved through targeted fine-tuning. To enable reuse by the research community, our codebase, datasets and models are publicly available[1].

2. Related Work

Argument Mining Argument mining [8] aims to identify, in text, arguments, their components, and dialectical relations (notably of attack or support) among them. In this paper we focus on relation-based argument mining [9], specifically on retrieving arguments supporting or attacking given claims. Rather than using LLMs to perform this task, e.g., as in [10], we rely on embedding models.

Dense Retrieval and Foundation Models Modern retrieval systems rely on the bi-encoder architecture to semantically match queries to related documents. In this framework, a query and a document are processed independently by neural embedding models to generate high-dimensional, real-valued dense vectors. This allows relevance to be evaluated mathematically, typically via cosine similarity, and enables computationally efficient offline indexing [11].

However, static document embeddings introduce a "representational bottleneck" [12] because they must compress all possible semantic interpretations into a single vector. To bypass this limitation, instruction-tuned models like Instructor [4], and more recently, decoder-only foundation models [5], were developed. By prepending a natural language instruction to the query, these architectures condition the embedding space to prioritise specific semantic features based on explicit user intent.

To train these foundation models for retrieval tasks, the standard industry practice relies on contrastive learning methodologies such as Multiple Negatives Ranking Loss (MNRL) [13], an adaptation of InfoNCE

[1]Code: https://github.com/as9122/cmna26-stance-aware-text-retrieval, datasets and models: https://huggingface.co/collections/as9122/cmna26-stance-aware-text-retrieval.

[14]. During training, MNRL forces the model to maximise similarity between a query and a positive document, while simultaneously minimising similarity with a batch of negative documents.

Lexical Bias and Instruction-Aware Retrieval While the architectural transition to foundation models suggests an inherent capacity for complex reasoning, empirical stress testing reveals that models often regress to a "bag-of-words" heuristic. Rather than processing an instruction as a logical constraint, models treat it merely as a source of additional keywords to be matched [15]. Benchmarks such as FollowIR [16] and InstructIR [15] have quantified this weakness, demonstrating the sensitivity of retrieval models to phrasing and surface-level lexical patterns.

In the context of text retrieval, subsets of documents can also act as "hard negatives" [17]. For instruction-tuned models, these are often documents that share high topical overlap with the query but explicitly violate the instruction (e.g., a document refuting climate change when instructed to find supporting arguments). Because the embedding space is dominated by lexical overlap, the "negative" document often contains an equal or higher density of query terms than a true "positive" document, causing its embedding to be undesirably close to the query. A common approach to address this issue is hard-negative mining, where a retrieval model is intentionally exposed to hard-negative documents during training [3, 18, 19]. This strategy can be optionally combined with other data-centric methods such as balanced sampling [20] and LLM data augmentation [21, 22, 23, 24].

3. Task Formulation and Experimental Setup

In this section, we provide an overview of the considered task of stance-aware argument retrieval while also describing the general experimental setup and evaluation metrics. This background will be helpful for our later analysis evaluating the performance of state-of-the-art embedding models on stance-aware argument retrieval as well as our investigation of techniques to improve this performance.

3.1. Task Definition

We consider the task of stance-aware argument retrieval, a search problem where an embedding model is used to retrieve relevant documents (arguments) from a background corpus, taking into account both the semantic topic of a target claim and, crucially, directional stance. Such stance can amount to either *supporting* or *attacking* the claim, in line with the two relation types in bipolar argumentation frameworks [25]. In particular, we define an input query Q as the concatenation of an instruction I, specifying the stance constraint (e.g., "Find arguments supporting...") and a claim C, specifying the core subject matter, such that $Q = I \oplus C$. This query can then be encoded using an instruction-tuned embedding model $\mathscr{M}$ and compared against the embeddings of documents in the background corpus $\mathscr{D}$, with those most similar in terms of cosine similarity $\text{sim}_{\mathscr{M}}$[2] being retrieved. Here, $\text{sim}_{\mathscr{M}}$ is defined as:

$$\text{sim}_{\mathscr{M}}(Q, D) = \frac{\mathscr{M}(Q) \cdot \mathscr{M}(D)}{\|\mathscr{M}(Q)\| \, \|\mathscr{M}(D)\|}$$

Given the requirements of the task, any document $D \in \mathscr{D}$ falls into one of four categories:

- Positive (D_{pos}): Matches both the claim topic and the requested stance.
- Hard Negative (D_{hn}): Matches the claim topic but directly violates the requested stance.
- Semi-Hard Negative (D_{shn}): Does not match the claim topic but matches the requested stance.
- Easy Negative (D_{en}): Does not match the claim topic nor the requested stance.

In our experiments, we treat arguments associated with the same claim but opposite stance as hard negatives, arguments associated with different claims but matching the requested stance as semi-hard negatives, and arguments associated with different claims and opposite stance as easy negatives.

The primary challenge of stance-aware argument retrieval lies in distinguishing between D_{pos} and D_{hn}, as both share a high degree of lexical overlap with the underlying claim C.

[2] In the subsequent text, we will drop the lower-index $\mathscr{M}$ when not referring to a specific embedding model.

3.2. Datasets and Setup

All our experiments maintain a strict separation between training, validation and evaluation resources. For training, we utilised an 80% split of the TFU Training Arguments dataset [26], paired with 20 standardised instructions (10 for supporting and 10 for attacking arguments) to ensure generalisation across varied phrasing. The remaining 20% held-out split, alongside two TFU Validation Arguments datasets, generated by GPT-5 and Qwen3-8B, respectively [26], were used for initial in-domain evaluation. Finally, out-of-domain capabilities were evaluated using the TFU Evaluation Arguments dataset [26], the AVeriTeC Arguments dataset [27, 26], and ArgTumour, a small domain-specific medical dataset focused on glioblastoma treatments [28]. To better assess generalisation, we also used additional unseen instructions (5 supporting and 5 attacking) for validation and evaluation. All instructions are provided in Appendix B. Dataset statistics and further details on our training setup are given in Appendix C.

To power experiments involving contrastive tuning, a triplet generation engine produced training triplets consisting of a query, a positive document and a negative document. A custom batch sampler was also utilised to ensure all triplets within a single training batch featured distinct claims, eliminating the risk of false in-batch negatives. In all our fine-tuning experiments, we used Low-Rank Adaptation (LoRA) [29] with a rank of $r = 64$, reducing the computational costs and enabling us to perform a wider range of experiments. To assess the generalisation of the proposed methods, we considered three pre-trained embedding models: BGE-Large [6], Instructor-XL [4] and Qwen3-Embedding-8B [5].

3.3. Evaluation Metrics

We evaluate retrieval performance using **Precision@R** as our primary metric, which dynamically scales the maximum achievable score based on the total number of positive documents R available for a query, as well as **NDCG@10**, which applies a logarithmic discount to penalise poorly ranked results.

To explain the behaviour of the considered embedding models and diagnose the underlying cause of the retrieval failures, we employ a structured word ablation protocol. Let Δ_w denote the raw change in cosine similarity between the query Q and document D when a single word $w \in Q$[3] is ablated:

$$\Delta_w(Q, D) = \text{sim}(Q \setminus \{w\}, D) - \text{sim}(Q, D)$$

Building upon this, we define two diagnostic metrics:

- **Relative Instruction/Claim Sensitivity (RIS/RCS)**: This measures the relative sensitivity of the model-driven similarity measure to the instruction and claim components of the query:

$$\text{RIS}(Q, D) = \frac{\sum_{w \in I} |\Delta_w(Q, D)|}{\sum_{w \in Q} |\Delta_w(Q, D)| + \epsilon}, \quad \text{RCS}(Q, D) = \frac{\sum_{w \in C} |\Delta_w(Q, D)|}{\sum_{w \in Q} |\Delta_w(Q, D)| + \epsilon}$$

  High RIS values indicate that the model is predominantly focusing on the instruction specifying the desired argumentative stance, while high RCS values indicate that the model is mainly focused on the claim topic, with $RIS(Q, D) + RCS(Q, D) \approx 1$. A small constant ϵ prevents division by zero.
- **Directional Impact (DI)**: This averages the raw, signed Δ_w values strictly for the subset of stance keywords $\mathcal{S}_I$ (e.g., "supporting", "attacking", etc) to test the model's relational logic:

$$\text{DI}(Q, D) = \frac{1}{|\mathcal{S}_I|} \sum_{w \in \mathcal{S}_I} \Delta_w(Q, D)$$

  For a hard negative D_{hn}, removing the stance word causes the instruction to become stance-agnostic, which should typically result in a higher similarity score and $\text{DI} > 0$.

We note that these metrics are solely based on the input-output behaviour of the model during word ablation rather than its internal representations. While this makes the metrics more widely applicable, it also causes them to be potentially sensitive to syntactic changes after ablations.

[3] We assume that w carries its index so that Q, I, C and D are ordered and repeated words are distinct.

Figure 2: Average Precision@R for base and tuned models evaluated on unseen instructions. The base models exhibit severe instruction blindness, retrieving a high proportion of stance-violating hard negatives. While fine-tuning successfully reduces hard-negative retrievals, it also increases the retrieval of semi-hard negatives. This suggests that the used contrastive curriculum forces the models to overly focus on instructional stance at the expense of the claim topic, resulting in a phenomenon that we call topical collapse.

4. Baseline Model Evaluation

4.1. Baseline Performance

Given the considered stance-aware retrieval task, we first evaluate the performance of base instruction-following embedding models and their counterparts optimised using a baseline fine-tuning strategy. This strategy optimises the models using a homogeneous curriculum comprised of hard-negative triplets using MNRL. The results (in Figure 2) show substantial performance differences between the models:

- **BGE-Large**: Fine-tuning substantially improved BGE-Large's in-domain Precision@R on the held-out TFU Training data from 0.419 to 0.822, reducing stance error to 8.8%. However, it remained the lowest-performing architecture overall, exhibiting a pronounced spike in semi-hard negatives on the TFU Validation sets (constituting 27.1% of the retrieved items on GPT-5 data).
- **Instructor-XL**: Instructor-XL benefited from fine-tuning, increasing its Precision@R on the held-out TFU Training data partition from 0.531 to 0.839 and dropping its stance error[4] to 9.4%.
- **Qwen3-Embedding-8B**: Qwen3-Embedding-8B demonstrated the overall best performance. On the held-out training data partition, the fine-tuned model achieved a Precision@R of 0.95, reducing the originally high stance error down from 39.7% to 2%.

Overall, we find that none of the evaluated base embedding models perform stance-aware argument retrieval reliably, despite being instructed about the desired argument stance. While fine-tuning

[4]We use the term *stance error* interchangeably with hard negative retrieval rate or Hard@R, since hard negatives match the claim topic but violate the requested stance.

successfully suppressed stance errors for Qwen and Instructor-XL, it also triggered an associated increase in retrieving semi-hard negatives (documents matching the stance but not the topic). On the GPT-5 validation split, Qwen's semi-hard negative rate more than doubled, from 5.6% to 12.9%. For Instructor-XL, this failure was even more pronounced, increasing from 7.5% to 22.6%. These results suggest that standard contrastive optimisation against a homogeneous dataset (i.e., with all triplet negatives being hard negatives) forces models to over-correct, unduly focusing on argument stance while partially neglecting the topic. We elaborate on this phenomenon in the following subsection.

In an additional experiment, we test whether the general shortcomings of bi-encoder embedding models can be mitigated by using an off-the-shelf cross-encoder reranker. Cross-encoders compute semantic similarity scores by applying full self-attention across the concatenated query and document, and can be used to refine the ranking of documents originally retrieved by a bi-encoder. Such a setup is commonly used to improve the reliability of retrieval pipelines [30]. However, empirical evaluation demonstrates that the off-the-shelf reranker fails at our task of stance-aware argument retrieval. When applied to the base Qwen3-Embedding-8B base model, the reranker, Qwen3-Reranker-8B, yielded negligible improvements, with stance errors remaining high (dropping marginally from 39.7% to 38.5%). More alarmingly, passing a candidate pool from the fine-tuned bi-encoder model to the base reranker actively counteracted the post-fine-tuning performance improvements. Because the cross-encoder relies on its own pre-trained lexical biases, its application reduced Precision@R on TFU Training from 0.95 to 0.645 and increased the stance error to 35.3%. These results show that even the application of a computationally demanding reranker may be insufficient for achieving high performance on stance-aware argument retrieval, prompting us to focus on improving bi-encoder models directly.

Full experimental results, including exact NDCG@10 scores and standard deviations across all models and datasets, are provided in Appendix D.

4.2. Topical Collapse

While the retrieval metrics demonstrated a clear reduction in stance error for models like Qwen and Instructor-XL after fine-tuning, the associated rise in semi-hard negatives warrants further investigation. To diagnose the cause, we use our word ablation metrics to evaluate word-level similarity shifts. Comprehensive results for word ablation metrics across all models/datasets are given in Appendix E.

First, we evaluate the Directional Impact (DI) to confirm that training succeeded in making models more sensitive to stance logic. In the base models, ablating instruction stance verbs caused a near-zero similarity shift on average (e.g., 0.036 for hard negatives from TFU Training for Qwen), showing the base models' tendency to largely ignore relational constraints. Conversely, after fine-tuning, the average DI for hard negatives has substantially increased (e.g., to 0.238 on TFU Training for Qwen) while the average DI for positive arguments has substantially decreased (e.g., to -0.199 on TFU Training for Qwen). This indicates that, for the fine-tuned models, the stance keywords cause the query embeddings to draw substantially closer to the positive arguments while moving further away from the hard negatives.

However, evaluating the model's RIS and RCS reveals the drawbacks of this stance focus. Considering Qwen as a representative example, the base version of the model exhibited a strong claim bias, resulting in an average RCS of 75.2% on the TFU Training Arguments dataset. Following contrastive fine-tuning, the model's average RIS on the dataset increased to 36%, with the corresponding RCS decrease to 64%. This suggests that fine-tuning has shifted model's focus to the stance instruction, but given the associated increase in semi-hard negative retrieval, this likely came at the cost of accurately representing the topic. This is corroborated by the full RCS distribution visualised in Figure 5, which shows a substantially heavier lower tail for the fine-tuned model (see Appendix F for more related figures). Individual word-ablation heatmaps, such as the example in Appendix G, provide additional visualisation of the stance-topic trade-off, demonstrating how models shift focus from topical nouns to stance verbs.

These findings suggest that undue focus on stance may be detrimental to assessing topical relevance. The model successfully learns to retrieve attacking arguments but loses some precision in representing the topic of the claim being attacked. We refer to this failure mode as the *topical collapse*.

5. Targeted Interventions for Stance-Aware Retrieval

5.1. Rationale

The analysis in Section 4.2 established that while contrastive fine-tuning successfully forces the bi-encoder to account for instruction constraints, it inadvertently triggers topical collapse. We hypothesise this is due to the used data mixture, with the embedding space shaped by the optimisation pressure to differentiate between the positive and hard negative samples. This pressure may result in interference [31], reducing the ability of the model to accurately represent the semantic topic as a by-product of becoming more sensitive to the stance. Additionally, the training may be negatively affected by spurious patterns in the training set, as positive and hard-negative documents commonly share similar vocabulary and thematic subject matter. Therefore, we posit that increasing the semantic diversity of the training data and reducing these spurious differences may mitigate topical collapse.

5.2. Data-Centric Interventions

Mixed Dataset The triplet negatives in the baseline, homogeneous curriculum consisted solely of hard negatives (D_{hn}). To test whether this strictly homogeneous training data caused topical collapse, we introduce a 'mixed' curriculum. In this distribution, the negative sample space is partitioned equally with one-third hard negatives (D_{hn}), one-third semi-hard negatives (D_{shn}) and one-third easy negatives (D_{en}). To isolate the effect of the data distribution, the number of triplets was fixed at the baseline ceiling of 31,800. Reintroducing D_{shn} forces the model to preserve its topical sensitivity.

Stance-Inverted Argument Augmentation While the mixed curriculum mitigates topical collapse, the remaining D_{hn} triplets still possess different sentence structures and lexical patterns compared to the positive documents (D_{pos}). To prevent the optimiser from exploiting these spurious differences, we generate synthetic, stance-inverted hard negatives to ensure near-identical sentence structure and substantial lexical overlap. However, naively prompting an LLM to invert an argument introduces new risks. For example, simply negating the arguments may make them inconsistent with real-world knowledge. Pre-trained embedding models, particularly those based on generative backbones, may identify such inconsistencies and use them as a spurious signal.

To construct helpful synthetic hard negatives, our generation pipeline uses a 'concede and sever' strategy. The generating LLM concedes any factual knowledge, but synthetically severs the causal link to the target claim while minimising lexical differences. To operationalise this, Gemini 3.5 Flash was instructed to generate candidate stance inversions using a one-shot prompt. The complete prompt is provided in Appendix H. We then replaced 50% of the natural hard negatives in the mixed curriculum with their synthetic counterparts. The D_{shn} and D_{en} data portions were left unchanged.

To validate this approach, we measured word-level Jaccard similarity. Standard argument pairs exhibited an average overlap of just 0.0938. In contrast, our synthetic pairs achieved 0.5866 (0.5896 with NLTK lemmatisation). While this falls short of the 90% target specified in the generation prompt, the discrepancy is likely an artefact of the Jaccard metric (intersection over union). Adding even a few words to invert the stance heavily penalises the score. Nonetheless, this six-fold increase confirms the successful generation of lexically similar arguments.

5.3. Hybrid Search

While the data-centric interventions attempt to improve the dense encoder's ability to balance topic and stance, vector embeddings inherently struggle with lexical matching for rare, domain-specific entities, such as complex medical terms in the ArgTumour dataset. To address this, we implemented a hybrid search pipeline based on a 'division of responsibility'. We fuse our tuned dense retriever with a traditional sparse lexical retriever (BM25 utilising local TF-IDF). Because the sparse model excels at lexical anchoring but is entirely stance-blind, and the tuned dense model excels at relational stance logic, combining them aims to result in more robust retrieval. We utilised Relative Score Fusion (RSF) with a

conservative sparse weighting ($\lambda \in [0.1, 0.15, 0.2]$). This configuration allows the BM25 component to act as a lightweight topical filter while heavily relying on the tuned bi-encoder for stance constraints.

6. Results

6.1. Intervention Evaluation

To evaluate the impact of the proposed interventions, we compare the retrieval metrics across the base models, the homogeneous fine-tuned baselines from Section 4, and models using our interventions. The resulting retrieval compositions (visualised in Figure 3) and metrics reveal substantial differences:

- **BGE-Large**: Results reveal that BGE-Large struggles to balance topic and stance. While the homogeneous curriculum achieved a Precision@R of 0.627 on the TFU Evaluation dataset, our interventions traded stance sensitivity for topical focus. Applying the Mixed + Aug curriculum successfully reduced semi-hard negatives (from 20.7% to 8.5%), but triggered a noticeable regression in relational logic. This nearly doubled the hard negative rate to 23.2%, suggesting that BGE-Large lacks the capacity to robustly maintain both boundaries simultaneously.
- **Instructor-XL**: Instructor-XL demonstrated a strong dependency on high-density hard-negative signals. In its base state on the TFU Evaluation dataset, it was largely stance-blind with a hard-negative rate of 41.1%. While the homogeneous curriculum reduced this down to 17.5%, it also triggered topical collapse, increasing the semi-hard negative rate from 2.7% to 10.1%. When the 'Mixed' curriculum was applied to restore topic-sensitivity, the semi-hard negative rate successfully dropped to 5.1%, but the model seemingly lacked the capacity to robustly learn relational logic. Consequently, its stance error rate regressed to 31.4%.
- **Qwen3-Embedding-8B**: The decoder-only model emerged as the architecture most capable in stance-topic adaptation. The homogeneous curriculum fixed its base stance blindness (reducing stance error from 33.5% to 3.9% on TFU Evaluation) but introduced the expected topical collapse. However, fine-tuning on the 'Mixed' curriculum successfully reduced the semi-hard negative rate from 11.7% to 9.8%, while maintaining good stance sensitivity. Furthermore, introducing the synthetic stance inversions ('Mixed + Aug') further reduced this rate to 7.8%. This resulted in overall best performance, with a Precision@R of 0.845 on the TFU Evaluation dataset.

Finally, evaluating the hybrid fusion (dense 'Mixed + Aug' combined with BM25 via RSF) confirmed our hypothesis regarding the division of responsibility. The addition of the sparse retriever provided a topical safety net for specialised domains, improving Qwen3-Embedding-8B's Precision@R on the ArgTumour dataset from 0.690 up to 0.723 (at $\lambda = 0.1$). However, because BM25 is inherently stance-blind, this lexical boost came at the direct cost of inflating the stance error across all models as the sparse weight (λ) increased. Consequently, hybrid fusion proved highly effective for dense, domain-specific corpora (ArgTumour, AVeriTeC) where entity matching is critical, but offered diminishing returns on general-domain datasets (TFU Evaluation) where the unhindered stance logic of the tuned bi-encoder was already sufficient. Results capturing hybrid retrieval performance are provided in Figure 9.

6.2. Word Ablation Validation

To diagnose the shifting retrieval distributions observed in Section 6.1, we again applied the word ablation metrics. Figure 4 visualises the Relative Claim Sensitivity (RCS) for Qwen on the AVeriTeC dataset. The base model exhibited a severe topic bias, allocating a vast majority of its representational capacity to the claim, resulting in a mean RCS of 0.783. Applying the homogeneous curriculum resulted in topical collapse: the mean RCS dropped sharply to 0.637, creating a notable secondary density peak as the model representations became less sensitive to the topic compared to the instruction.

The application of the data-centric interventions counteracted this phenomenon. By introducing the 'Mixed + Aug' curriculum, the collapse peak was flattened, and semantic focus partially shifted back toward the claim, causing the mean RCS to increase to 0.661. Crucially, evaluating the model's

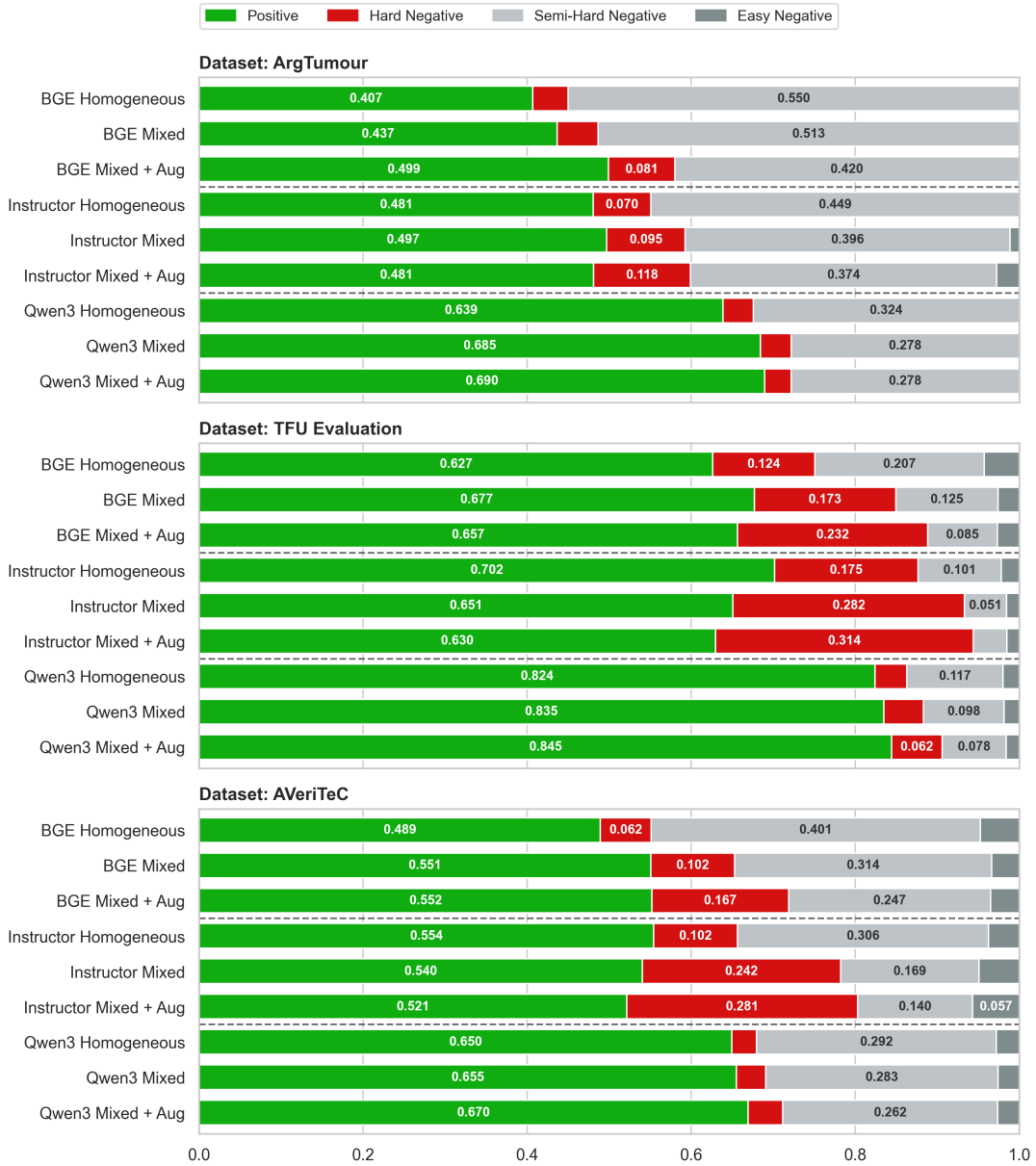

Figure 3: Retrieval composition across the ArgTumour, TFU Evaluation, and AVeriTeC datasets. The chart contrasts the homogeneous baseline against the data-centric interventions (Mixed and Mixed + Aug) across the three model architectures. The balanced curriculum (Mixed) mitigates topical collapse by uniformly reducing the proportion of semi-hard negatives (light grey segments) across all models. However, architectural responses diverge significantly: while Qwen3-Embedding-8B slightly expands its positive retrieval (green segments), both BGE-Large and Instructor-XL exhibit a severe stance regression (inflated red segments) when the hard-negative pressure is reduced. More detailed results can be found in Appendix D

Directional Impact (DI) confirmed that this recovered topical sensitivity did not come at the cost of relational logic. While the base model practically ignored stance verbs (DI for hard negatives: 0.024), the 'Mixed + Aug' model maintained a strong, asymmetric stance boundary (DI for hard negatives: 0.144, DI for positives: −0.120). Computing the ablation metrics across the broader evaluation suite confirmed these general trends across all datasets. Full ablation metric results are given in Appendix E.

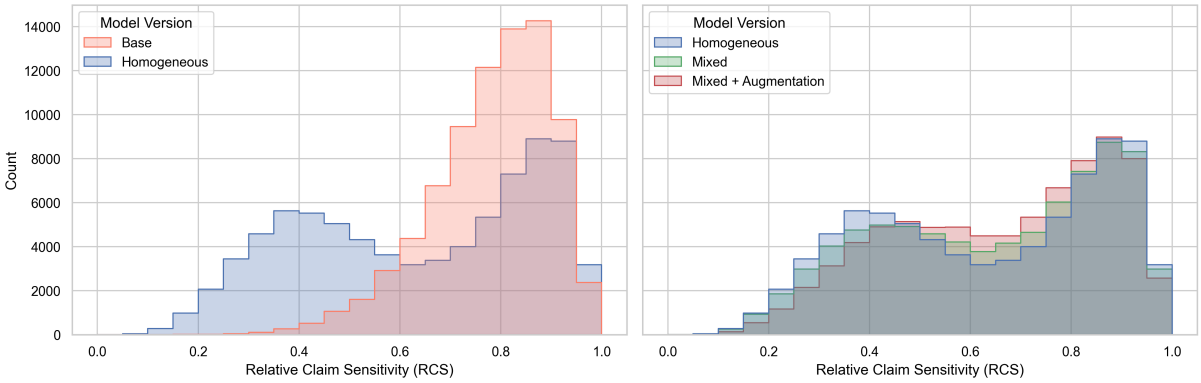

Figure 4: Relative Claim Sensitivity (RCS) distributions for Qwen3-Embedding-8B on AVeriTeC data. **Left**: The Base model exhibits a severe noun bias, allocating the vast majority of its representational capacity to the claim (red distribution). Applying the homogeneous curriculum triggers topical collapse, shifting focus away from the claim and creating a notable secondary peak with low claim sensitivity (blue distribution). **Right**: The data-centric interventions regularise the model's instruction focus. The Mixed (green) and Mixed + Augmentation (red) curricula progressively flatten the peak and pull semantic focus back toward the underlying topic.

7. Conclusion

In this paper, we diagnosed and addressed the inability of base embedding models to balance topical subject matter with directional stance during argument retrieval. Baseline evaluations confirmed that models initially exhibit a severe topic bias, disregarding stance constraints in the instructions. While contrastive fine-tuning on a homogeneous curriculum of hard negatives improves stance sensitivity, it can also degrade the ability of models to correctly represent the topic, resulting in *topical collapse*.

To resolve this collapse, we proposed a two-stage data-centric pipeline. A balanced curriculum increased topical sensitivity across all evaluated architectures, while synthetic stance inversions reduced residual lexical shortcuts. This approach proved particularly effective for Qwen3-Embedding-8B, improving its ability to jointly encode stance and topical relevance as well as the overall retrieval performance. Furthermore, integrating a BM25 sparse retriever provided an additional topical safety net, patching the dense model's remaining blind spots in highly specialised domains.

Future work could further explore applications of our tuned models to argument mining from large text corpora or for argumentative fact verification (e.g., in the spirit of [32, 33]). Additionally, while our data-centric paradigm was reasonably effective at counteracting topical collapse, future research could explore model-centric interventions, such as attention steering, to further improve performance. Finally, it would be interesting to explore dynamically adapting the weighting factor λ in our hybrid search pipeline to specific queries or corpora.

Acknowledgments

Dejl and Toni were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX). Toni was also funded by EPSRC (grant UKRI3928, NeSyDebates).

Declaration on Generative AI

The authors used Gemini 3 Pro to generate standardised training/evaluation instructions and for code assistance (boilerplate PyTorch code, code generating Matplotlib figures and LaTeX tables). Further, the authors used Gemini 3.5 Flash to generate synthetic, stance-inverted arguments for dataset augmentation. Finally, GPT-5.5 was used to provide general feedback on the draft manuscript. After using these tools, the authors reviewed and edited the outputs as needed and take full responsibility for the paper content.

References

[1] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2019) 765–818. URL: https://aclanthology.org/J19-4006/. doi:10.1162/coli_a_00364.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550/. doi:10.18653/v1/2020.emnlp-main.550.

[4] H. Su, W. Shi, J. Kasai, Y. Wang, Y. Hu, M. Ostendorf, W.-t. Yih, N. A. Smith, L. Zettlemoyer, T. Yu, One embedder, any task: Instruction-finetuned text embeddings, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1102–1121. URL: https://aclanthology.org/2023.findings-acl.71/. doi:10.18653/v1/2023.findings-acl.71.

[5] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, CoRR abs/2506.05176 (2025). URL: https://doi.org/10.48550/arXiv.2506.05176. doi:10.48550/ARXIV.2506.05176. arXiv:2506.05176.

[6] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, J.-Y. Nie, C-Pack: Packed resources for general chinese embeddings, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 641–649. URL: https://doi.org/10.1145/3626772.3657878. doi:10.1145/3626772.3657878.

[7] K. Sinha, P. Parthasarathi, J. Pineau, A. Williams, UnNatural Language Inference, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7329–7346. URL: https://aclanthology.org/2021.acl-long.569/. doi:10.18653/v1/2021.acl-long.569.

[8] J. Lawrence, C. Reed, Argument mining: A survey, Comput. Linguistics 45 (2019) 765–818. URL: https://doi.org/10.1162/coli_a_00364. doi:10.1162/COLI_A_00364.

[9] L. Carstens, F. Toni, Towards relation based argumentation mining, in: Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA, The Association for Computational Linguistics, 2015, pp. 29–34. URL: https://doi.org/10.3115/v1/w15-0504. doi:10.3115/V1/W15-0504.

[10] D. Gorur, A. Rago, F. Toni, Can large language models perform relation-based argument mining?, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025, Association for Computational Linguistics, 2025, pp. 8518–8534. URL: https://aclanthology.org/2025.coling-main.569/.

[11] Z. Xu, F. Mo, Z. Huang, C. Zhang, P. Yu, B. W. Phillips, J. Lin, V. Srikumar, A survey of model architectures in information retrieval, Trans. Mach. Learn. Res. 2026 (2026). URL: https://openreview.net/forum?id=xAIbTbHRrX.

[12] R. Nogueira, K. Cho, Passage re-ranking with BERT, CoRR abs/1901.04085 (2019). URL: http://arxiv.org/abs/1901.04085. arXiv:1901.04085.

[13] M. L. Henderson, R. Al-Rfou, B. Strope, Y. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, R. Kurzweil, Efficient natural language response suggestion for smart reply, CoRR abs/1705.00652 (2017). URL: http://arxiv.org/abs/1705.00652. arXiv:1705.00652.

[14] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding,

CoRR abs/1807.03748 (2018). URL: http://arxiv.org/abs/1807.03748. arXiv:1807.03748.

[15] H. Oh, H. Lee, S. Ye, H. Shin, H. Jang, C. Jun, M. Seo, INSTRUCTIR: A benchmark for instruction following of information retrieval models, CoRR abs/2402.14334 (2024). URL: https://doi.org/10.48550/arXiv.2402.14334. doi:10.48550/ARXIV.2402.14334. arXiv:2402.14334.

[16] O. Weller, B. Chang, S. MacAvaney, K. Lo, A. Cohan, B. Van Durme, D. Lawrie, L. Soldaini, FollowIR: Evaluating and teaching information retrieval models to follow instructions, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 11926–11942. URL: https://aclanthology.org/2025.naacl-long.597/. doi:10.18653/v1/2025.naacl-long.597.

[17] Y. Zhuang, A. Trinh, R. Qiang, H. Sun, C. Zhang, H. Dai, B. Dai, Towards better instruction following retrieval models, CoRR abs/2505.21439 (2025). URL: https://doi.org/10.48550/arXiv.2505.21439. doi:10.48550/ARXIV.2505.21439. arXiv:2505.21439.

[18] L. Xiong, C. Xiong, Y. Li, K. Tang, J. Liu, P. N. Bennett, J. Ahmed, A. Overwijk, Approximate nearest neighbor negative contrastive learning for dense text retrieval, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=zeFrfgyZln.

[19] G. d. S. P. Moreira, R. Osmulski, M. Xu, R. Ak, B. Schifferer, E. Oldridge, Improving text embedding models with positive-aware hard-negative mining, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2169–2178. URL: https://doi.org/10.1145/3746252.3761254. doi:10.1145/3746252.3761254.

[20] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, A. Hanbury, Efficiently teaching an effective dense retriever with balanced topic aware sampling, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 113–122. URL: https://doi.org/10.1145/3404835.3462891. doi:10.1145/3404835.3462891.

[21] L. Bonifacio, H. Abonizio, M. Fadaee, R. Nogueira, InPars: Unsupervised dataset generation for information retrieval, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 2387–2392. URL: https://doi.org/10.1145/3477495.3531863. doi:10.1145/3477495.3531863.

[22] Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. B. Hall, M. Chang, Promptagator: Few-shot dense retrieval from 8 examples, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=gmL46YMpu2J.

[23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[24] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[25] C. Cayrol, M. C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 378–389.

[26] G. Freedman, A. Dejl, A. Gould, Mansi, L. Chen, J. Jiang, F. Toni, Neurosymbolic learning for inference-time argumentation, CoRR abs/2605.20098 (2026). URL: https://doi.org/10.48550/arXiv.2605.20098. doi:10.48550/ARXIV.2605.20098. arXiv:2605.20098.

[27] M. S. Schlichtkrull, Z. Guo, A. Vlachos, AVeriTeC: A dataset for real-world claim verification with evidence from the web, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/cd86a30526cd1aff61d6f89f107634e4-Abstract-Datasets_and_Benchmarks.html.

[28] A. Dejl, H. Ayoobi, C. Cherrington, D. Gardner, F. Toni, L. Pakzad-Shahabi, M. Williams, ArgTumour: Integrating large language models and computational argumentation to discuss treatment options for high-grade glioma, Neuro-Oncology 27 (2025) ii26–ii26. URL: https://doi.org/10.1093/neuonc/noaf185.104. doi:10.1093/neuonc/noaf185.104.

[29] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[30] T. Pandit, S. Mahendru, M. Raval, D. Upadhyay, The evolution of reranking models in information retrieval: From heuristic methods to large language models, in: X. Ding, Y. Dong (Eds.), Computational Linguistics and Natural Language Processing, Springer Nature Singapore, Singapore, 2026, pp. 56–67.

[31] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, volume 24 of *Psychology of Learning and Motivation*, Academic Press, 1989, pp. 109–165. URL: https://www.sciencedirect.com/science/article/pii/S0079742108605368. doi:https://doi.org/10.1016/S0079-7421(08)60536-8.

[32] G. Freedman, A. Dejl, D. Gorur, X. Yin, A. Rago, F. Toni, Argumentative large language models for explainable and contestable claim verification, Proceedings of the AAAI Conference on Artificial Intelligence 39 (2025) 14930–14939. URL: https://ojs.aaai.org/index.php/AAAI/article/view/33637. doi:10.1609/aaai.v39i14.33637.

[33] Y. Zhu, N. Potyka, D. Hernández, Y. He, Z. Ding, B. Xiong, D. Zhou, E. Kharlamov, S. Staab, Argrag: Explainable retrieval augmented generation using quantitative bipolar argumentation, in: L. H. Gilpin, E. Giunchiglia, P. Hitzler, E. van Krieken (Eds.), Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning, volume 284 of *Proceedings of Machine Learning Research*, PMLR, 2025, pp. 697–718. URL: https://proceedings.mlr.press/v284/zhu25a.html.

A. Appendices Outline

The appendices provide supplementary materials, detailed experimental results, and extended evaluations to support the findings in the main text. Appendix B details the complete set of seen and unseen instructions utilised for the stance-aware retrieval tasks, and Appendix C provides the experimental details. Appendices D and E supply comprehensive tabular breakdowns of the retrieval performance and word-ablation metrics across all evaluated models and datasets. Subsequent sections offer expanded visual diagnostics, including supplementary Relative Claim Sensitivity (RCS) and Directional Impact (DI) plots in Appendix F, as well as detailed word-ablation heatmaps and their corresponding argument texts in Appendix G. Finally, Appendix H provides the exact one-shot prompt utilised for synthetic data generation, and Appendix I visualises the retrieval results of the hybrid retrieval pipeline.

B. Instruction Sets

Overview of the seen/unseen (i.e., training/testing) instructions used for support/attack stance retrieval tasks respectively.

Seen Instructions

Support

- Find evidence backing
- Retrieve arguments in favor of
- Search for statements validating
- Highlight information corroborating
- Uncover points upholding
- Show claims affirming
- Provide reasoning that supports
- Identify excerpts substantiating
- Extract statements advocating for
- Gather facts verifying

Attack

- Find evidence refuting
- Retrieve arguments against
- Search for statements invalidating
- Highlight information contradicting
- Uncover points opposing
- Show claims challenging
- Provide reasoning that undermines
- Identify excerpts disputing
- Extract statements arguing against
- Gather facts debunking

Unseen Instructions

Support

- Discover sources promoting
- Retrieve evidence confirming
- Locate specific arguments justifying
- Find statements endorsing
- Show points backing up

Attack

- Discover sources dismissing
- Retrieve evidence disproving
- Locate specific arguments attacking
- Find statements rejecting
- Show points doubting

C. Experimental Details

This section outlines the hardware, hyperparameters, and dataset statistics utilised during our experiments.

Hardware Model training and inference were distributed across two high-performance compute clusters. Experiments utilised L40S GPUs hosted on the Imperial College HPC Cluster, as well as A40 GPUs hosted on the Department of Computing (Doc) GPU Cluster.

Training Hyperparameters All fine-tuning experiments were optimised using Multiple Negatives Ranking Loss (MNRL) [13]. The models were trained for 2 epochs with a per-device batch size of 8 and a gradient accumulation step of 2. The optimisation process utilised a learning rate of 2×10^{-5} with a warmup ratio of 0.1. Depending on the architecture, training utilised either FP16 or BF16 precision. For random seeds, the training code relied on the standard default initialisation provided by the Transformers library, with dynamic claim shuffling handled at the batch-sampler level.

LoRA Configuration For fine-tuning via Low-Rank Adaption (LoRA), the α scaling parameter was dynamically set to $2 \times r$ (where $r = 64$), alongside a dropout rate of 0.05. The target modules varied by architecture:

- **BGE-Large:** `query`, `key`, `value`, `dense`
- **Instructor-XL:** q, k, v, o
- **Qwen3-Embedding-8B:** `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`

Dataset Statistics Table 1 provides a comprehensive breakdown of the datasets used for validation and zero-shot evaluation.

Table 1
Dataset statistics detailing the number of claims, total supporting/attacking arguments, and the average number of supporting/attacking arguments per claim (with minimum and maximum bounds).

Dataset	Claims	Total Sup.	Avg Sup. (Min-Max)	Total Att.	Avg Att. (Min-Max)
TFU Training Arguments (80% for training)	283	1051	3.71 (0-7)	1146	4.05 (0-7)
TFU Training Arguments (20% held-out)	71	295	4.15 (1-6)	303	4.27 (2-6)
TFU Validation Arguments (GPT-5)	150	622	4.15 (0-12)	882	5.88 (0-13)
TFU Validation Arguments (Qwen3-8B)	150	374	2.49 (0-5)	485	3.23 (0-5)
ArgTumour Arguments	14	40	2.86 (0-16)	58	4.14 (2-10)
TFU Evaluation Arguments	750	1819	2.43 (0-5)	2321	3.09 (0-5)
AVeriTeC Arguments	1746	3180	1.82 (0-5)	5310	3.04 (0-5)

D. Retrieval Metrics Tables

The following tables present comprehensive stance retrieval metrics across all evaluated datasets and model families. Results represent the mean (μ) and standard deviation (σ). Hard@R, Semi@R, and Easy@R denote the proportion of retrieved hard, semi-hard, and easy negatives within the top R results, respectively. The absolute best score (highest for NDCG/Precision, lowest for error rates) is highlighted in **bold**, while the best score within each dense/hybrid subgroup is underlined. Note that all Hybrid Fusion (RSF) strategies operate exclusively over the Mixed + Aug base models. Note that the reported standard deviations (σ) reflect the variance across different evaluation queries and *unseen* instructions.

Table 2
Retrieval performance on the **TFU Training Arguments (20% held-out)** dataset using **BGE-Large**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.613 ± 0.28	0.438 ± 0.27	0.342 ± 0.26	0.181 ± 0.27	0.038 ± 0.12
	Support	0.605 ± 0.27	0.401 ± 0.27	0.420 ± 0.25	0.043 ± 0.12	0.136 ± 0.23
	Overall	0.609 ± 0.27	0.419 ± 0.27	0.381 ± 0.26	0.112 ± 0.22	0.087 ± 0.19
Homogeneous	Attack	0.883 ± 0.22	0.795 ± 0.27	**0.087 ± 0.17**	0.112 ± 0.19	0.006 ± 0.03
	Support	**0.936 ± 0.12**	**0.850 ± 0.20**	**0.088 ± 0.16**	0.053 ± 0.13	0.010 ± 0.05
	Overall	0.909 ± 0.18	0.822 ± 0.24	**0.088 ± 0.17**	0.082 ± 0.17	0.008 ± 0.04
Mixed	Attack	0.897 ± 0.19	**0.812 ± 0.26**	0.100 ± 0.19	0.084 ± 0.16	0.004 ± 0.03
	Support	0.933 ± 0.11	0.835 ± 0.21	0.144 ± 0.20	0.017 ± 0.07	0.004 ± 0.03
	Overall	0.915 ± 0.16	**0.824 ± 0.23**	0.122 ± 0.19	0.050 ± 0.13	0.004 ± 0.03
Mixed + Aug	Attack	0.898 ± 0.19	0.805 ± 0.26	0.131 ± 0.22	0.058 ± 0.13	0.006 ± 0.03
	Support	0.927 ± 0.12	0.820 ± 0.22	0.157 ± 0.21	0.020 ± 0.08	0.002 ± 0.02
	Overall	0.912 ± 0.16	0.813 ± 0.24	0.144 ± 0.22	0.039 ± 0.11	0.004 ± 0.03
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	**0.902 ± 0.17**	0.794 ± 0.26	0.148 ± 0.23	0.053 ± 0.13	0.005 ± 0.03
	Support	0.929 ± 0.11	0.817 ± 0.22	0.167 ± 0.21	**0.016 ± 0.08**	**0.000 ± 0.00**
	Overall	**0.915 ± 0.15**	0.806 ± 0.24	0.157 ± 0.22	0.034 ± 0.11	0.003 ± 0.02
Hybrid ($\lambda = 0.15$)	Attack	0.900 ± 0.17	0.797 ± 0.25	0.151 ± 0.23	0.048 ± 0.12	0.005 ± 0.03
	Support	0.925 ± 0.12	0.807 ± 0.22	0.176 ± 0.21	0.017 ± 0.08	**0.000 ± 0.00**
	Overall	0.913 ± 0.15	0.802 ± 0.24	0.164 ± 0.22	0.032 ± 0.10	0.003 ± 0.02
Hybrid ($\lambda = 0.20$)	Attack	0.897 ± 0.17	0.783 ± 0.24	0.171 ± 0.22	**0.042 ± 0.11**	**0.004 ± 0.03**
	Support	0.921 ± 0.12	0.799 ± 0.23	0.185 ± 0.22	**0.016 ± 0.07**	**0.000 ± 0.00**
	Overall	0.909 ± 0.14	0.791 ± 0.23	0.178 ± 0.22	**0.029 ± 0.09**	**0.002 ± 0.02**

Table 3
Retrieval performance on the **TFU Training Arguments (20% held-out)** dataset using **Instructor-XL**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.725 ± 0.14	0.482 ± 0.19	0.507 ± 0.20	**0.008 ± 0.06**	0.003 ± 0.02
	Support	0.804 ± 0.14	0.581 ± 0.19	0.410 ± 0.19	0.006 ± 0.04	0.003 ± 0.03
	Overall	0.764 ± 0.15	0.531 ± 0.20	0.459 ± 0.20	**0.007 ± 0.05**	0.003 ± 0.02
Homogeneous	Attack	**0.872 ± 0.23**	**0.773 ± 0.29**	**0.139 ± 0.24**	0.080 ± 0.17	0.009 ± 0.04
	Support	**0.965 ± 0.09**	**0.905 ± 0.17**	**0.049 ± 0.12**	0.030 ± 0.08	0.015 ± 0.07
	Overall	**0.919 ± 0.18**	**0.839 ± 0.24**	**0.094 ± 0.20**	0.055 ± 0.13	0.012 ± 0.06
Mixed	Attack	0.827 ± 0.21	0.654 ± 0.26	0.302 ± 0.26	0.041 ± 0.12	**0.002 ± 0.02**
	Support	0.951 ± 0.08	0.844 ± 0.18	0.141 ± 0.18	0.009 ± 0.04	0.007 ± 0.04
	Overall	0.889 ± 0.17	0.749 ± 0.24	0.222 ± 0.24	0.025 ± 0.09	0.004 ± 0.03
Mixed + Aug	Attack	0.799 ± 0.20	0.602 ± 0.26	0.367 ± 0.26	0.028 ± 0.10	0.003 ± 0.03
	Support	0.937 ± 0.09	0.810 ± 0.19	0.180 ± 0.19	**0.005 ± 0.03**	0.005 ± 0.03
	Overall	0.868 ± 0.17	0.706 ± 0.25	0.273 ± 0.25	0.017 ± 0.07	0.004 ± 0.03
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.793 ± 0.18	0.590 ± 0.25	0.387 ± 0.25	0.021 ± 0.07	**0.002 ± 0.02**
	Support	0.928 ± 0.09	0.783 ± 0.19	0.203 ± 0.19	0.011 ± 0.06	0.003 ± 0.03
	Overall	0.861 ± 0.16	0.686 ± 0.24	0.295 ± 0.24	0.016 ± 0.07	**0.003 ± 0.02**
Hybrid ($\lambda = 0.15$)	Attack	0.788 ± 0.18	0.593 ± 0.23	0.385 ± 0.23	0.017 ± 0.06	0.004 ± 0.03
	Support	0.922 ± 0.09	0.759 ± 0.19	0.226 ± 0.19	0.013 ± 0.06	0.003 ± 0.03
	Overall	0.855 ± 0.16	0.676 ± 0.23	0.306 ± 0.23	0.015 ± 0.06	0.003 ± 0.03
Hybrid ($\lambda = 0.20$)	Attack	0.780 ± 0.18	0.582 ± 0.22	0.398 ± 0.22	0.014 ± 0.05	0.006 ± 0.04
	Support	0.917 ± 0.09	0.737 ± 0.19	0.248 ± 0.19	0.013 ± 0.06	**0.001 ± 0.02**
	Overall	0.849 ± 0.16	0.660 ± 0.22	0.323 ± 0.22	0.014 ± 0.06	0.004 ± 0.03

Table 4
Retrieval performance on the **TFU Training Arguments (20% held-out)** dataset using **Qwen3-Embedding-8B**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.792 ± 0.15	0.559 ± 0.22	0.430 ± 0.22	0.006 ± 0.03	0.004 ± 0.03
	Support	0.866 ± 0.13	0.634 ± 0.21	0.363 ± 0.21	0.003 ± 0.02	**0.000 ± 0.00**
	Overall	0.829 ± 0.14	0.597 ± 0.21	0.397 ± 0.22	0.005 ± 0.03	0.002 ± 0.02
Zero-Shot + Reranker	Attack	0.779 ± 0.16	0.535 ± 0.23	0.465 ± 0.24	**0.000 ± 0.00**	**0.001 ± 0.01**
	Support	0.874 ± 0.13	0.695 ± 0.22	0.305 ± 0.22	**0.000 ± 0.00**	**0.000 ± 0.00**
	Overall	0.826 ± 0.15	0.615 ± 0.24	0.385 ± 0.24	**0.000 ± 0.00**	**0.000 ± 0.01**
Homogeneous	Attack	0.979 ± 0.07	0.934 ± 0.15	0.023 ± 0.10	0.038 ± 0.11	0.006 ± 0.04
	Support	0.993 ± 0.03	0.966 ± 0.10	0.018 ± 0.08	0.012 ± 0.05	0.004 ± 0.03
	Overall	0.986 ± 0.05	0.950 ± 0.13	0.020 ± 0.09	0.025 ± 0.09	0.005 ± 0.03
Homogeneous + Reranker	Attack	0.808 ± 0.16	0.591 ± 0.25	0.406 ± 0.26	**0.000 ± 0.00**	0.003 ± 0.03
	Support	0.884 ± 0.13	0.700 ± 0.21	0.300 ± 0.21	**0.000 ± 0.00**	**0.000 ± 0.00**
	Overall	0.846 ± 0.15	0.645 ± 0.24	0.353 ± 0.24	**0.000 ± 0.00**	0.002 ± 0.02
Mixed	Attack	0.979 ± 0.07	0.933 ± 0.14	**0.019 ± 0.09**	0.045 ± 0.11	0.003 ± 0.03
	Support	**0.994 ± 0.02**	0.972 ± 0.09	**0.016 ± 0.07**	0.010 ± 0.05	0.002 ± 0.02
	Overall	0.987 ± 0.05	0.952 ± 0.12	**0.018 ± 0.08**	0.027 ± 0.09	0.003 ± 0.02
Mixed + Aug	Attack	**0.985 ± 0.04**	**0.939 ± 0.13**	0.026 ± 0.09	0.031 ± 0.10	0.004 ± 0.03
	Support	0.993 ± 0.03	**0.973 ± 0.09**	0.020 ± 0.08	0.005 ± 0.04	0.001 ± 0.02
	Overall	**0.989 ± 0.04**	**0.956 ± 0.11**	0.023 ± 0.09	0.018 ± 0.08	0.003 ± 0.02
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.985 ± 0.04	0.932 ± 0.14	0.033 ± 0.10	0.028 ± 0.09	0.007 ± 0.05
	Support	0.992 ± 0.03	0.967 ± 0.10	0.023 ± 0.08	0.003 ± 0.02	0.007 ± 0.05
	Overall	0.989 ± 0.04	0.949 ± 0.12	0.028 ± 0.09	0.016 ± 0.07	0.007 ± 0.05
Hybrid ($\lambda = 0.15$)	Attack	0.984 ± 0.04	0.927 ± 0.14	0.039 ± 0.11	0.026 ± 0.09	0.009 ± 0.05
	Support	0.992 ± 0.03	0.959 ± 0.10	0.026 ± 0.08	0.007 ± 0.04	0.007 ± 0.05
	Overall	0.988 ± 0.04	0.943 ± 0.12	0.033 ± 0.10	0.017 ± 0.07	0.008 ± 0.05
Hybrid ($\lambda = 0.20$)	Attack	0.982 ± 0.04	0.915 ± 0.14	0.052 ± 0.12	0.023 ± 0.08	0.010 ± 0.06
	Support	0.990 ± 0.03	0.949 ± 0.11	0.036 ± 0.09	0.010 ± 0.04	0.005 ± 0.04
	Overall	0.986 ± 0.04	0.932 ± 0.13	0.044 ± 0.11	0.017 ± 0.07	0.008 ± 0.05

Table 5
Retrieval performance on the **TFU Validation Arguments generated by GPT-5** dataset using **BGE-Large**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.562 ± 0.31	0.448 ± 0.30	0.314 ± 0.28	**0.131 ± 0.18**	0.106 ± 0.15
	Support	0.678 ± 0.22	0.523 ± 0.21	0.261 ± 0.19	0.120 ± 0.18	0.096 ± 0.15
	Overall	0.610 ± 0.28	0.480 ± 0.27	0.292 ± 0.25	0.126 ± 0.18	0.102 ± 0.15
Homogeneous	Attack	0.761 ± 0.25	0.655 ± 0.27	**0.022 ± 0.07**	0.317 ± 0.26	**0.006 ± 0.03**
	Support	0.822 ± 0.19	0.730 ± 0.24	**0.046 ± 0.11**	0.207 ± 0.22	0.017 ± 0.07
	Overall	0.787 ± 0.23	0.686 ± 0.26	**0.032 ± 0.09**	0.271 ± 0.25	0.011 ± 0.05
Mixed	Attack	0.804 ± 0.23	**0.709 ± 0.26**	0.061 ± 0.14	0.224 ± 0.23	0.007 ± 0.04
	Support	**0.846 ± 0.17**	**0.745 ± 0.22**	0.084 ± 0.13	0.157 ± 0.21	0.014 ± 0.05
	Overall	**0.822 ± 0.21**	**0.724 ± 0.24**	0.071 ± 0.13	0.196 ± 0.22	**0.010 ± 0.04**
Mixed + Aug	Attack	0.796 ± 0.23	0.701 ± 0.26	0.086 ± 0.16	0.205 ± 0.22	0.009 ± 0.05
	Support	0.836 ± 0.18	0.733 ± 0.22	0.120 ± 0.15	0.136 ± 0.19	**0.011 ± 0.05**
	Overall	0.813 ± 0.21	0.714 ± 0.24	0.100 ± 0.15	0.176 ± 0.21	0.010 ± 0.05
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.804 ± 0.23	0.704 ± 0.25	0.120 ± 0.19	0.166 ± 0.21	0.010 ± 0.05
	Support	0.835 ± 0.18	0.734 ± 0.22	0.135 ± 0.16	0.116 ± 0.18	0.015 ± 0.05
	Overall	0.817 ± 0.21	0.717 ± 0.24	0.126 ± 0.18	0.145 ± 0.20	0.012 ± 0.05
Hybrid ($\lambda = 0.15$)	Attack	**0.806 ± 0.23**	0.697 ± 0.26	0.136 ± 0.20	0.156 ± 0.20	0.011 ± 0.05
	Support	0.831 ± 0.18	0.728 ± 0.22	0.152 ± 0.16	0.104 ± 0.17	0.016 ± 0.05
	Overall	0.817 ± 0.21	0.710 ± 0.24	0.143 ± 0.18	0.134 ± 0.19	0.013 ± 0.05
Hybrid ($\lambda = 0.20$)	Attack	0.805 ± 0.23	0.700 ± 0.25	0.148 ± 0.20	0.141 ± 0.19	0.012 ± 0.05
	Support	0.823 ± 0.19	0.724 ± 0.22	0.162 ± 0.17	**0.097 ± 0.16**	0.017 ± 0.05
	Overall	0.813 ± 0.21	0.710 ± 0.24	0.154 ± 0.19	**0.123 ± 0.18**	0.014 ± 0.05

Table 6
Retrieval performance on the **TFU Validation Arguments generated by GPT-5** dataset using **Instructor-XL**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.662 ± 0.26	0.529 ± 0.32	0.336 ± 0.31	**0.079 ± 0.14**	0.056 ± 0.13
	Support	0.694 ± 0.24	0.558 ± 0.23	0.319 ± 0.20	**0.070 ± 0.15**	0.054 ± 0.11
	Overall	0.675 ± 0.25	0.541 ± 0.29	0.328 ± 0.27	**0.075 ± 0.15**	0.055 ± 0.12
Homogeneous	Attack	**0.786 ± 0.24**	**0.687 ± 0.28**	**0.043 ± 0.13**	0.263 ± 0.26	**0.007 ± 0.05**
	Support	**0.841 ± 0.19**	**0.754 ± 0.23**	**0.048 ± 0.10**	0.177 ± 0.21	0.021 ± 0.06
	Overall	**0.809 ± 0.23**	**0.715 ± 0.26**	**0.045 ± 0.12**	0.227 ± 0.25	**0.013 ± 0.06**
Mixed	Attack	0.768 ± 0.24	0.646 ± 0.29	0.166 ± 0.24	0.171 ± 0.23	0.017 ± 0.07
	Support	0.821 ± 0.20	0.717 ± 0.23	0.132 ± 0.16	0.132 ± 0.19	**0.019 ± 0.07**
	Overall	0.790 ± 0.23	0.676 ± 0.27	0.152 ± 0.21	0.155 ± 0.22	0.018 ± 0.07
Mixed + Aug	Attack	0.732 ± 0.25	0.605 ± 0.30	0.224 ± 0.27	0.144 ± 0.21	0.026 ± 0.09
	Support	0.813 ± 0.20	0.696 ± 0.22	0.166 ± 0.17	0.117 ± 0.19	0.021 ± 0.07
	Overall	0.766 ± 0.23	0.644 ± 0.27	0.200 ± 0.24	0.133 ± 0.20	0.024 ± 0.08
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.736 ± 0.24	0.609 ± 0.30	0.241 ± 0.27	0.120 ± 0.19	0.030 ± 0.09
	Support	0.800 ± 0.21	0.686 ± 0.22	0.192 ± 0.17	0.100 ± 0.16	0.023 ± 0.07
	Overall	0.763 ± 0.23	0.641 ± 0.27	0.221 ± 0.24	0.111 ± 0.18	0.027 ± 0.09
Hybrid ($\lambda = 0.15$)	Attack	0.738 ± 0.24	0.611 ± 0.29	0.247 ± 0.27	0.110 ± 0.18	0.031 ± 0.10
	Support	0.793 ± 0.21	0.676 ± 0.21	0.207 ± 0.17	0.094 ± 0.16	0.023 ± 0.07
	Overall	0.761 ± 0.23	0.638 ± 0.26	0.230 ± 0.24	0.104 ± 0.17	0.028 ± 0.09
Hybrid ($\lambda = 0.20$)	Attack	0.736 ± 0.24	0.609 ± 0.29	0.254 ± 0.27	0.103 ± 0.17	0.034 ± 0.10
	Support	0.788 ± 0.21	0.663 ± 0.22	0.220 ± 0.18	0.092 ± 0.16	0.025 ± 0.08
	Overall	0.758 ± 0.23	0.632 ± 0.27	0.239 ± 0.24	0.098 ± 0.16	0.030 ± 0.09

Table 7
Retrieval performance on the **TFU Validation Arguments generated by GPT-5** dataset using **Qwen3-Embedding-8B**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.730 ± 0.24	0.600 ± 0.31	0.314 ± 0.29	**0.057 ± 0.13**	0.030 ± 0.10
	Support	0.803 ± 0.18	0.667 ± 0.21	0.256 ± 0.19	**0.055 ± 0.15**	0.022 ± 0.06
	Overall	0.760 ± 0.22	0.628 ± 0.27	0.289 ± 0.26	**0.056 ± 0.14**	0.027 ± 0.09
Homogeneous	Attack	0.906 ± 0.14	0.828 ± 0.19	**0.012 ± 0.06**	0.154 ± 0.19	0.006 ± 0.04
	Support	0.926 ± 0.13	0.877 ± 0.19	**0.020 ± 0.06**	0.095 ± 0.18	0.008 ± 0.03
	Overall	0.914 ± 0.14	0.849 ± 0.19	**0.015 ± 0.06**	0.129 ± 0.19	0.007 ± 0.04
Mixed	Attack	0.915 ± 0.14	0.840 ± 0.20	0.013 ± 0.06	0.141 ± 0.19	0.006 ± 0.04
	Support	**0.926 ± 0.13**	**0.884 ± 0.18**	0.024 ± 0.07	0.085 ± 0.17	0.007 ± 0.03
	Overall	0.920 ± 0.13	0.859 ± 0.19	0.018 ± 0.06	0.118 ± 0.19	0.006 ± 0.04
Mixed + Aug	Attack	0.914 ± 0.14	0.843 ± 0.20	0.026 ± 0.09	0.124 ± 0.19	0.006 ± 0.04
	Support	0.925 ± 0.13	0.876 ± 0.19	0.027 ± 0.07	0.089 ± 0.18	0.008 ± 0.03
	Overall	0.919 ± 0.13	0.857 ± 0.19	0.027 ± 0.08	0.110 ± 0.18	0.007 ± 0.04
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	**0.918 ± 0.13**	**0.845 ± 0.19**	0.039 ± 0.11	0.111 ± 0.17	0.005 ± 0.04
	Support	0.925 ± 0.13	0.878 ± 0.18	0.038 ± 0.09	0.077 ± 0.17	**0.007 ± 0.03**
	Overall	**0.921 ± 0.13**	**0.859 ± 0.19**	0.039 ± 0.10	0.097 ± 0.17	**0.006 ± 0.04**
Hybrid ($\lambda = 0.15$)	Attack	0.916 ± 0.13	0.841 ± 0.19	0.049 ± 0.13	0.104 ± 0.17	0.005 ± 0.05
	Support	0.924 ± 0.13	0.870 ± 0.18	0.045 ± 0.10	0.077 ± 0.17	0.008 ± 0.03
	Overall	0.919 ± 0.13	0.853 ± 0.19	0.048 ± 0.12	0.093 ± 0.17	0.006 ± 0.04
Hybrid ($\lambda = 0.20$)	Attack	0.913 ± 0.13	0.837 ± 0.20	0.058 ± 0.13	0.100 ± 0.16	**0.005 ± 0.05**
	Support	0.921 ± 0.13	0.858 ± 0.19	0.055 ± 0.11	0.078 ± 0.16	0.010 ± 0.04
	Overall	0.916 ± 0.13	0.846 ± 0.19	0.057 ± 0.13	0.091 ± 0.16	0.007 ± 0.04

Table 8
Retrieval performance on the **TFU Validation Arguments generated by Qwen3-8B** dataset using **BGE-Large**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.619 ± 0.24	0.408 ± 0.31	0.481 ± 0.30	**0.036 ± 0.11**	0.075 ± 0.16
	Support	0.823 ± 0.22	0.611 ± 0.29	0.296 ± 0.29	0.067 ± 0.16	0.026 ± 0.09
	Overall	0.715 ± 0.25	0.503 ± 0.32	0.394 ± 0.31	**0.051 ± 0.14**	0.052 ± 0.13
Homogeneous	Attack	0.807 ± 0.22	0.658 ± 0.26	**0.064 ± 0.14**	0.270 ± 0.27	**0.008 ± 0.05**
	Support	0.844 ± 0.24	0.734 ± 0.33	**0.104 ± 0.23**	0.134 ± 0.26	0.028 ± 0.12
	Overall	0.824 ± 0.23	0.694 ± 0.30	**0.083 ± 0.19**	0.206 ± 0.27	0.017 ± 0.09
Mixed	Attack	**0.861 ± 0.17**	**0.716 ± 0.24**	0.105 ± 0.18	0.170 ± 0.23	0.008 ± 0.05
	Support	0.872 ± 0.21	**0.745 ± 0.31**	0.146 ± 0.25	0.082 ± 0.20	0.026 ± 0.12
	Overall	**0.866 ± 0.19**	**0.730 ± 0.27**	0.124 ± 0.22	0.129 ± 0.22	0.017 ± 0.09
Mixed + Aug	Attack	0.858 ± 0.18	0.697 ± 0.26	0.170 ± 0.22	0.117 ± 0.20	0.015 ± 0.07
	Support	0.854 ± 0.21	0.695 ± 0.31	0.210 ± 0.28	0.076 ± 0.19	0.018 ± 0.08
	Overall	0.856 ± 0.19	0.696 ± 0.28	0.189 ± 0.25	0.098 ± 0.20	0.016 ± 0.07
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.853 ± 0.18	0.690 ± 0.25	0.216 ± 0.21	0.077 ± 0.17	0.017 ± 0.07
	Support	0.868 ± 0.20	0.712 ± 0.29	0.202 ± 0.26	0.072 ± 0.18	0.014 ± 0.07
	Overall	0.860 ± 0.19	0.700 ± 0.27	0.209 ± 0.24	0.075 ± 0.18	0.015 ± 0.07
Hybrid ($\lambda = 0.15$)	Attack	0.836 ± 0.19	0.676 ± 0.26	0.241 ± 0.22	0.064 ± 0.15	0.018 ± 0.08
	Support	0.870 ± 0.20	0.712 ± 0.28	0.206 ± 0.26	0.070 ± 0.18	0.011 ± 0.07
	Overall	0.852 ± 0.19	0.693 ± 0.27	0.225 ± 0.24	0.067 ± 0.16	0.015 ± 0.07
Hybrid ($\lambda = 0.20$)	Attack	0.824 ± 0.19	0.672 ± 0.26	0.255 ± 0.23	0.055 ± 0.13	0.017 ± 0.08
	Support	**0.875 ± 0.20**	0.719 ± 0.28	0.209 ± 0.26	**0.062 ± 0.17**	**0.010 ± 0.06**
	Overall	0.848 ± 0.20	0.694 ± 0.27	0.234 ± 0.25	0.058 ± 0.15	**0.014 ± 0.07**

Table 9
Retrieval performance on the **TFU Validation Arguments generated by Qwen3-8B** dataset using **Instructor-XL**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.710 ± 0.22	0.492 ± 0.33	0.445 ± 0.30	**0.025 ± 0.09**	0.038 ± 0.12
	Support	0.820 ± 0.21	0.595 ± 0.30	0.342 ± 0.31	0.041 ± 0.13	0.022 ± 0.09
	Overall	0.762 ± 0.22	0.540 ± 0.32	0.397 ± 0.31	**0.033 ± 0.11**	0.030 ± 0.11
Homogeneous	Attack	**0.844 ± 0.19**	**0.687 ± 0.26**	**0.100 ± 0.16**	0.209 ± 0.25	**0.004 ± 0.04**
	Support	**0.895 ± 0.21**	**0.783 ± 0.31**	**0.103 ± 0.23**	0.106 ± 0.24	**0.008 ± 0.06**
	Overall	**0.868 ± 0.20**	**0.732 ± 0.29**	**0.101 ± 0.20**	0.161 ± 0.25	**0.006 ± 0.05**
Mixed	Attack	0.818 ± 0.21	0.646 ± 0.27	0.236 ± 0.24	0.103 ± 0.20	0.015 ± 0.07
	Support	0.878 ± 0.21	0.717 ± 0.31	0.201 ± 0.29	0.067 ± 0.18	0.015 ± 0.08
	Overall	0.846 ± 0.21	0.679 ± 0.29	0.219 ± 0.27	0.086 ± 0.19	0.015 ± 0.07
Mixed + Aug	Attack	0.785 ± 0.22	0.603 ± 0.31	0.299 ± 0.27	0.071 ± 0.16	0.027 ± 0.10
	Support	0.870 ± 0.21	0.686 ± 0.31	0.239 ± 0.30	0.057 ± 0.15	0.018 ± 0.09
	Overall	0.825 ± 0.22	0.642 ± 0.31	0.271 ± 0.29	0.065 ± 0.16	0.023 ± 0.09
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.772 ± 0.22	0.589 ± 0.31	0.330 ± 0.27	0.052 ± 0.14	0.028 ± 0.11
	Support	0.870 ± 0.21	0.685 ± 0.31	0.252 ± 0.30	0.045 ± 0.13	0.019 ± 0.09
	Overall	0.818 ± 0.22	0.634 ± 0.31	0.293 ± 0.29	0.049 ± 0.14	0.024 ± 0.10
Hybrid ($\lambda = 0.15$)	Attack	0.766 ± 0.22	0.580 ± 0.31	0.345 ± 0.27	0.047 ± 0.13	0.029 ± 0.11
	Support	0.869 ± 0.21	0.684 ± 0.31	0.259 ± 0.30	0.039 ± 0.13	0.018 ± 0.09
	Overall	0.814 ± 0.22	0.629 ± 0.31	0.304 ± 0.29	0.043 ± 0.13	0.024 ± 0.10
Hybrid ($\lambda = 0.20$)	Attack	0.760 ± 0.22	0.568 ± 0.31	0.360 ± 0.28	0.044 ± 0.13	0.029 ± 0.11
	Support	0.869 ± 0.20	0.679 ± 0.30	0.266 ± 0.30	**0.038 ± 0.13**	0.017 ± 0.08
	Overall	0.811 ± 0.22	0.620 ± 0.31	0.316 ± 0.29	0.041 ± 0.13	0.023 ± 0.10

Table 10
Retrieval performance on the **TFU Validation Arguments generated by Qwen3-8B** dataset using **Qwen3-Embedding-8B**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.780 ± 0.20	0.566 ± 0.30	0.387 ± 0.28	**0.027 ± 0.09**	0.020 ± 0.09
	Support	0.881 ± 0.20	0.705 ± 0.31	0.242 ± 0.31	**0.038 ± 0.11**	0.015 ± 0.08
	Overall	0.828 ± 0.20	0.631 ± 0.31	0.319 ± 0.30	**0.032 ± 0.10**	0.018 ± 0.08
Homogeneous	Attack	0.932 ± 0.12	0.832 ± 0.21	0.025 ± 0.10	0.136 ± 0.20	0.007 ± 0.05
	Support	0.903 ± 0.25	0.836 ± 0.30	**0.024 ± 0.12**	0.125 ± 0.27	0.014 ± 0.09
	Overall	0.918 ± 0.19	0.834 ± 0.26	**0.025 ± 0.11**	0.131 ± 0.24	0.010 ± 0.07
Mixed	Attack	0.929 ± 0.12	0.823 ± 0.21	**0.023 ± 0.09**	0.148 ± 0.20	0.006 ± 0.05
	Support	0.917 ± 0.22	0.850 ± 0.29	0.046 ± 0.19	0.093 ± 0.24	0.011 ± 0.07
	Overall	0.923 ± 0.17	0.835 ± 0.25	0.034 ± 0.14	0.122 ± 0.22	0.008 ± 0.06
Mixed + Aug	Attack	0.945 ± 0.09	0.844 ± 0.20	0.032 ± 0.10	0.118 ± 0.19	0.006 ± 0.05
	Support	0.921 ± 0.21	**0.852 ± 0.29**	0.066 ± 0.22	0.072 ± 0.21	0.009 ± 0.07
	Overall	0.934 ± 0.16	0.848 ± 0.25	0.048 ± 0.17	0.097 ± 0.20	0.008 ± 0.06
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	**0.954 ± 0.08**	**0.856 ± 0.19**	0.050 ± 0.12	0.089 ± 0.16	0.005 ± 0.04
	Support	0.930 ± 0.19	0.851 ± 0.29	0.068 ± 0.22	0.071 ± 0.20	0.010 ± 0.07
	Overall	**0.943 ± 0.15**	**0.854 ± 0.24**	0.059 ± 0.18	0.080 ± 0.18	0.007 ± 0.05
Hybrid ($\lambda = 0.15$)	Attack	0.952 ± 0.09	0.852 ± 0.19	0.066 ± 0.13	0.077 ± 0.15	**0.005 ± 0.04**
	Support	**0.932 ± 0.19**	0.848 ± 0.29	0.073 ± 0.22	0.071 ± 0.20	0.008 ± 0.06
	Overall	0.943 ± 0.14	0.850 ± 0.24	0.070 ± 0.18	0.074 ± 0.18	0.006 ± 0.05
Hybrid ($\lambda = 0.20$)	Attack	0.949 ± 0.09	0.844 ± 0.19	0.080 ± 0.15	0.071 ± 0.15	**0.005 ± 0.04**
	Support	0.931 ± 0.19	0.846 ± 0.28	0.083 ± 0.23	0.064 ± 0.18	**0.007 ± 0.05**
	Overall	0.941 ± 0.14	0.845 ± 0.24	0.081 ± 0.19	0.068 ± 0.17	**0.006 ± 0.04**

Table 11
Retrieval performance on the **ArgTumour Arguments** dataset using **BGE-Large**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.485 ± 0.40	0.317 ± 0.32	0.128 ± 0.25	0.513 ± 0.31	0.043 ± 0.09
	Support	0.338 ± 0.30	0.259 ± 0.29	0.089 ± 0.13	**0.381 ± 0.12**	0.271 ± 0.15
	Overall	0.441 ± 0.38	0.299 ± 0.31	0.116 ± 0.22	0.473 ± 0.27	0.111 ± 0.15
Homogeneous	Attack	0.567 ± 0.34	0.444 ± 0.31	**0.021 ± 0.08**	0.535 ± 0.30	**0.000 ± 0.00**
	Support	0.421 ± 0.32	0.321 ± 0.25	0.094 ± 0.14	0.585 ± 0.23	**0.000 ± 0.00**
	Overall	0.523 ± 0.34	0.407 ± 0.30	**0.043 ± 0.11**	0.550 ± 0.28	**0.000 ± 0.00**
Mixed	Attack	0.596 ± 0.36	0.492 ± 0.36	0.036 ± 0.13	0.472 ± 0.34	**0.000 ± 0.00**
	Support	0.475 ± 0.30	0.306 ± 0.24	**0.083 ± 0.13**	0.610 ± 0.20	**0.000 ± 0.00**
	Overall	0.560 ± 0.34	0.437 ± 0.34	0.050 ± 0.13	0.513 ± 0.31	**0.000 ± 0.00**
Mixed + Aug	Attack	0.655 ± 0.38	0.576 ± 0.40	0.068 ± 0.19	0.356 ± 0.36	**0.000 ± 0.00**
	Support	0.494 ± 0.30	0.320 ± 0.25	0.111 ± 0.16	0.569 ± 0.22	**0.000 ± 0.00**
	Overall	0.607 ± 0.36	0.499 ± 0.38	0.081 ± 0.18	0.420 ± 0.34	**0.000 ± 0.00**
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	**0.672 ± 0.38**	0.589 ± 0.39	0.071 ± 0.20	0.340 ± 0.36	**0.000 ± 0.00**
	Support	0.518 ± 0.26	0.321 ± 0.23	0.183 ± 0.20	0.496 ± 0.14	**0.000 ± 0.00**
	Overall	0.626 ± 0.36	0.508 ± 0.37	0.105 ± 0.21	0.387 ± 0.31	**0.000 ± 0.00**
Hybrid ($\lambda = 0.15$)	Attack	0.669 ± 0.38	**0.601 ± 0.40**	0.066 ± 0.18	**0.333 ± 0.36**	**0.000 ± 0.00**
	Support	0.532 ± 0.23	0.345 ± 0.20	0.206 ± 0.20	0.449 ± 0.08	**0.000 ± 0.00**
	Overall	0.628 ± 0.35	**0.524 ± 0.37**	0.108 ± 0.20	0.368 ± 0.31	**0.000 ± 0.00**
Hybrid ($\lambda = 0.20$)	Attack	0.670 ± 0.38	0.595 ± 0.39	0.062 ± 0.17	0.343 ± 0.36	**0.000 ± 0.00**
	Support	**0.541 ± 0.21**	**0.356 ± 0.19**	0.222 ± 0.19	0.421 ± 0.08	**0.000 ± 0.00**
	Overall	**0.631 ± 0.34**	0.523 ± 0.36	0.110 ± 0.19	**0.367 ± 0.30**	**0.000 ± 0.00**

Table 12
Retrieval performance on the **ArgTumour Arguments** dataset using **Instructor-XL**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.579 ± 0.38	0.503 ± 0.39	0.157 ± 0.26	0.287 ± 0.30	0.054 ± 0.10
	Support	0.473 ± 0.25	0.371 ± 0.27	0.121 ± 0.13	0.413 ± 0.22	0.095 ± 0.18
	Overall	0.547 ± 0.35	0.463 ± 0.36	0.146 ± 0.23	0.324 ± 0.28	0.066 ± 0.13
Homogeneous	Attack	0.567 ± 0.42	0.507 ± 0.41	**0.053 ± 0.14**	0.440 ± 0.37	**0.000 ± 0.00**
	Support	**0.534 ± 0.30**	**0.419 ± 0.26**	**0.111 ± 0.18**	0.470 ± 0.20	**0.000 ± 0.00**
	Overall	0.557 ± 0.39	0.481 ± 0.38	**0.070 ± 0.15**	0.449 ± 0.33	**0.000 ± 0.00**
Mixed	Attack	0.581 ± 0.41	0.537 ± 0.41	0.089 ± 0.21	0.358 ± 0.35	0.016 ± 0.07
	Support	0.489 ± 0.30	0.403 ± 0.26	**0.111 ± 0.17**	0.486 ± 0.31	**0.000 ± 0.00**
	Overall	0.553 ± 0.38	0.497 ± 0.38	0.095 ± 0.19	0.396 ± 0.35	0.011 ± 0.06
Mixed + Aug	Attack	0.586 ± 0.40	0.530 ± 0.41	0.119 ± 0.23	0.312 ± 0.33	0.039 ± 0.10
	Support	0.478 ± 0.28	0.367 ± 0.25	0.117 ± 0.17	0.517 ± 0.28	**0.000 ± 0.00**
	Overall	0.554 ± 0.37	0.481 ± 0.38	0.118 ± 0.22	0.373 ± 0.33	0.027 ± 0.09
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.599 ± 0.40	0.542 ± 0.42	0.121 ± 0.23	0.290 ± 0.33	0.047 ± 0.12
	Support	0.510 ± 0.25	0.379 ± 0.25	0.150 ± 0.19	0.471 ± 0.25	**0.000 ± 0.00**
	Overall	0.572 ± 0.36	0.493 ± 0.39	0.130 ± 0.22	0.344 ± 0.32	0.033 ± 0.10
Hybrid ($\lambda = 0.15$)	Attack	0.602 ± 0.40	0.556 ± 0.41	0.128 ± 0.23	0.269 ± 0.32	0.046 ± 0.11
	Support	0.518 ± 0.25	0.402 ± 0.26	0.206 ± 0.20	0.393 ± 0.21	**0.000 ± 0.00**
	Overall	0.577 ± 0.36	0.510 ± 0.38	0.151 ± 0.22	0.306 ± 0.29	0.033 ± 0.10
Hybrid ($\lambda = 0.20$)	Attack	**0.612 ± 0.39**	**0.560 ± 0.41**	0.137 ± 0.23	**0.256 ± 0.30**	0.046 ± 0.11
	Support	0.522 ± 0.24	0.402 ± 0.26	0.211 ± 0.19	**0.387 ± 0.20**	**0.000 ± 0.00**
	Overall	**0.585 ± 0.36**	**0.512 ± 0.38**	0.159 ± 0.22	**0.296 ± 0.28**	0.033 ± 0.10

Table 13
Retrieval performance on the **ArgTumour Arguments** dataset using **Qwen3-Embedding-8B**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.705 ± 0.38	0.683 ± 0.39	0.183 ± 0.29	**0.125 ± 0.23**	0.009 ± 0.04
	Support	0.782 ± 0.19	0.625 ± 0.19	0.218 ± 0.19	**0.118 ± 0.12**	0.038 ± 0.13
	Overall	0.728 ± 0.33	0.666 ± 0.34	0.193 ± 0.26	**0.123 ± 0.20**	0.017 ± 0.08
Homogeneous	Attack	0.771 ± 0.23	0.626 ± 0.31	0.050 ± 0.18	0.324 ± 0.30	**0.000 ± 0.00**
	Support	0.783 ± 0.15	0.669 ± 0.17	0.006 ± 0.03	0.325 ± 0.17	**0.000 ± 0.00**
	Overall	0.775 ± 0.20	0.639 ± 0.28	0.037 ± 0.16	0.324 ± 0.27	**0.000 ± 0.00**
Mixed	Attack	0.798 ± 0.22	0.684 ± 0.30	0.054 ± 0.19	0.262 ± 0.29	**0.000 ± 0.00**
	Support	0.809 ± 0.12	0.686 ± 0.16	**0.000 ± 0.00**	0.314 ± 0.16	**0.000 ± 0.00**
	Overall	0.801 ± 0.20	0.685 ± 0.27	0.037 ± 0.16	0.278 ± 0.26	**0.000 ± 0.00**
Mixed + Aug	Attack	0.788 ± 0.23	0.682 ± 0.30	0.046 ± 0.17	0.271 ± 0.28	**0.000 ± 0.00**
	Support	0.811 ± 0.13	0.707 ± 0.15	**0.000 ± 0.00**	0.293 ± 0.15	**0.000 ± 0.00**
	Overall	0.795 ± 0.21	0.690 ± 0.26	**0.033 ± 0.15**	0.278 ± 0.25	**0.000 ± 0.00**
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.798 ± 0.23	0.717 ± 0.30	**0.039 ± 0.15**	0.244 ± 0.28	**0.000 ± 0.00**
	Support	**0.838 ± 0.13**	**0.737 ± 0.20**	0.028 ± 0.06	0.235 ± 0.18	**0.000 ± 0.00**
	Overall	0.810 ± 0.21	0.723 ± 0.27	0.036 ± 0.13	0.241 ± 0.26	**0.000 ± 0.00**
Hybrid ($\lambda = 0.15$)	Attack	0.800 ± 0.23	0.721 ± 0.29	**0.039 ± 0.15**	0.239 ± 0.27	**0.000 ± 0.00**
	Support	0.831 ± 0.13	0.730 ± 0.21	0.028 ± 0.06	0.242 ± 0.19	**0.000 ± 0.00**
	Overall	0.809 ± 0.21	0.724 ± 0.27	0.036 ± 0.13	0.240 ± 0.25	**0.000 ± 0.00**
Hybrid ($\lambda = 0.20$)	Attack	**0.804 ± 0.23**	**0.729 ± 0.29**	**0.039 ± 0.15**	0.232 ± 0.28	**0.000 ± 0.00**
	Support	0.828 ± 0.14	0.730 ± 0.21	0.028 ± 0.06	0.242 ± 0.19	**0.000 ± 0.00**
	Overall	**0.811 ± 0.20**	**0.729 ± 0.26**	0.036 ± 0.13	0.235 ± 0.25	**0.000 ± 0.00**

Table 14
Retrieval performance on the **TFU Evaluation Arguments** dataset using **BGE-Large**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.590 ± 0.27	0.411 ± 0.30	0.416 ± 0.30	0.107 ± 0.19	0.066 ± 0.16
	Support	0.766 ± 0.26	0.591 ± 0.33	0.296 ± 0.31	0.071 ± 0.18	0.042 ± 0.13
	Overall	0.673 ± 0.28	0.497 ± 0.33	0.359 ± 0.31	0.090 ± 0.19	0.055 ± 0.14
Homogeneous	Attack	0.734 ± 0.29	0.599 ± 0.31	**0.104 ± 0.20**	0.261 ± 0.28	0.036 ± 0.12
	Support	0.775 ± 0.31	0.658 ± 0.36	**0.147 ± 0.28**	0.146 ± 0.27	0.049 ± 0.15
	Overall	0.753 ± 0.30	0.627 ± 0.33	**0.124 ± 0.24**	0.207 ± 0.28	0.043 ± 0.14
Mixed	Attack	0.793 ± 0.26	**0.656 ± 0.31**	0.149 ± 0.24	0.170 ± 0.24	0.024 ± 0.09
	Support	**0.820 ± 0.28**	**0.701 ± 0.33**	0.198 ± 0.30	0.074 ± 0.19	0.027 ± 0.11
	Overall	**0.806 ± 0.27**	**0.677 ± 0.32**	0.173 ± 0.27	0.125 ± 0.23	0.026 ± 0.10
Mixed + Aug	Attack	0.802 ± 0.25	0.655 ± 0.31	0.197 ± 0.27	0.118 ± 0.20	0.030 ± 0.11
	Support	0.806 ± 0.28	0.658 ± 0.34	0.272 ± 0.32	0.048 ± 0.15	0.022 ± 0.09
	Overall	0.804 ± 0.26	0.657 ± 0.32	0.232 ± 0.30	0.085 ± 0.18	0.027 ± 0.10
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	**0.810 ± 0.23**	0.650 ± 0.30	0.236 ± 0.27	0.089 ± 0.18	0.025 ± 0.10
	Support	0.800 ± 0.27	0.641 ± 0.33	0.301 ± 0.32	0.041 ± 0.14	0.017 ± 0.08
	Overall	0.806 ± 0.25	0.646 ± 0.32	0.267 ± 0.30	0.066 ± 0.17	0.021 ± 0.09
Hybrid ($\lambda = 0.15$)	Attack	0.809 ± 0.23	0.645 ± 0.30	0.253 ± 0.28	0.080 ± 0.17	0.022 ± 0.09
	Support	0.792 ± 0.27	0.623 ± 0.33	0.321 ± 0.32	0.039 ± 0.14	**0.017 ± 0.08**
	Overall	0.801 ± 0.25	0.634 ± 0.31	0.285 ± 0.30	0.060 ± 0.16	0.020 ± 0.09
Hybrid ($\lambda = 0.20$)	Attack	0.806 ± 0.22	0.638 ± 0.30	0.269 ± 0.28	**0.073 ± 0.17**	**0.020 ± 0.09**
	Support	0.784 ± 0.27	0.608 ± 0.33	0.337 ± 0.33	**0.038 ± 0.13**	0.018 ± 0.08
	Overall	0.796 ± 0.25	0.624 ± 0.31	0.301 ± 0.30	**0.056 ± 0.15**	**0.019 ± 0.09**

Table 15
Retrieval performance on the **TFU Evaluation Arguments** dataset using **Instructor-XL**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.740 ± 0.22	0.538 ± 0.30	0.413 ± 0.29	**0.028 ± 0.10**	0.021 ± 0.09
	Support	0.757 ± 0.27	0.548 ± 0.32	0.408 ± 0.32	**0.025 ± 0.10**	0.018 ± 0.08
	Overall	0.748 ± 0.24	0.543 ± 0.31	0.411 ± 0.31	**0.027 ± 0.10**	0.020 ± 0.09
Homogeneous	Attack	**0.797 ± 0.25**	**0.649 ± 0.31**	**0.196 ± 0.27**	0.137 ± 0.22	0.018 ± 0.08
	Support	**0.861 ± 0.27**	**0.760 ± 0.32**	**0.152 ± 0.29**	0.062 ± 0.17	0.026 ± 0.10
	Overall	**0.827 ± 0.26**	**0.702 ± 0.32**	**0.175 ± 0.28**	0.101 ± 0.21	0.022 ± 0.09
Mixed	Attack	0.792 ± 0.23	0.611 ± 0.30	0.310 ± 0.29	0.064 ± 0.15	0.015 ± 0.08
	Support	0.842 ± 0.26	0.695 ± 0.32	0.252 ± 0.31	0.036 ± 0.12	0.017 ± 0.08
	Overall	0.816 ± 0.25	0.651 ± 0.31	0.282 ± 0.30	0.051 ± 0.14	0.016 ± 0.08
Mixed + Aug	Attack	0.779 ± 0.22	0.594 ± 0.30	0.344 ± 0.30	0.049 ± 0.13	**0.014 ± 0.07**
	Support	0.830 ± 0.26	0.670 ± 0.32	0.281 ± 0.32	0.033 ± 0.12	0.016 ± 0.08
	Overall	0.804 ± 0.24	0.630 ± 0.31	0.314 ± 0.31	0.041 ± 0.13	**0.015 ± 0.07**
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.779 ± 0.21	0.588 ± 0.30	0.361 ± 0.29	0.036 ± 0.11	0.015 ± 0.08
	Support	0.819 ± 0.26	0.645 ± 0.31	0.311 ± 0.31	0.028 ± 0.11	**0.016 ± 0.07**
	Overall	0.798 ± 0.24	0.615 ± 0.31	0.338 ± 0.30	0.032 ± 0.11	0.016 ± 0.08
Hybrid ($\lambda = 0.15$)	Attack	0.776 ± 0.21	0.584 ± 0.29	0.368 ± 0.29	0.032 ± 0.10	0.015 ± 0.08
	Support	0.808 ± 0.26	0.627 ± 0.31	0.330 ± 0.32	0.026 ± 0.10	0.017 ± 0.08
	Overall	0.791 ± 0.24	0.604 ± 0.30	0.350 ± 0.30	0.030 ± 0.10	0.016 ± 0.08
Hybrid ($\lambda = 0.20$)	Attack	0.772 ± 0.21	0.579 ± 0.30	0.373 ± 0.29	0.032 ± 0.11	0.016 ± 0.08
	Support	0.798 ± 0.26	0.613 ± 0.31	0.341 ± 0.32	0.027 ± 0.10	0.018 ± 0.08
	Overall	0.784 ± 0.24	0.595 ± 0.30	0.358 ± 0.30	0.030 ± 0.10	0.017 ± 0.08

Table 16
Retrieval performance on the **TFU Evaluation Arguments** dataset using **Qwen3-Embedding-8B**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.787 ± 0.20	0.581 ± 0.30	0.381 ± 0.30	**0.025 ± 0.09**	0.013 ± 0.07
	Support	0.843 ± 0.26	0.680 ± 0.33	0.285 ± 0.32	**0.021 ± 0.10**	**0.014 ± 0.07**
	Overall	0.814 ± 0.23	0.628 ± 0.32	0.335 ± 0.32	**0.023 ± 0.10**	**0.014 ± 0.07**
Homogeneous	Attack	0.918 ± 0.17	0.835 ± 0.24	**0.038 ± 0.13**	0.114 ± 0.21	0.014 ± 0.07
	Support	0.871 ± 0.29	0.813 ± 0.32	**0.040 ± 0.17**	0.121 ± 0.27	0.026 ± 0.12
	Overall	0.896 ± 0.23	0.824 ± 0.28	**0.039 ± 0.15**	0.117 ± 0.24	0.020 ± 0.09
Mixed	Attack	0.922 ± 0.16	0.839 ± 0.24	0.041 ± 0.14	0.107 ± 0.20	0.013 ± 0.07
	Support	0.887 ± 0.27	0.831 ± 0.31	0.057 ± 0.21	0.089 ± 0.23	0.024 ± 0.11
	Overall	0.906 ± 0.22	0.835 ± 0.27	0.049 ± 0.17	0.098 ± 0.22	0.018 ± 0.09
Mixed + Aug	Attack	0.932 ± 0.15	**0.850 ± 0.23**	0.049 ± 0.15	0.088 ± 0.18	0.013 ± 0.07
	Support	0.894 ± 0.26	**0.839 ± 0.31**	0.076 ± 0.24	0.067 ± 0.20	0.018 ± 0.10
	Overall	0.914 ± 0.21	**0.845 ± 0.27**	0.062 ± 0.20	0.078 ± 0.19	0.016 ± 0.08
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	**0.936 ± 0.14**	0.850 ± 0.23	0.068 ± 0.17	0.070 ± 0.16	**0.012 ± 0.06**
	Support	**0.897 ± 0.26**	0.838 ± 0.30	0.100 ± 0.27	0.044 ± 0.15	0.018 ± 0.09
	Overall	**0.918 ± 0.21**	0.844 ± 0.26	0.083 ± 0.22	0.058 ± 0.16	0.015 ± 0.08
Hybrid ($\lambda = 0.15$)	Attack	0.935 ± 0.14	0.843 ± 0.23	0.081 ± 0.18	0.063 ± 0.15	0.012 ± 0.06
	Support	0.896 ± 0.26	0.830 ± 0.30	0.109 ± 0.28	0.043 ± 0.14	0.018 ± 0.09
	Overall	0.916 ± 0.20	0.837 ± 0.26	0.094 ± 0.23	0.054 ± 0.15	0.015 ± 0.07
Hybrid ($\lambda = 0.20$)	Attack	0.932 ± 0.14	0.830 ± 0.23	0.096 ± 0.19	0.061 ± 0.15	0.013 ± 0.06
	Support	0.892 ± 0.26	0.816 ± 0.30	0.121 ± 0.28	0.044 ± 0.14	0.019 ± 0.09
	Overall	0.913 ± 0.20	0.824 ± 0.27	0.107 ± 0.24	0.053 ± 0.15	0.016 ± 0.08

Table 17
Retrieval performance on the **AVeriTeC Arguments** dataset using **BGE-Large**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.457 ± 0.30	0.322 ± 0.28	0.315 ± 0.29	**0.202 ± 0.26**	0.161 ± 0.24
	Support	0.617 ± 0.33	0.453 ± 0.36	0.300 ± 0.36	0.139 ± 0.26	0.108 ± 0.24
	Overall	0.530 ± 0.33	0.382 ± 0.33	0.308 ± 0.33	0.173 ± 0.26	0.137 ± 0.24
Homogeneous	Attack	0.590 ± 0.31	0.464 ± 0.31	**0.059 ± 0.15**	0.452 ± 0.32	0.024 ± 0.09
	Support	0.636 ± 0.38	0.519 ± 0.40	**0.065 ± 0.20**	0.340 ± 0.39	0.075 ± 0.22
	Overall	0.611 ± 0.35	0.489 ± 0.35	**0.062 ± 0.18**	0.401 ± 0.36	0.047 ± 0.16
Mixed	Attack	0.647 ± 0.30	0.507 ± 0.31	0.076 ± 0.18	0.397 ± 0.31	**0.019 ± 0.08**
	Support	0.730 ± 0.34	**0.603 ± 0.39**	0.132 ± 0.28	0.215 ± 0.34	0.050 ± 0.18
	Overall	0.685 ± 0.32	0.551 ± 0.35	0.102 ± 0.23	0.314 ± 0.34	0.033 ± 0.14
Mixed + Aug	Attack	0.673 ± 0.29	0.525 ± 0.31	0.129 ± 0.22	0.325 ± 0.31	0.020 ± 0.08
	Support	0.731 ± 0.33	0.584 ± 0.38	0.211 ± 0.33	0.153 ± 0.29	0.051 ± 0.18
	Overall	0.699 ± 0.31	0.552 ± 0.35	0.167 ± 0.28	0.247 ± 0.31	0.035 ± 0.14
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.702 ± 0.28	0.554 ± 0.30	0.166 ± 0.23	0.258 ± 0.29	0.023 ± 0.09
	Support	**0.748 ± 0.32**	0.594 ± 0.38	0.249 ± 0.35	0.115 ± 0.25	0.042 ± 0.16
	Overall	0.723 ± 0.30	**0.572 ± 0.34**	0.204 ± 0.30	0.192 ± 0.28	**0.032 ± 0.13**
Hybrid ($\lambda = 0.15$)	Attack	0.707 ± 0.28	**0.559 ± 0.30**	0.182 ± 0.24	0.234 ± 0.28	0.025 ± 0.09
	Support	0.748 ± 0.32	0.586 ± 0.38	0.266 ± 0.35	0.107 ± 0.24	**0.041 ± 0.15**
	Overall	**0.725 ± 0.30**	0.572 ± 0.34	0.220 ± 0.30	0.176 ± 0.27	0.032 ± 0.13
Hybrid ($\lambda = 0.20$)	Attack	**0.708 ± 0.27**	0.559 ± 0.30	0.194 ± 0.24	0.221 ± 0.27	0.026 ± 0.10
	Support	0.744 ± 0.31	0.576 ± 0.38	0.283 ± 0.36	**0.100 ± 0.23**	0.042 ± 0.16
	Overall	0.724 ± 0.29	0.567 ± 0.34	0.235 ± 0.30	**0.165 ± 0.26**	0.033 ± 0.13

Table 18
Retrieval performance on the **AVeriTeC Arguments** dataset using **Instructor-XL**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.625 ± 0.27	0.452 ± 0.30	0.346 ± 0.30	**0.120 ± 0.21**	0.081 ± 0.18
	Support	0.689 ± 0.31	0.488 ± 0.36	0.383 ± 0.36	**0.069 ± 0.19**	0.060 ± 0.19
	Overall	0.655 ± 0.29	0.469 ± 0.33	0.363 ± 0.33	**0.097 ± 0.20**	0.072 ± 0.18
Homogeneous	Attack	0.637 ± 0.31	**0.496 ± 0.31**	**0.122 ± 0.22**	0.357 ± 0.32	**0.025 ± 0.10**
	Support	0.737 ± 0.35	**0.623 ± 0.40**	**0.079 ± 0.23**	0.246 ± 0.36	0.052 ± 0.18
	Overall	0.683 ± 0.33	**0.554 ± 0.36**	**0.102 ± 0.22**	0.306 ± 0.34	**0.037 ± 0.14**
Mixed	Attack	0.652 ± 0.29	0.488 ± 0.31	0.251 ± 0.28	0.209 ± 0.28	0.051 ± 0.14
	Support	**0.759 ± 0.31**	0.602 ± 0.38	0.231 ± 0.34	0.121 ± 0.26	0.046 ± 0.17
	Overall	0.701 ± 0.30	0.540 ± 0.35	0.242 ± 0.31	0.169 ± 0.27	0.049 ± 0.16
Mixed + Aug	Attack	0.644 ± 0.28	0.474 ± 0.30	0.293 ± 0.29	0.169 ± 0.25	0.063 ± 0.16
	Support	0.750 ± 0.31	0.578 ± 0.38	0.267 ± 0.35	0.107 ± 0.24	0.048 ± 0.17
	Overall	0.693 ± 0.30	0.521 ± 0.34	0.281 ± 0.32	0.140 ± 0.25	0.057 ± 0.17
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.660 ± 0.27	0.486 ± 0.30	0.307 ± 0.29	0.147 ± 0.23	0.060 ± 0.16
	Support	0.750 ± 0.31	0.568 ± 0.37	0.303 ± 0.36	0.084 ± 0.22	0.045 ± 0.17
	Overall	0.701 ± 0.29	0.524 ± 0.34	0.305 ± 0.32	0.118 ± 0.23	0.054 ± 0.16
Hybrid ($\lambda = 0.15$)	Attack	0.664 ± 0.27	0.489 ± 0.30	0.311 ± 0.29	0.141 ± 0.23	0.059 ± 0.16
	Support	0.746 ± 0.30	0.562 ± 0.37	0.315 ± 0.36	0.077 ± 0.21	0.046 ± 0.17
	Overall	**0.701 ± 0.29**	0.522 ± 0.33	0.313 ± 0.33	0.112 ± 0.22	0.053 ± 0.16
Hybrid ($\lambda = 0.20$)	Attack	**0.666 ± 0.27**	0.490 ± 0.30	0.313 ± 0.29	0.138 ± 0.22	0.059 ± 0.15
	Support	0.739 ± 0.30	0.553 ± 0.37	0.328 ± 0.36	0.074 ± 0.20	**0.045 ± 0.16**
	Overall	0.699 ± 0.29	0.519 ± 0.33	0.320 ± 0.33	0.109 ± 0.22	0.053 ± 0.16

Table 19
Retrieval performance on the **AVeriTeC Arguments** dataset using **Qwen3-Embedding-8B**.

Strategy	Stance	NDCG@10	Prec@R	Hard@R	Semi@R	Easy@R
GROUP 1: Dense Retrieval						
Zero-Shot	Attack	0.689 ± 0.25	0.497 ± 0.30	0.334 ± 0.30	**0.116 ± 0.21**	0.053 ± 0.14
	Support	0.756 ± 0.30	0.571 ± 0.37	0.324 ± 0.37	**0.065 ± 0.19**	0.040 ± 0.15
	Overall	0.720 ± 0.28	0.531 ± 0.33	0.330 ± 0.33	**0.093 ± 0.20**	0.047 ± 0.15
Homogeneous	Attack	0.778 ± 0.25	0.642 ± 0.30	**0.042 ± 0.13**	0.295 ± 0.29	0.020 ± 0.09
	Support	0.741 ± 0.37	0.658 ± 0.40	**0.016 ± 0.11**	0.289 ± 0.39	0.037 ± 0.16
	Overall	0.761 ± 0.31	0.650 ± 0.35	**0.030 ± 0.12**	0.292 ± 0.34	0.028 ± 0.13
Mixed	Attack	0.776 ± 0.25	0.636 ± 0.30	0.042 ± 0.14	0.304 ± 0.29	0.018 ± 0.08
	Support	0.765 ± 0.36	0.679 ± 0.40	0.028 ± 0.14	0.258 ± 0.37	0.035 ± 0.16
	Overall	0.771 ± 0.30	0.655 ± 0.35	0.036 ± 0.14	0.283 ± 0.33	0.026 ± 0.12
Mixed + Aug	Attack	0.794 ± 0.24	0.654 ± 0.30	0.045 ± 0.14	0.284 ± 0.29	0.017 ± 0.08
	Support	0.779 ± 0.35	0.688 ± 0.40	0.039 ± 0.17	0.237 ± 0.36	0.037 ± 0.16
	Overall	0.787 ± 0.29	0.670 ± 0.35	0.042 ± 0.16	0.262 ± 0.33	0.026 ± 0.12
GROUP 2: Hybrid Fusion (RSF)						
Hybrid ($\lambda = 0.10$)	Attack	0.819 ± 0.22	0.686 ± 0.28	0.059 ± 0.15	0.240 ± 0.27	**0.016 ± 0.08**
	Support	0.804 ± 0.33	0.711 ± 0.38	0.057 ± 0.21	0.199 ± 0.34	0.034 ± 0.15
	Overall	0.812 ± 0.28	0.697 ± 0.33	0.058 ± 0.18	0.221 ± 0.31	0.024 ± 0.12
Hybrid ($\lambda = 0.15$)	Attack	0.824 ± 0.22	**0.689 ± 0.28**	0.069 ± 0.16	0.226 ± 0.27	0.016 ± 0.07
	Support	0.811 ± 0.32	**0.715 ± 0.38**	0.072 ± 0.23	0.182 ± 0.32	0.031 ± 0.14
	Overall	0.818 ± 0.27	**0.701 ± 0.33**	0.070 ± 0.20	0.206 ± 0.30	0.023 ± 0.11
Hybrid ($\lambda = 0.20$)	Attack	**0.824 ± 0.22**	0.687 ± 0.28	0.082 ± 0.17	0.214 ± 0.26	0.016 ± 0.08
	Support	**0.813 ± 0.32**	0.713 ± 0.38	0.089 ± 0.25	0.169 ± 0.31	**0.029 ± 0.14**
	Overall	**0.819 ± 0.27**	0.699 ± 0.33	0.085 ± 0.21	0.193 ± 0.29	**0.022 ± 0.11**

E. Ablation Metrics Tables

The following tables present the word-ablation diagnostic metrics (RIS, RCS, and DI) across all evaluated datasets and models. Note that, similar to the retrieval results, the reported standard deviations (σ) reflect the variance across different evaluation queries and *unseen* instructions.

Table 20
Ablation metrics on the **TFU Training Arguments (20% held-out)** dataset. Results present the mean (μ) and standard deviation (σ). Metrics include Relative Instruction Sensitivity (RIS), Relative Claim Sensitivity (RCS), and Directional Impact (DI) calculated for positive and hard negative pairs.

Model	Strategy	RIS	RCS	DI (Pos)	DI (Hard Neg)
BGE-Large	Zero-Shot	0.195 ± 0.10	0.805 ± 0.10	-0.010 ± 0.01	-0.002 ± 0.01
	Homogeneous	0.354 ± 0.25	0.646 ± 0.25	-0.123 ± 0.16	0.179 ± 0.17
	Mixed	0.331 ± 0.22	0.669 ± 0.22	-0.092 ± 0.13	0.157 ± 0.15
	Mixed + Aug	0.315 ± 0.21	0.685 ± 0.21	-0.077 ± 0.12	0.138 ± 0.13
Instructor-XL	Zero-Shot	0.139 ± 0.09	0.861 ± 0.09	-0.004 ± 0.01	0.001 ± 0.01
	Homogeneous	0.374 ± 0.22	0.626 ± 0.22	-0.085 ± 0.10	0.124 ± 0.12
	Mixed	0.299 ± 0.20	0.701 ± 0.20	-0.040 ± 0.06	0.059 ± 0.07
	Mixed + Aug	0.265 ± 0.19	0.735 ± 0.19	-0.031 ± 0.04	0.041 ± 0.05
Qwen3-Embedding-8B	Zero-Shot	0.248 ± 0.14	0.752 ± 0.14	-0.008 ± 0.03	0.035 ± 0.04
	Homogeneous	0.360 ± 0.26	0.640 ± 0.26	-0.199 ± 0.21	0.238 ± 0.25
	Mixed	0.362 ± 0.25	0.638 ± 0.25	-0.164 ± 0.18	0.233 ± 0.23
	Mixed + Aug	0.356 ± 0.24	0.644 ± 0.24	-0.147 ± 0.16	0.214 ± 0.21

Table 21
Ablation metrics on the **TFU Validation Arguments generated by GPT-5** dataset. Results present the mean (μ) and standard deviation (σ). Metrics include Relative Instruction Sensitivity (RIS), Relative Claim Sensitivity (RCS), and Directional Impact (DI) calculated for positive and hard negative pairs.

Model	Strategy	RIS	RCS	DI (Pos)	DI (Hard Neg)
BGE-Large	Zero-Shot	0.187 ± 0.13	0.813 ± 0.13	-0.009 ± 0.01	-0.003 ± 0.01
	Homogeneous	0.362 ± 0.25	0.638 ± 0.25	-0.135 ± 0.14	0.171 ± 0.16
	Mixed	0.318 ± 0.21	0.682 ± 0.21	-0.091 ± 0.10	0.137 ± 0.12
	Mixed + Aug	0.295 ± 0.20	0.705 ± 0.20	-0.075 ± 0.09	0.117 ± 0.11
Instructor-XL	Zero-Shot	0.134 ± 0.09	0.866 ± 0.09	-0.004 ± 0.01	0.001 ± 0.01
	Homogeneous	0.389 ± 0.21	0.611 ± 0.21	-0.086 ± 0.09	0.125 ± 0.10
	Mixed	0.320 ± 0.20	0.680 ± 0.20	-0.047 ± 0.05	0.064 ± 0.07
	Mixed + Aug	0.284 ± 0.19	0.716 ± 0.19	-0.036 ± 0.04	0.047 ± 0.05
Qwen3-Embedding-8B	Zero-Shot	0.287 ± 0.13	0.713 ± 0.13	0.006 ± 0.03	0.046 ± 0.03
	Homogeneous	0.413 ± 0.26	0.587 ± 0.26	-0.205 ± 0.19	0.221 ± 0.22
	Mixed	0.409 ± 0.25	0.591 ± 0.25	-0.161 ± 0.15	0.217 ± 0.21
	Mixed + Aug	0.402 ± 0.24	0.598 ± 0.24	-0.148 ± 0.14	0.193 ± 0.18

Table 22
Ablation metrics on the **TFU Validation Arguments generated by Qwen3-8B** dataset. Results present the mean (μ) and standard deviation (σ). Metrics include Relative Instruction Sensitivity (RIS), Relative Claim Sensitivity (RCS), and Directional Impact (DI) calculated for positive and hard negative pairs.

Model	Strategy	RIS	RCS	DI (Pos)	DI (Hard Neg)
BGE-Large	Zero-Shot	0.134 ± 0.11	0.866 ± 0.11	-0.007 ± 0.01	-0.002 ± 0.01
	Homogeneous	0.323 ± 0.24	0.677 ± 0.24	-0.069 ± 0.12	0.157 ± 0.17
	Mixed	0.283 ± 0.21	0.717 ± 0.21	-0.037 ± 0.08	0.130 ± 0.14
	Mixed + Aug	0.251 ± 0.20	0.749 ± 0.20	-0.020 ± 0.07	0.106 ± 0.11
Instructor-XL	Zero-Shot	0.103 ± 0.07	0.897 ± 0.07	-0.002 ± 0.01	0.003 ± 0.01
	Homogeneous	0.341 ± 0.20	0.659 ± 0.20	-0.051 ± 0.08	0.115 ± 0.10
	Mixed	0.273 ± 0.19	0.727 ± 0.19	-0.017 ± 0.04	0.067 ± 0.07
	Mixed + Aug	0.238 ± 0.18	0.762 ± 0.18	-0.014 ± 0.03	0.049 ± 0.06
Qwen3-Embedding-8B	Zero-Shot	0.280 ± 0.13	0.720 ± 0.13	-0.006 ± 0.04	0.051 ± 0.04
	Homogeneous	0.400 ± 0.26	0.600 ± 0.26	-0.177 ± 0.20	0.214 ± 0.23
	Mixed	0.400 ± 0.25	0.600 ± 0.25	-0.126 ± 0.15	0.218 ± 0.22
	Mixed + Aug	0.384 ± 0.23	0.616 ± 0.23	-0.104 ± 0.13	0.195 ± 0.19

Table 23
Ablation metrics on the **ArgTumour Arguments** dataset. Results present the mean (μ) and standard deviation (σ). Metrics include Relative Instruction Sensitivity (RIS), Relative Claim Sensitivity (RCS), and Directional Impact (DI) calculated for positive and hard negative pairs.

Model	Strategy	RIS	RCS	DI (Pos)	DI (Hard Neg)
BGE-Large	Zero-Shot	0.152 ± 0.08	0.848 ± 0.08	-0.004 ± 0.01	-0.001 ± 0.01
	Homogeneous	0.372 ± 0.24	0.628 ± 0.24	-0.090 ± 0.13	0.091 ± 0.10
	Mixed	0.330 ± 0.21	0.670 ± 0.21	-0.036 ± 0.10	0.092 ± 0.09
	Mixed + Aug	0.301 ± 0.19	0.699 ± 0.19	-0.030 ± 0.09	0.077 ± 0.07
Instructor-XL	Zero-Shot	0.141 ± 0.08	0.859 ± 0.08	-0.004 ± 0.01	0.004 ± 0.01
	Homogeneous	0.395 ± 0.21	0.605 ± 0.21	-0.076 ± 0.12	0.064 ± 0.08
	Mixed	0.346 ± 0.20	0.654 ± 0.20	-0.039 ± 0.08	0.054 ± 0.06
	Mixed + Aug	0.326 ± 0.19	0.674 ± 0.19	-0.043 ± 0.08	0.044 ± 0.05
Qwen3-Embedding-8B	Zero-Shot	0.259 ± 0.12	0.741 ± 0.12	-0.002 ± 0.03	0.045 ± 0.03
	Homogeneous	0.375 ± 0.23	0.625 ± 0.23	-0.221 ± 0.17	0.125 ± 0.15
	Mixed	0.374 ± 0.24	0.626 ± 0.24	-0.174 ± 0.13	0.133 ± 0.15
	Mixed + Aug	0.350 ± 0.23	0.650 ± 0.23	-0.143 ± 0.11	0.131 ± 0.14

Table 24
Ablation metrics on the **TFU Evaluation Arguments** dataset. Results present the mean (μ) and standard deviation (σ). Metrics include Relative Instruction Sensitivity (RIS), Relative Claim Sensitivity (RCS), and Directional Impact (DI) calculated for positive and hard negative pairs.

Model	Strategy	RIS	RCS	DI (Pos)	DI (Hard Neg)
BGE-Large	Zero-Shot	0.216 ± 0.11	0.784 ± 0.11	-0.004 ± 0.02	0.004 ± 0.02
	Homogeneous	0.313 ± 0.22	0.687 ± 0.22	-0.053 ± 0.14	0.131 ± 0.18
	Mixed	0.265 ± 0.20	0.735 ± 0.20	-0.022 ± 0.10	0.114 ± 0.14
	Mixed + Aug	0.238 ± 0.18	0.762 ± 0.18	0.000 ± 0.08	0.098 ± 0.12
Instructor-XL	Zero-Shot	0.133 ± 0.08	0.867 ± 0.08	0.000 ± 0.01	0.005 ± 0.01
	Homogeneous	0.315 ± 0.21	0.685 ± 0.21	-0.022 ± 0.08	0.103 ± 0.11
	Mixed	0.254 ± 0.18	0.746 ± 0.18	-0.004 ± 0.04	0.056 ± 0.07
	Mixed + Aug	0.225 ± 0.16	0.775 ± 0.16	-0.008 ± 0.03	0.039 ± 0.05
Qwen3-Embedding-8B	Zero-Shot	0.251 ± 0.14	0.749 ± 0.14	-0.006 ± 0.04	0.043 ± 0.04
	Homogeneous	0.339 ± 0.24	0.661 ± 0.24	-0.162 ± 0.19	0.192 ± 0.23
	Mixed	0.330 ± 0.24	0.670 ± 0.24	-0.125 ± 0.16	0.188 ± 0.21
	Mixed + Aug	0.315 ± 0.22	0.685 ± 0.22	-0.104 ± 0.14	0.173 ± 0.19

Table 25
Ablation metrics on the **AVeriTeC Arguments** dataset. Results present the mean (μ) and standard deviation (σ). Metrics include Relative Instruction Sensitivity (RIS), Relative Claim Sensitivity (RCS), and Directional Impact (DI) calculated for positive and hard negative pairs.

Model	Strategy	RIS	RCS	DI (Pos)	DI (Hard Neg)
BGE-Large	Zero-Shot	0.127 ± 0.09	0.873 ± 0.09	-0.006 ± 0.01	-0.002 ± 0.01
	Homogeneous	0.311 ± 0.21	0.689 ± 0.21	-0.083 ± 0.13	0.103 ± 0.17
	Mixed	0.275 ± 0.19	0.725 ± 0.19	-0.047 ± 0.09	0.098 ± 0.14
	Mixed + Aug	0.228 ± 0.17	0.772 ± 0.17	-0.027 ± 0.07	0.074 ± 0.10
Instructor-XL	Zero-Shot	0.107 ± 0.07	0.893 ± 0.07	-0.002 ± 0.01	0.002 ± 0.01
	Homogeneous	0.312 ± 0.19	0.688 ± 0.19	-0.047 ± 0.08	0.093 ± 0.11
	Mixed	0.231 ± 0.16	0.769 ± 0.16	-0.019 ± 0.04	0.046 ± 0.06
	Mixed + Aug	0.197 ± 0.14	0.803 ± 0.14	-0.016 ± 0.03	0.030 ± 0.05
Qwen3-Embedding-8B	Zero-Shot	0.217 ± 0.12	0.783 ± 0.12	-0.010 ± 0.03	0.024 ± 0.03
	Homogeneous	0.363 ± 0.23	0.637 ± 0.23	-0.185 ± 0.20	0.159 ± 0.23
	Mixed	0.354 ± 0.23	0.646 ± 0.23	-0.143 ± 0.17	0.161 ± 0.22
	Mixed + Aug	0.339 ± 0.21	0.661 ± 0.21	-0.120 ± 0.15	0.144 ± 0.19

F. Ablation Metrics Figures

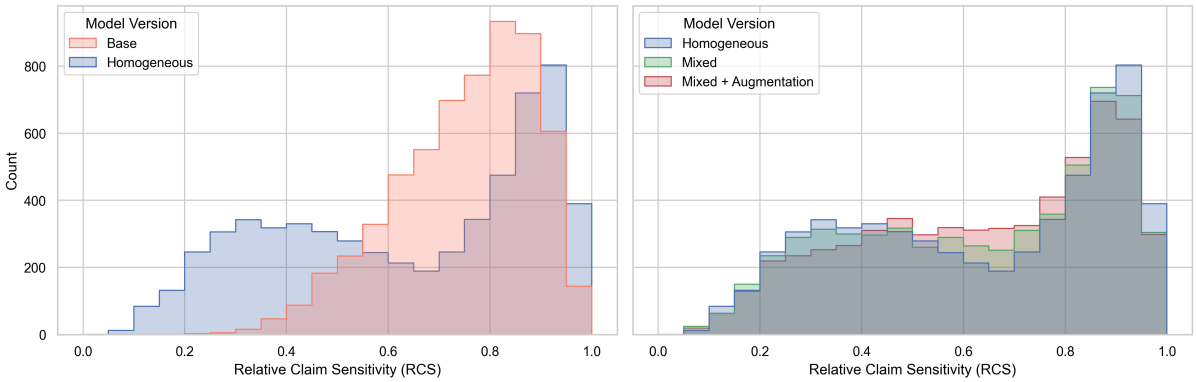

Figure 5: Relative Claim Sensitivity (RCS) distributions for Qwen3-Embedding-8B on the 20% held-out TFU Training Arguments dataset. **Left**: The Base model exhibits a strong noun bias, focusing primarily on the claim (red distribution). The Homogeneous curriculum triggers topical collapse, shifting attention away from the claim to hyper-fixate on the instruction (blue distribution). **Right**: The data-centric interventions (Mixed and Mixed + Augmentation) successfully mitigate this collapse, flattening the secondary peak and organically restoring attention to the underlying topic.

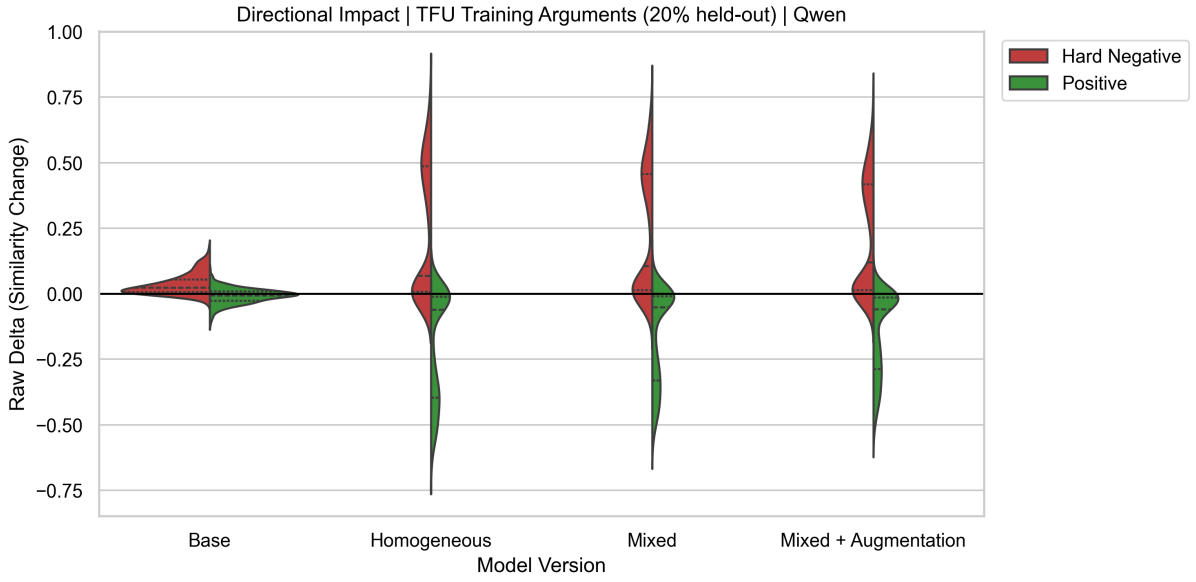

Figure 6: Directional Impact (DI) distributions for Qwen3-Embedding-8B evaluated on the 20% held-out TFU Training Arguments dataset. The Base model exhibits near-zero similarity shifts when stance keywords are ablated, confirming its blindness to relational logic. Following fine-tuning, all training paradigms (Homogeneous, Mixed, and Mixed + Augmentation) demonstrate strong asymmetric stance boundaries, successfully pulling positive documents closer (green) while repelling stance-violating hard negatives (red).

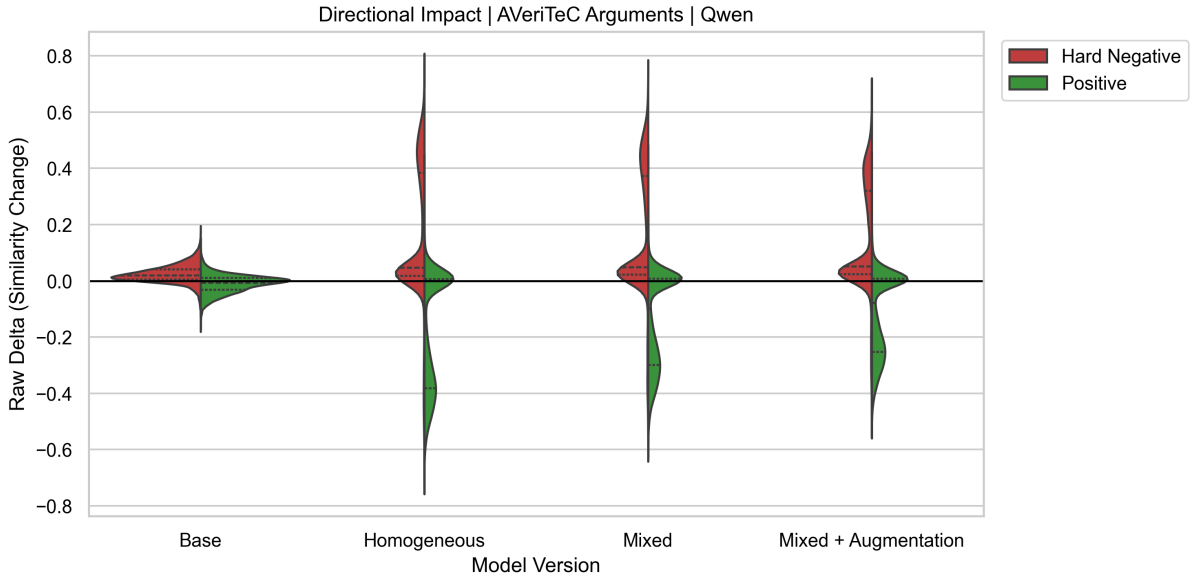

Figure 7: Directional Impact (DI) distributions for Qwen3-Embedding-8B on the out-of-domain AVeriTeC Arguments dataset. Mirroring the in-domain results, the Base model fails to process directional instructions. Crucially, the Mixed and Mixed + Augmentation models sustain their wide distributional splits between positive and hard-negative documents, showing that the data-centric interventions maintain strict, generalisable stance boundaries even when recovering from topical collapse.

G. Word-Ablation Heatmap

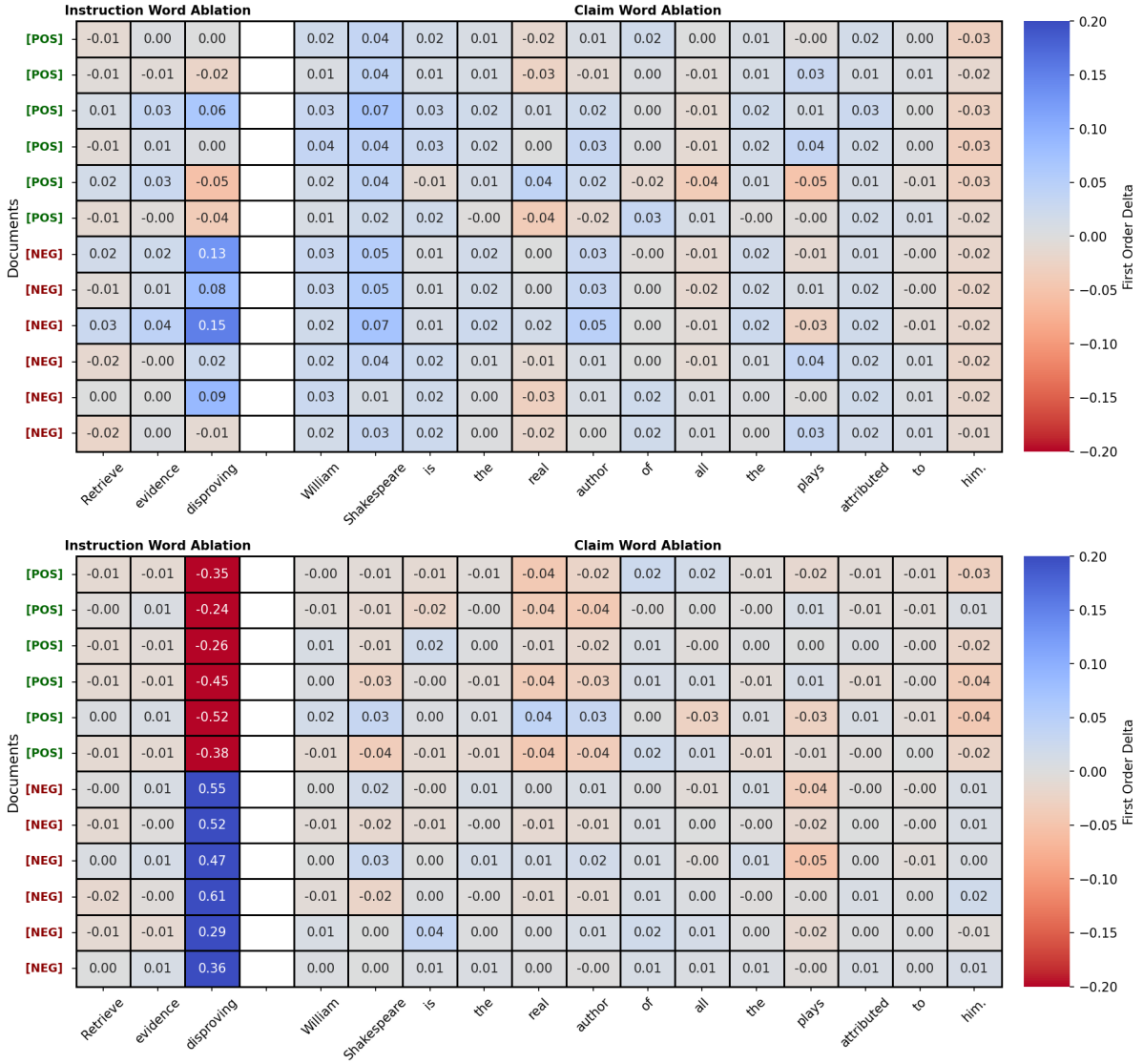

Figure 8: Word-ablation grids analysing the impact of removing individual words on document retrieval scores for base and tuned (homogeneous) Qwen3-Embedding-8B models. The *x*-axis displays the removed words from both the instruction and the claim. The *y*-axis represents the evaluation documents, separated into those attacking ([POS]) and supporting ([NEG]) the claim. The cell values and colours indicate the first-order delta in retrieval score when a specific word is ablated. The arguments (in-order) are given on the next page.

Word-level ablation heatmaps provide a direct, visual confirmation of the topical collapse phenomenon. Figure 8 isolates a specific query instance ("Retrieve evidence disproving" + "William Shakespeare is the real...") on base and tuned (homogeneous) variants of Qwen3-Embedding-8B to illustrate how representational capacity is redistributed after fine-tuning. In the base model, the stance verb "disproving" exerts inconsistent and frequently minimal influence, confirming baseline instruction blindness.

Conversely, the tuned model causes a significant shift in the delta values. The instruction verb "disproving" transforms into a salient, high-contrast anchor, heavily penalising stance-violating documents while pulling correct ones closer. However, this hyper-fixation comes at the direct expense of the claim's semantic anchors. The influence of the topical nouns collapses to near-zero, visually corroborating the macro-level RIS/RCS shifts. Ultimately, the model successfully learns to enforce the "disproving" constraint but at the cost of reducing its ability to verify the underlying subject matter.

Instruction: Retrieve evidence disproving
Claim: William Shakespeare is the real author of all the plays attributed to him.
Positive (Attacking) Arguments:

- The claim is false because there is a major discrepancy between the known life of the Stratford man—a provincial businessman—and the vast, expert knowledge of law, foreign languages, court politics, and aristocratic pursuits detailed in the plays.
- The claim is false because for the most prolific writer of the age, there is a complete lack of a literary paper trail; no personal letters, manuscripts, or books from his library have ever been found, which is highly improbable.
- The claim is false because Shakespeare's detailed will makes no mention whatsoever of any literary assets, such as personal books, plays in progress, or manuscripts, an inexplicable omission for a professional writer.
- The claim is false because the six surviving signatures of Shakespeare are shaky and inconsistent, appearing more like the labored scrawl of a barely literate person than the hand of a literary master.
- The claim's assertion that he authored 'all' the plays is false because modern textual analysis provides strong evidence that several plays, including 'Titus Andronicus' and 'Henry VIII,' were collaborations, meaning he was not the sole author.
- The claim is false because alternative candidates, such as Edward de Vere, Earl of Oxford, possessed the elite education, travel experience, and intimate knowledge of court life that are consistently and accurately reflected in the plays, making them a more plausible match for the author's profile.

Hard Negative (Supporting) Arguments:

- The claim is true because Shakespeare's name was printed on the title pages of numerous plays published during his lifetime, establishing a direct, public link between the man and the works.
- The claim is true because Shakespeare's friends and fellow actors, Ben Jonson, John Heminge, and Henry Condell, explicitly identified him as the author in the First Folio, the definitive collection of his plays. They knew him personally and would have had no reason to perpetuate a fraud.
- The claim is true because financial and legal records confirm that William Shakespeare of Stratford was a shareholder and leading actor in the company that owned and performed the plays, directly connecting the man to the business of the works.
- The claim is true because no one during Shakespeare's life or for two centuries after his death expressed any doubt about his authorship. The idea of an alternative author is a modern theory unsupported by any contemporary evidence.
- The claim is true because Shakespeare's grammar school education in Stratford would have provided a rigorous foundation in Latin, rhetoric, and classical literature, which is sufficient to explain the knowledge in the plays without requiring a university degree or noble birth.
- The claim is true because arguments against his authorship are often rooted in classist snobbery, presuming that a commoner or a 'glover's son' could not possess such genius, which is an assumption about creative ability, not a fact.

H. Stance Inversion Prompt

Prompt submitted to Gemini 3.5 Flash to generate stance-inverted variants of existing arguments:

You are an expert NLP data engineer tasked with generating "Synthetic Stance-Inversions" for a contrastive learning dataset. Your goal is to take a human-written document that supports (or attacks) a claim and rewrite it to explicitly REVERSE its stance.

CRITICAL CONSTRAINTS:

1. **Minimum Edit Distance:** You must maintain a 90%+ proportional lexical overlap with the original text. Keep all nouns, historical entities, subjects, and sentence structures completely identical.
2. **The Relational Flip:** You must ONLY alter the relational verbs, polarity adjectives, or conjunctions to reverse the argumentative polarity.
3. **Factual Integrity (No Hallucinations):** Do not invent false historical facts or corrupt entity knowledge (e.g., do not say a peace activist hated peace). Instead, sever the causal link (e.g., "absence of evidence does not disprove...").
4. **No Explanatory Bloat:** You are strictly forbidden from inventing new reasons, rationalisations, or concluding clauses. Do not explain *why* the claim is false/true. Simply negate the existing causal link.
5. **Semantic Diversity (No Lazy Negation):** Do not rely exclusively on inserting "not" or "does not" before a verb, as this creates logical paradoxes and brittle syntax. Use semantic antonyms or direct verb inversions. Do not over-rely on "Although" templates.
6. **Output Format:** You must output strictly valid JSON containing the new text.

EXAMPLE

- **Claim:** Atlantis was a real, technologically advanced civilization.
- **Original Document:** The claim is false because there is a complete absence of archaeological evidence—such as ruins, tools, pottery, or inscriptions—to support the existence of a large, technologically advanced civilization in 9,000 BCE.
- **Output:**
  - **attack_flipped:** ["The claim is true because a complete absence of archaeological evidence—such as ruins, tools, pottery, or inscriptions—fails to definitively disprove the existence of a large, technologically advanced civilization in 9,000 BCE."]

YOUR TASK You will be given a list of support and a list of attack for a given claim and should return the support_flipped and the attack_flipped lists.

I. Hybrid Retrieval Results

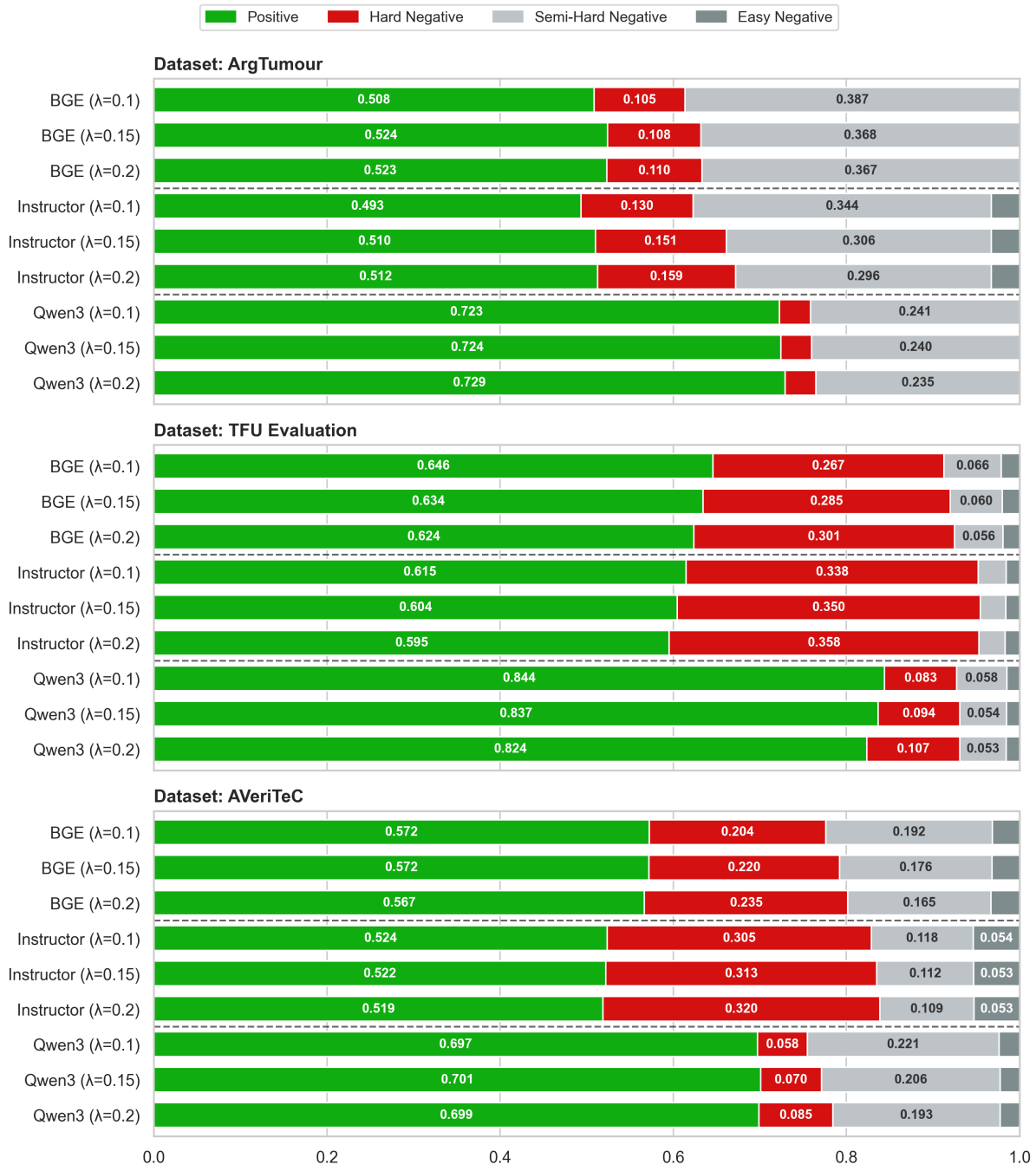

Figure 9: Hybrid retrieval performance combining the optimally tuned dense bi-encoder (Mixed + Augmentation) with a BM25 sparse retriever via Relative Score Fusion (RSF). The visualisations illustrate the impact of incrementally increasing the sparse weighting ($\lambda \in [0.1, 0.15, 0.2]$). Introducing the sparse retriever acts as a topical safety net, actively improving precision on highly specialised, entity-dense corpora like ArgTumour. However, because BM25 is fundamentally stance-blind, increasing its influence directly correlates with a higher number of stance errors across all architectures. This posits a strategic trade-off: hybrid search is highly effective for domain-specific entity matching but yields diminishing returns on general-domain datasets (e.g., TFU Evaluation) where the logic of the tuned dense model is already sufficient.