# Conditional Dichotomy Quantification via Geometric Embedding

**Shaobo Cui**
EPFL
shaobo.cui@epfl.ch

**Wenqing Liu**
EPFL
wenqing.liu@epfl.ch

**Yiyang Feng**
EPFL
yiyang.feng@epfl.ch

**Jiawei Zhou**
Stony Brook University
jiawei.zhou.1@stonybrook.edu

**Boi Faltings**
EPFL
boi.faltings@epfl.ch

## Abstract

Conditional dichotomy, the contrast between two outputs conditioned on the same context, is vital for applications such as debate, defeasible natural language inference, and causal reasoning. Existing methods that rely on semantic similarity often fail to capture the nuanced oppositional dynamics essential for these applications. Motivated by these limitations, we introduce a novel task, *Conditional Dichotomy Quantification* (ConDQ), which formalizes the direct measurement of conditional dichotomy and provides carefully constructed datasets covering debate, defeasible natural language inference, and causal reasoning scenarios. To address this task, we develop the Dichotomy-oriented Geometric Embedding (DoGE) framework, which leverages complex-valued embeddings and a dichotomous objective to model and quantify these oppositional relationships effectively. Extensive experiments validate the effectiveness and versatility of DoGE, demonstrating its potential in understanding and quantifying conditional dichotomy across diverse NLP applications. Our code and datasets are available at https://github.com/cui-shaobo/conditional-dichotomy-quantification.

## 1 Introduction

Conditional dichotomy refers to the contrast between two outputs that are conditioned on the same context, highlighting both their opposition and interconnectedness (Apothéloz et al., 1993; Hidey and McKeown, 2019). In many NLP tasks, the ability to generate and assess contrasting outputs conditioned on the same context is crucial for applications like debate (Chen et al., 2019; Liang et al., 2024), defeasible natural language inference (NLI) (Forbes et al., 2020; Rudinger et al., 2020), and causal reasoning (Kiciman et al., 2024; Cui et al., 2024). Despite its importance and broad applications, conditional dichotomy has not been thoroughly explored in the literature.

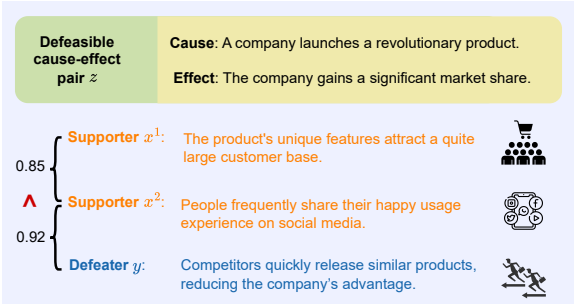



Figure 1: A motivational example illustrating the challenge of assessing conditional dichotomy through semantic similarity. The similarity score between two supporters (0.85) is lower than the similarity score between one supporter and one defeater (0.92).

An intuitive approach to measuring conditional dichotomy primarily relies on semantic textual similarity (STS) (Reimers and Gurevych, 2019; Gao et al., 2021), operating under the assumption that a lower similarity score indicates a higher degree of dichotomy. However, as illustrated in Figure 1, this assumption does not always hold. For the same cause-effect pair, the similarity score between two supporters (0.85) is unexpectedly lower than the similarity score between a supporter and a defeater (0.92).[1] This counterintuitive result demonstrates that using semantic textual similarity as an indirect measure of conditional dichotomy is both unreliable and ineffective, as it fails to capture the genuine dichotomy between conditioned outputs.

To bridge the gap in benchmarking, evaluation, and methods in conditional dichotomy, we introduce a novel task termed *Conditional Dichotomy Quantification* (ConDQ), which aims to directly measure the degree of dichotomy between two outputs conditioned on the same context. We instantiate this task across three representative scenarios,

[1]We refer to a *supporter* as an argument that reinforces a cause-effect relation, and a *defeater* as one that weakens it. The scores are computed by the state-of-the-art STS model AoE (Li and Li, 2024).



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7765–7791
July 27 - August 1, 2025 

as illustrated in Figure 2: (i) debate: dichotomy between supporting and opposing arguments for given debate topics (Chen et al., 2019; Liang et al., 2024); (ii) defeasible NLI: dichotomy between strengtheners and weakeners in defeasible natural language inference (Rudinger et al., 2020); (iii) causal reasoning: dichotomy between supporters and defeaters for given cause-effect pairs (Cui et al., 2024). We benchmark each scenario with datasets consisting of quadruples of (context, positive, negative, neutral).

To comprehensively evaluate the proposed task, we introduce two novel evaluation metrics: (i) *Dichotomy Consistency Frequency (DCF)*, which assesses whether embeddings preserve the relative positional relationships among positive, negative, and neutral arguments, i.e., relational consistency; and (ii) *Oppo-Angle*, which quantifies the angular separation between oppositional arguments, i.e., absolute opposition. These metrics address the limitations of conventional similarity measures by offering quantification of both relational consistency and absolute opposition, thus providing a holistic framework for assessing dichotomous structures in embedding spaces.

To overcome the limitations of STS-based methods, we propose the Dichotomy-oriented Geometric Embedding (DoGE) framework for quantifying conditional dichotomy. Inspired by previous works (Arora et al., 2017; Li and Li, 2024), DoGE adopts a *complex-valued* embedding framework in the mathematical sense: each sentence vector has real and imaginary components, offering a richer geometric space for modeling dichotomy. Moreover, DoGE introduces an innovative dichotomous objective that *geometrically* positions neutral arguments between positive and negative ones within the embedding space (Figure 4). This structure improves both the representational quality and the precision of dichotomy quantification.

Our extensive experiments across the debate, defeasible NLI, and causal reasoning scenarios demonstrate the effectiveness and versatility of DoGE. Compared to other embedding methods, DoGE achieves significant improvements in both DCF and Oppo-Angle. Moreover, visualizations of DoGE's embedding space reveal clear angular separations among positive, negative, and neutral arguments, highlighting its capability to disentangle and model dichotomous relationships effectively.

Our contributions are as follows:

1. **Formalization of the conditional dichotomy quantification task**: We introduce a novel NLP task that measures the nuanced opposition between outputs conditioned on a shared context, thereby establishing the foundation for studying oppositional perspectives in natural language.

2. **Novel benchmark datasets and evaluation metrics**: We develop datasets for debate, defeasible NLI, and causal reasoning scenarios, and introduce two new metrics (DCF and Oppo-Angle) to assess relational consistency and absolute opposition. Together, these resources establish a systematic evaluation framework for the proposed task.

3. **Proposal of the Dichotomy-oriented Geometric Embedding (DoGE) framework**: We present a novel embedding framework that operates within the complex-valued embedding space, incorporating a unique dichotomous objective to capture nuanced dichotomy. This framework delivers geometrically precise positioning of arguments and dynamically adapts to contextual variations.

4. **Extensive experiments with diverse embedding methods and backbones**: We conduct comprehensive experiments using diverse embedding methods and various backbone models across multiple scenarios. DoGE consistently outperforms strong baselines in quantifying conditional dichotomy, demonstrating both effectiveness and versatility.

## 2 Task: Conditional Dichotomy Quantification (ConDQ)

### 2.1 Formal Task Definition

The task of Conditional Dichotomy Quantification (ConDQ) aims to measure the dichotomy degree between two outputs, $X$ and $Y$, that are derived from the same context $Z$. Formally, the dichotomy degree is:

$$\Phi_{(XY|Z)} = f(\Delta_{(XY|Z)}) \quad (1)$$

where $\Phi_{(XY|Z)}$ represents the dichotomy degree between outputs $X$ and $Y$ given the context $Z$. The term $\Delta_{(XY|Z)}$ [2] represents the angular distance be-

[2] In this paper, $\Delta_{(XY|Z)}$ denotes the angular distance between $X$ and $Y$ conditioned on $Z$ in the real-valued space, while $\Gamma_{(XY|Z)}$ (defined in 4.1) denotes the distance between $X$ and $Y$ conditioned on $Z$ in the complex-valued space.

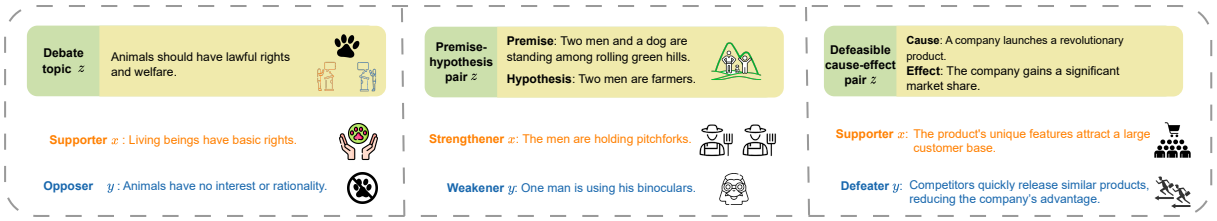



Figure 2: Instantiated scenarios under the umbrella of conditional dichotomy: debate (left), defeasible natural language inference (middle), and causal reasoning (right).

tween the embeddings of $X$ and $Y$ within the embedding space that is influenced by $Z$. The function $f$ maps this angular distance to a dichotomy degree, with larger distances reflecting stronger dichotomy.

ConDQ differs from related tasks by explicitly quantifying opposition between conditioned outputs, rather than simply measuring semantic similarity. A detailed comparison is provided in App. B.

## 2.2 Instantiated Scenarios

We define three concrete scenarios under the ConDQ framework, each capturing contextual dichotomy in a different domain.

**Scenario A: Supporting and Opposing Arguments in Debate**. This scenario focuses on measuring the dichotomous degree between a supporting argument $X$ and an opposing argument $Y$ given a debate topic $Z$. In Figure 2 (left), these arguments represent oppositional viewpoints on the issue of animals' lawful rights. The supporting argument emphasizes the *basic rights* of animals, while the opposing argument challenges their interest or rationality. We reconstruct PERSPECTRUM dataset (Chen et al., 2019) for this scenario.

**Scenario B: Strengtheners and Weakeners in Defeasible Natural Language Inference**. In this scenario, as shown in Figure 2 (middle), the shared context $Z$ consists of a premise and a hypothesis. The task measures the dichotomous degree between a strengthener argument $X$ and a weakener argument $Y$ for the given premise-hypothesis relationship. The opposition lies in how $X$ and $Y$ focus on the actions described, with $X$ supporting and $Y$ weakening the connection between the premise and the hypothesis. For example, the selected tools, pitchforks and binoculars, have opposite effects on justifying the inference. We evaluate this scenario using the $\delta$-NLI dataset (Rudinger et al., 2020).

**Scenario C: Supporting and Defeating Arguments in Causal Reasoning**. This scenario involves measuring the dichotomous degree between a supporter $X$ and a defeater $Y$ for a given cause-effect pair $Z$, as shown in Figure 2 (right). It captures how oppositional influences interact with a shared causal context. This scenario is supported with the $\delta$-CAUSAL dataset (Cui et al., 2024).

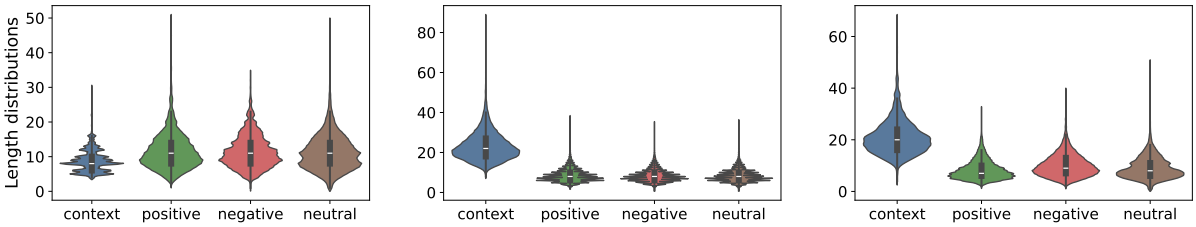
## 2.3 Evaluation Metrics

To evaluate the embedding space structure, we introduce a **neutral argument** $W$ for each context. $W$ is unaligned with positive $X$ or negative $Y$ arguments, ensuring $(X, Y)$'s distinction and positioning $W$ between them. We propose two metrics to assess conditional dichotomy from complementary perspectives: (i) DCF measures the relational consistency between dichotomous and non-dichotomous pairs; and (ii) Oppo-Angle directly quantifies explicit opposition degree.

**Dichotomy Consistency Frequency (DCF)**. DCF evaluates whether the embeddings preserve positional relationships among dichotomous and non-dichotomous pairs, i.e., relational consistency. Dichotomous pairs $(X, Y)$ are positive and negative arguments. We also generate neutral arguments $W$ detailed in § 3, and create non-dichotomous pairs $(X, W)$ and $(Y, W)$. Specifically, DCF measures the percentage of test samples where (i) the angular distance between positive and neutral embeddings is smaller than the angular distance between positive and negative embeddings, and (ii) the angular distance between negative and neutral embeddings is smaller than the angular distance between positive and negative embeddings. Mathematically,

$$\text{DCF} = \frac{100}{N}\sum_{i=1}^{N}\mathbb{1}\Big(\Delta_{(X_iY_i|Z_i)} > \Delta_{(X_iW_i|Z_i)} \wedge \Delta_{(X_iY_i|Z_i)} > \Delta_{(Y_iW_i|Z_i)}\Big) \quad (2)$$

Here, $N$ is the number of examples. $\Delta_{(X_iY_i|Z_i)}$ represents the angular distance between arguments $X_i$ and $Y_i$ conditioned on the same $Z_i$. The final metric is the computed ratio multiplied by 100




(a) Scenario A: Debate. (b) Scenario B: Defeasible NLI. (c) Scenario C: Causal Reasoning.

Figure 3: Sentence length distributions for contexts, positive, negative, and neutral arguments across datasets.

for intuitive interpretation. Submetrics include: (i) $\underline{\text{DCF}_{\text{positive}}}$: validates that the positive-neutral angular distance is smaller than the positive-negative angular distance; and (ii) $\underline{\text{DCF}_{\text{negative}}}$: validates that the negative-neutral angular distance is smaller than the positive-negative angular distance.

$$\text{DCF}_{\text{positive}} = \frac{100}{N}\sum_{i=1}^{N}\mathbb{1}\left(\Delta_{(X_iY_i|Z_i)} > \Delta_{(X_iW_i|Z_i)}\right)$$
$$\text{DCF}_{\text{negative}} = \frac{100}{N}\sum_{i=1}^{N}\mathbb{1}\left(\Delta_{(X_iY_i|Z_i)} > \Delta_{(Y_iW_i|Z_i)}\right) \quad (3)$$

**Direct Angular Quantification of Positive-Negative Embedding Opposition (Oppo-Angle).** Oppo-Angle directly quantifies the angular separation between positive and negative arguments, offering an explicit measure of absolute opposition. Unlike DCF, which evaluates relational consistency, Oppo-Angle focuses solely on the magnitude of dichotomous degree:

$$\text{Oppo-Angle} = \frac{100}{N}\sum_{i=1}^{N}\left(1 - \cos(\mathbf{E}_{X_i|Z_i}, \mathbf{E}_{Y_i|Z_i})\right) \quad (4)$$

Similarly, the factor of 100 scales the score to a more interpretable and readable range.

## 3 Supportive Datasets

### 3.1 Construction of Datasets

**Raw Dataset Collection.** Our dataset construction process starts from the raw data described in § 2.2. For each context $Z$, we gather two distinct sets of arguments: positive arguments $\mathcal{X}$ and negative arguments $\mathcal{Y}$. To ensure comprehensive coverage of oppositional perspectives, we generate all possible triples $(Z, X, Y)$ by selecting one $X$ from $\mathcal{X}$ and one $Y$ from $\mathcal{Y}$. Namely, the dataset encompasses all triples drawn from the Cartesian product $\{Z\} \otimes \mathcal{X} \otimes \mathcal{Y}$, encompassing a wide range of oppositional viewpoints for each context.

**Neutral Argument Collection.** Beyond positive and negative arguments, neutral arguments $W$ play a vital role in our evaluation metrics by offering a reference point. We generate $W$ for each triple $(Z, X, Y)$ via a meticulous three-step process: (i) We analyze the linguistic features of $X$ and $Y$ (including the number of words, noun chunks, and verb chunks) using the spaCy library (Honnibal and Montani, 2017); (ii) We use GPT-4o (OpenAI, 2023) to generate neutral arguments with word counts similar to $X$ or $Y$. This ensures that the model focuses on the semantic content rather than exploiting lexical cues like word counts. Furthermore, we randomly keep half of the noun and verb chunks from $X$ or $Y$ and prompt the model to generate arguments incorporating those chunks. This design enables the neutral arguments to mimic either $X$ or $Y$ in their forms (e.g., sentence lengths) and partially in their contents (e.g., shared nouns or verbs), rather than being entirely unrelated to the context, thus avoiding the encoding of spurious correlations; (iii) To ensure the neutrality of the generated arguments, two annotators independently verified them. Their high level of agreement confirms the reliability of these neutral arguments.

This three-step process guarantees the quality of neutral arguments and mitigates the risk of spurious correlations. Details on neutral argument generation and licensing information for datasets, tools, and software are provided in App. C and App. E.

| Statistic | Debate | Defeasible NLI | Causal Reasoning |
|---|---|---|---|
| # Overall | 95,524 | 440,744 | 47,518 |
| # Train | 58,058 | 8,462 | 14,008 |
| # Valid | 21,316 | 8,656 | 17,944 |
| # Test | 16,150 | 423,626 | 15,566 |
| avg. len(context) | 8.79 | 23.12 | 20.99 |
| avg. len(positive) | 11.59 | 8.52 | 8.35 |
| avg. len(negative) | 11.51 | 8.32 | 10.08 |
| avg. len(neutral) | 11.16 | 8.37 | 9.07 |

Table 1: Dataset statistics of each scenario.

### 3.2 Statistics of Collected Datasets

Table 1 provides a statistical summary of the collected benchmarks. Sentence length distribu-

tions are visualized in Figure 3. We find that (i) *Diverse length distributions across scenarios*: Across all scenarios, the length distributions are well-diversified, ensuring that the datasets cover a wide range of argument complexities and structures. This diversity is essential for robustly evaluating the proposed ConDQ task; and (ii) *Aligned length distributions of arguments within scenarios:* Within each scenario, the length distributions of positive, negative, and neutral arguments are closely aligned. This alignment prevents models from exploiting superficial length differences to distinguish between argument types, thereby enforcing a focus on genuine dichotomous relationships rather than superficial cues. More analysis of dataset statistics is presented in App. F.

## 4 Dichotomy-oriented Geometric Embedding (DoGE) Framework

The Dichotomy-oriented Geometric Embedding (DoGE) framework comprises several components: representing text in a complex-valued embedding space (§4.1), implementing a dichotomous objective for geometric positioning of different arguments in the complex-valued space (§4.2), incorporating contrastive learning to enhance opposition (§4.3), and establishing training and inference procedures (§4.4).

### 4.1 Representing Text in Complex-Valued Embedding Space

Existing embedding approaches predominantly rely on real-valued representations that encode semantic information. However, measuring dichotomy requires distinguishing not only surface semantic differences but also essential oppositions. Existing research shows that complex-valued embeddings are suitable for preserving essential information (Arora et al., 2017; Li and Li, 2024). Inspired by them, DoGE leverages a complex-valued embedding space that represents each sentence with both real and imaginary components, enhancing the model's ability to capture conditional dichotomy.

**Text Representation**. Each input text is encoded into a complex-valued embedding by first obtaining a $2d$-dimensional real vector $\mathbf{E} \in \mathbb{R}^{2d}$ from a Transformer encoder such as BERT (Devlin et al., 2019) or LLaMA (Touvron et al., 2023). We partition $\mathbf{E}$ into two $d$-dimensional components: the real part $\mathbf{E}^{\text{re}} = \mathbf{E}_{0:d} \in \mathbb{R}^d$ and the imaginary part $\mathbf{E}^{\text{im}} = \mathbf{E}_{d:2d} \in \mathbb{R}^d$. These components together define a complex-valued embedding.

**Distance in Complex-Valued Space**. Consider two embedding vectors in the complex-valued space $\mathbb{C}$:

$$\begin{aligned} \mathbf{X} &= \mathbf{a} + \mathbf{b}\mathrm{i} \in \mathbb{C} \\ \mathbf{W} &= \mathbf{c} + \mathbf{d}\mathrm{i} \in \mathbb{C} \end{aligned}. \tag{5}$$

The distance (Li and Li, 2024) between $X$ and $W$ in the complex-valued space $\Gamma_{(XW|Z)}$ is:[3]

$$\begin{aligned} \Gamma_{(XW|Z)} &= \text{abs}\left(\frac{\mathbf{X}}{\mathbf{W}} \times \frac{\sqrt{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2}}{\sqrt{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2}}\right) \\ &= \text{abs}\left[\frac{(\mathbf{a}\cdot\mathbf{c} + \mathbf{b}\cdot\mathbf{d}) + (\mathbf{b}\cdot\mathbf{c} - \mathbf{a}\cdot\mathbf{d})\mathrm{i}}{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2} \times \frac{\sqrt{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2}}{\sqrt{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2}}\right] \\ &= \text{abs}\left[\frac{(\mathbf{a}\cdot\mathbf{c} + \mathbf{b}\cdot\mathbf{d}) + (\mathbf{b}\cdot\mathbf{c} - \mathbf{a}\cdot\mathbf{d})\mathrm{i}}{\sqrt{(\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2)(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)}}\right] \end{aligned}. \tag{6}$$

### 4.2 Dichotomous Objective for Geometrical Positioning in Complex-Valued Space

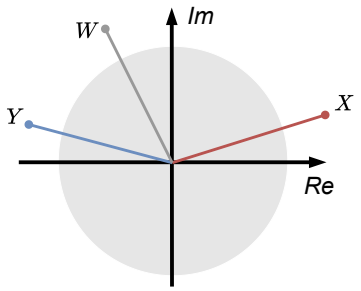



Figure 4: Illustration of the dichotomous objective: the neutral argument ($W$) is geometrically positioned between the positive ($X$) and negative ($Y$) arguments, establishing a geometrically balanced arrangement in the complex-valued embedding space. A low-dimensional case ($d = 1$) is shown for clarity, whereas actual embeddings are high-dimensional.

To ensure that embeddings reflect the intended geometric relationships among positive, negative, and neutral arguments, we introduce a dichotomous objective function. This objective geometrically positions neutral arguments between positive and negative arguments in the embedding space, as illustrated in Figure 4. Specifically, consider a quadruple $(Z, X, Y, W)$, where $X$ and $Y$ represent positive and negative arguments, respectively, conditioned on the context $Z$, and $W$ is a neutral argument. Given the context $Z$, the distances between these arguments in the complex-valued space are defined as: (i) $\Gamma_{(XY|Z)}$ (the complex-valued distance between $X$ and $Y$ conditioned on $Z$); (ii) $\Gamma_{(XW|Z)}$ (the complex-valued distance between

[3]Full proof is provided in App. G.

| Scenario | *Scenario A: Debate* | | | | *Scenario B: Defeasible NLI* | | | | *Scenario C: Defeasible Causality* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | DCF | $\text{DCF}_{\text{positive}}$ | $\text{DCF}_{\text{negative}}$ | Oppo-Angle | DCF | $\text{DCF}_{\text{positive}}$ | $\text{DCF}_{\text{negative}}$ | Oppo-Angle | DCF | $\text{DCF}_{\text{positive}}$ | $\text{DCF}_{\text{negative}}$ | Oppo-Angle |
| Open-source models | | | | | | | | | | | | |
| InferSent-*GloVe* | 36.19 | 64.22 | 61.47 | 1.58 | 23.11 | 49.83 | 49.06 | 0.39 | 26.71 | 57.75 | 48.05 | 0.44 |
| InferSent-*fastText* | 42.02 | 68.41 | 65.52 | 4.56 | 27.66 | 54.97 | 53.05 | 1.42 | 32.36 | 64.78 | 52.63 | 1.44 |
| USE | 16.53 | 46.39 | 43.37 | 3.31 | 18.07 | 48.54 | 45.35 | 1.01 | 13.54 | 42.19 | 34.19 | 0.46 |
| BERT baselines | | | | | | | | | | | | |
| BERT | 31.37 | 54.72 | 52.58 | 0.18 | 11.99 | 27.11 | 26.68 | 0.25 | 27.17 | 54.01 | 42.11 | 0.26 |
| CoSENT | 38.49 | 63.63 | 58.92 | 0.64 | 26.86 | 55.07 | 52.29 | 0.28 | 30.07 | 59.83 | 47.41 | 0.14 |
| SBERT | 31.61 | 58.07 | 55.02 | 1.50 | 22.89 | 51.29 | 48.50 | 0.64 | 22.68 | 51.61 | 41.23 | 0.43 |
| SimCSE | 30.59 | 61.51 | 52.89 | 2.78 | 13.91 | 44.85 | 39.83 | 0.93 | 25.15 | 66.25 | 42.54 | 1.30 |
| AoE | 26.27 | 55.35 | 49.34 | 0.48 | 24.02 | 51.63 | 47.07 | 0.10 | 30.09 | 63.45 | 44.98 | 0.11 |
| RoBERTa baselines | | | | | | | | | | | | |
| RoBERTa | 43.61 | 64.41 | 64.76 | 0.00 | 12.6 | 34.12 | 32.82 | 0.00 | 24.06 | 53.37 | 40.68 | 0.00 |
| SimCSE | 30.84 | 61.60 | 51.16 | 2.42 | 12.78 | 42.41 | 37.24 | 0.64 | 27.01 | 67.45 | 42.88 | 1.28 |
| LLaMA baselines | | | | | | | | | | | | |
| $\text{LLaMA-2}_{(7B)}$ | 30.46 | 50.50 | 51.65 | 16.99 | 21.25 | 39.72 | 39.22 | 8.65 | 32.80 | 58.45 | 45.81 | 5.67 |
| $\text{LLaMA-2}_{(13B)}$ | 47.42 | 65.04 | 64.13 | 11.24 | 30.27 | 43.65 | 43.54 | 4.59 | 34.05 | 55.53 | 45.85 | 2.56 |
| $\text{AoE}_{(7B)}$ | 38.92 | 58.82 | 55.10 | 14.85 | 20.01 | 41.13 | 37.14 | 8.22 | 27.20 | 57.13 | 40.13 | 4.03 |
| $\text{AoE}_{(13B)}$ | 44.88 | 62.28 | 60.21 | 9.73 | 28.72 | 42.45 | 40.63 | 3.20 | 30.89 | 54.43 | 42.01 | 1.58 |
| $\text{LLaMA-3.1}_{(8B)}$ | 39.81 | 58.73 | 56.13 | 10.86 | 21.7 | 38.20 | 37.56 | 5.33 | 26.38 | 45.22 | 39.15 | 2.67 |
| $\text{LLaMA-3.1}_{(70B)}$ | 34.47 | 55.15 | 52.98 | 13.74 | 15.95 | 37.27 | 34.67 | 6.83 | 25.23 | 46.13 | 40.43 | 3.84 |
| Our Method: DoGE (BERT version) with ablation versions. | | | | | | | | | | | | |
| DoGE (ours) | 46.97 $\pm$ 0.491 | 69.96 $\pm$ 0.975 | 63.43 $\pm$ 0.776 | 30.66 $\pm$ 6.671 | 41.72 $\pm$ 10.432 | 57.42 $\pm$ 9.596 | 57.44 $\pm$ 9.457 | 3.25 $\pm$ 0.766 | 67.59 $\pm$ 1.151 | 85.08 $\pm$ 0.845 | 77.34 $\pm$ 0.507 | 20.69 $\pm$ 1.821 |
| DoGE w/o DICT | 35.72 $\pm$ 2.154 | 63.19 $\pm$ 1.485 | 55.12 $\pm$ 1.398 | 82.72 $\pm$ 0.290 | 34.77 $\pm$ 1.355 | 54.10 $\pm$ 1.240 | 54.14 $\pm$ 0.941 | 42.47 $\pm$ 1.328 | 53.98 $\pm$ 3.349 | 77.12 $\pm$ 1.344 | 66.19 $\pm$ 2.508 | 103.80 $\pm$ 3.456 |
| DoGE w/o CL | 47.78 $\pm$ 0.584 | 70.95 $\pm$ 0.287 | 63.94 $\pm$ 0.605 | 18.02 $\pm$ 0.769 | 40.71 $\pm$ 4.636 | 57.21 $\pm$ 4.089 | 56.74 $\pm$ 4.167 | 2.71 $\pm$ 0.308 | 68.78 $\pm$ 0.570 | 85.59 $\pm$ 0.529 | 78.02 $\pm$ 0.408 | 19.58 $\pm$ 1.145 |
| Our Method: DoGE (RoBERTa version) with ablation versions. | | | | | | | | | | | | |
| DoGE (ours) | 55.93 $\pm$ 0.903 | 74.99 $\pm$ 0.664 | 68.72 $\pm$ 0.812 | 83.67 $\pm$ 3.126 | 47.27 $\pm$ 24.340 | 59.16 $\pm$ 25.820 | 58.74 $\pm$ 25.500 | 0.63 $\pm$ 0.471 | 76.55 $\pm$ 0.500 | 90.39 $\pm$ 0.543 | 83.11 $\pm$ 0.157 | 5.06 $\pm$ 0.568 |
| DoGE w/o DICT | 43.97 $\pm$ 2.740 | 63.97 $\pm$ 3.962 | 61.76 $\pm$ 2.845 | 31.59 $\pm$ 44.670 | 11.62 $\pm$ 1.048 | 20.19 $\pm$ 1.650 | 19.88 $\pm$ 1.610 | 0.01 $\pm$ 0.012 | 24.73 $\pm$ 1.561 | 40.04 $\pm$ 4.326 | 37.19 $\pm$ 1.817 | 0.00 $\pm$ 0.005 |
| DoGE w/o CL | 62.98 $\pm$ 1.776 | 81.42 $\pm$ 1.318 | 74.79 $\pm$ 1.224 | 3.62 $\pm$ 0.296 | 47.31 $\pm$ 24.440 | 59.39 $\pm$ 25.720 | 58.77 $\pm$ 25.440 | 0.67 $\pm$ 0.438 | 76.64 $\pm$ 1.557 | 90.35 $\pm$ 0.391 | 83.21 $\pm$ 1.512 | 2.08 $\pm$ 0.316 |

Table 2: Comparative study of different embedding methods on the conditional dichotomy quantification task (§ 5.2). *The last few rows of this table present the ablation study results,* in which key components of DoGE (our proposed method) are removed to illustrate their individual impact (§ 5.3).

$X$ and $W$ conditioned on $Z$); and (iii) $\Gamma_{(YW|Z)}$ (the complex-valued distance between $Y$ and $W$ conditioned on $Z$). The constraints that place $W$ between $X$ and $Y$ in the complex-valued space are:

$$\begin{cases} \Gamma_{(XW|Z)} < \Gamma_{(XY|Z)} \\ \Gamma_{(YW|Z)} < \Gamma_{(XY|Z)} \end{cases} \tag{7}$$

We encode these two constraints using a dichotomous loss:

$$\mathcal{L}_{\text{dichotomous}} = \sum \log\left(1 + e^{\frac{\Gamma_{(XW|Z)} - \Gamma_{(XY|Z)}}{\tau_{\text{dichotomous}}}}\right) + \sum \log\left(1 + e^{\frac{\Gamma_{(YW|Z)} - \Gamma_{(XY|Z)}}{\tau_{\text{dichotomous}}}}\right) \tag{8}$$

where $\tau_{\text{dichotomous}}$ is a temperature parameter. Minimizing this loss encourages the neutral argument to lie geometrically between the positive and negative arguments, ensuring a balanced and interpretable representation of the dichotomous structure.

### 4.3 Contrastive Learning Mechanism for Enhanced Opposition

While the dichotomous objective ensures proper geometric relations, it alone may not fully capture the intensity of oppositional relationships. To address this, we incorporate a contrastive loss that pushes positive and negative arguments farther apart, strengthening their separability.

For each pair of positive and negative samples $(X_i, Y_i)$ sharing context $Z_i$, we define the contrastive loss. Here, $W_j$ is a neutral sample. The contrastive loss is formulated as:

$$\mathcal{L}_{\text{cl}} = -\sum_b \sum_i^m \left( \log \left[ \frac{e^{\frac{\Delta_{(X_i Y_i|Z_i)}}{\tau_{\text{cl}}}}}{e^{\frac{\Delta_{(X_i Y_i|Z_i)}}{\tau_{\text{cl}}}} + \sum_{j \neq i}^m e^{\frac{\Delta_{(X_i W_j|Z_i)}}{\tau_{\text{cl}}}}} \right] + \log \left[ \frac{e^{\frac{\Delta_{(X_i Y_i|Z_i)}}{\tau_{\text{cl}}}}}{e^{\frac{\Delta_{(X_i Y_i|Z_i)}}{\tau_{\text{cl}}}} + \sum_{j \neq i}^m e^{\frac{\Delta_{(Y_i W_j|Z_i)}}{\tau_{\text{cl}}}}} \right] \right) \tag{9}$$

where $\tau_{\text{cl}}$ is a temperature hyperparameter. $m$ represents the number of samples in the $b$-th batch. Minimizing $\mathcal{L}_{\text{cl}}$ maximizes the angular distance between dichotomous pairs while minimizing the angular distance between non-dichotomous pairs, enabling more effective utilization of contrastive learning and enhancing the model's ability to capture and emphasize their dichotomous relationship.

### 4.4 Training and Inference of DoGE

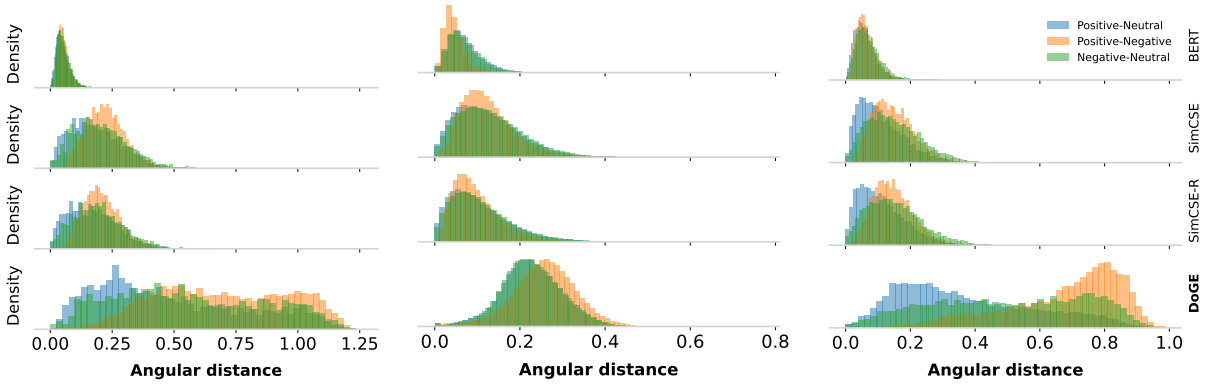
The final training objective combines the dichotomous and contrastive losses:

$$\mathcal{L} = w_1 \cdot \mathcal{L}_{\text{dichotomous}} + w_2 \cdot \mathcal{L}_{\text{cl}} \tag{10}$$

where $w_1$ and $w_2$ are hyperparameters balancing these two objectives. During inference, given two arguments $X$ and $Y$ conditioned on the same context $Z$, we measure their dichotomous degree as:

$$\Phi_{(XY|Z)} = 1 - \cos(\mathbf{E}_{X|Z}, \mathbf{E}_{Y|Z}) \tag{11}$$

where cos is the cosine similarity taken over the concatenated real and imaginary parts of the embeddings $\mathbf{E}_{X|Z}$ and $\mathbf{E}_{Y|Z}$. Each embedding $\mathbf{E}_{X|Z}$




(a) Scenario A: Debate. (b) Scenario B: Defeasible NLI. (c) Scenario C: Causal Reasoning.

Figure 5: Distributions of angular distance between positive, neutral, and negative argument embeddings (Positive-Neutral: $\Delta_{(XW|Z)}$, Positive-Negative: $\Delta_{(XY|Z)}$, Negative-Neutral: $\Delta_{(YW|Z)}$) across BERT, SimCSE (in BERT base backbone), SimCSE-R (SimCSE in RoBERTa base backbone), and DoGE under different scenarios.

is obtained by encoding the concatenated text of context $Z$ and argument $X$. A larger value of $\Phi_{(XY|Z)}$ indicates a stronger dichotomy. By integrating complex-valued embeddings, geometric constraints, and contrastive learning, DoGE ensures an interpretable quantification of conditional dichotomy, paving the way for analysis of oppositional perspectives in NLP tasks.

## 5 Empirical Study

In this section, we empirically evaluate our DoGE framework. We start with the experimental setup (§5.1), present the main results in comparison with baseline models (§5.2), conduct an ablation study to assess the contributions of individual components (§5.3), and finish with embedding space visualizations illustrating DoGE's capabilities (§5.4).

### 5.1 Experimental Setup

We evaluate DoGE against baselines including InferSent (Conneau et al., 2017), USE (Cer et al., 2018), SBERT (Reimers and Gurevych, 2019), SimCSE (Gao et al., 2021), CoSENT (Su, 2022), and AoE (Li and Li, 2024) across BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), LLaMA-2 (Touvron et al., 2023), and LLaMA-3 (Grattafiori et al., 2024) backbones. DoGE is trained for 5 epochs with hyperparameters $w_1 = 1.0$ and $w_2 = 3.0$ (see more details about experimental setup in App. H).

### 5.2 Main Results

**Superior Performance of DoGE**. Table 2 presents the performance of all models across the three scenarios using four metrics: DCF, $\text{DCF}_{\text{positive}}$, $\text{DCF}_{\text{negative}}$, and Oppo-Angle. The results are categorized under open-source models, BERT baselines, RoBERTa baselines, LLaMA baselines, and our proposed DoGE with its ablated variants. Key observations include: (i) *Superior DCF and Oppo-Angle scores*: DoGE surpasses all open-source, BERT, and RoBERTa baselines, achieving higher DCF and Oppo-Angle scores across all scenarios, indicating better relational consistency among positive, negative, and neutral arguments, as well as clearer angular separations between dichotomous pairs. Remarkably, these scores are even higher than, or comparable to, those of LLaMA variants, which have substantially more parameters; (ii) *Robustness across backbones*: Performance gains are consistent across different backbones, underscoring DoGE's versatility.

**Visualizing Angular Distance Distributions**. Figure 5 illustrates the distributions of angular distances between different pairs of argument embeddings (positive-negative, negative-neutral, and positive-neutral) for various models. The score distribution of DoGE distinctly positions the positive-negative angular distance towards the far right, whereas the other baseline models exhibit blended distributions without clear separation. This demonstrates that DoGE ultimately captures the expected dichotomous relationships, delving deeper into the distinction between oppositional arguments compared to existing methods.

Further analyses, including out-of-domain generalization and the role of context, are detailed in App. I and App. J, respectively.

### 5.3 Ablation Study

To assess component impact, we evaluate two ablated versions of DoGE (last rows of Table 2).

1. DoGE w/o DICT (without Dichotomous Objective): Removing the Dichotomous objective might increase the absolute magnitude of the dichotomous degree, as indicated by increased Oppo-Angle scores in BERT across all scenarios. However, this leads to a notable decrease in relational consistency, evidenced by DCF scores decreasing across all scenarios and backbone models. For instance, in Scenario A, DCF drops from 46.97 to 35.72 for the BERT version and from 55.93 to 43.97 for the RoBERTa one, indicating a loss in relational consistency.

2. DoGE w/o CL (without Contrastive Learning): While removing the Contrastive Learning component retains or slightly improves DCF, it generally leads to a substantial decrease in Oppo-Angle scores across most scenarios and backbone models, highlighting the role of Contrastive Learning in maintaining angular separations. In Scenario A, the Oppo-Angle score for BERT drops from 30.66 to 18.02; for RoBERTa, it falls sharply from 83.67 to 3.62.

The ablation results underscore the importance of both DICT and CL for balancing strong relational consistency with significant opposition magnitude, enabling DoGE to model and quantify dichotomous relationships collaboratively.

### 5.4 Visualization of Embeddings

Figure 6 presents the t-SNE projection of embeddings $\mathbf{E}_{X|Z}$, $\mathbf{E}_{Y|Z}$, and $\mathbf{E}_{W|Z}$ ($\in \mathbb{R}^{2d}$), produced by DoGE model for positive ($X$), negative (Y), and neutral ($W$) arguments, each conditioned on the debate topic $Z$ in the debate scenario. The visualization demonstrates that DoGE effectively captures the underlying argument dichotomy: (i) *In-context separation*: Within each debate topic $Z$, DoGE places the positive (●) and its paired negative (■) on opposite sides of the embedding space. The occasional red-blue overlap occurs only when points from *different* topics are projected close together (e.g., a positive from topic A near a negative from topic B). (ii) *Intermediary positioning of neutral arguments*: Neutral arguments (▲) are well positioned between positive (●) and negative (■) arguments, reflecting their intermediary stance; and (iii) *Consistency across scenarios*: The spatial patterns persist across scenarios, underscoring the robustness of DoGE. More visualizations are shown in App. J.

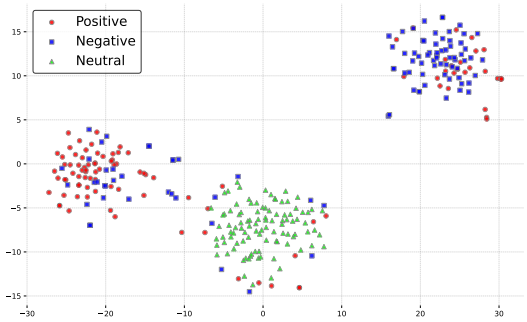



Figure 6: The t-SNE visualization of embeddings produced by DoGE (BERT version). Positive (●) and negative (■) arguments are distinctly separated, while neutral (▲) arguments are positioned between them. This highlights DoGE's effectiveness in representing dichotomous relationships.

## 6 Related Work

Our work is closely related to semantic textual similarity (STS) and contrastive learning. STS methods (Reimers and Gurevych, 2019; Gao et al., 2021; Li and Li, 2024) primarily focus on measuring semantic proximity between text pairs, using similarity scores to assess their degree of alignment. However, these methods are limited in capturing faithful dichotomous relationships, where dichotomous outputs are not merely dissimilar but fundamentally oppositional. Contrastive learning methods have shown success in learning textual representations that distinguish positive and negative pairs (Gao et al., 2021; Lee et al., 2021). However, these works often merely classify or contrasting examples rather than explicitly quantifying the opposition degree between text pairs. Unlike STS and contrastive learning, our work introduces a novel task: conditional dichotomy quantification, which focuses on context-conditioned opposition rather than semantic proximity or direct contrast, thereby expanding the scope of NLP research topics.

Recent studies have examined dichotomy in contexts such as defeasible NLI (Rudinger et al., 2020), causal reasoning (Cui et al., 2024), debate (Chen et al., 2019), and semantic opposition (de Silva and Dou, 2019; Vahtola et al., 2022). However, these works lack a standard measurement for the dichotomous relationships. Our proposed task, ConDQ, and embedding framework, DoGE, address this gap by offering a novel paradigm specifically designed to quantify the degree of opposition between

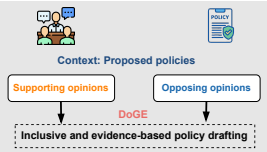
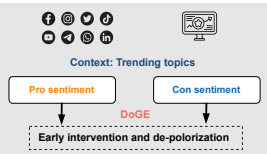
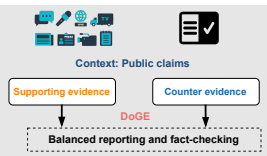
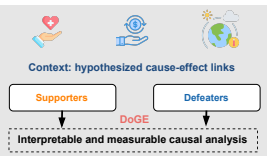
| Domain | Representative Scenario | | Societal Implications |
|---|---|---|---|
| Public Governance and Policy Making | Cluster public opinions that support or oppose a proposed regulation/policy during public hearings or citizen assemblies. |  | (i) Foster inclusive civic participation regarding public topics through balanced aggregation of competing viewpoints; (ii) Support more inclusive and evidence-based policy drafting by surfacing and contrasting diverse civic arguments. |
| Social Media and Online Discourse | Detect polarized clusters (pro and con) and disentangle mixed sentiments (positive and negative) during a controversial product launch or public topics in social media. |  | (i) Strengthen societal resilience to digital polarization by identifying polarized opinion clusters; (ii) Support responsible platform governance through early de-escalation of toxic or divisive dynamics in crisis-prone topics. |
| Journalism and Information Verification | Investigate a trending claim (e.g., on public health or election integrity) by retrieving its most robust counter-evidence and gauging the strength of opposition across published sources. |  | (i) Promote the integrity of journalism by enabling rapid access to the most contested claims and strongest rebuttals; (ii) Provide automated dashboards and alerts that highlight claims facing strong opposition across published sources. |
| Causal Analysis (Finance, Health, Climate, etc) | Contrast evidence (supporters and defeaters) that distinctly influences the causal relation in critical incidents (e.g., factors underlying a market crash, disease outbreak, or extreme weather). |  | (i) Clarify and accelerate the root-cause analysis under uncertainty, informing better public responses; (ii) Support reliable legal judgments by rapidly identifying competing (causal) explanations in incident reports. |

Table 3: Real-world applications of conditional dichotomy quantification and the DoGE framework, highlighting their societal implications across policy making, social media, journalism, and causal analysis.

two context-conditioned texts.

## 7 Broader Impact

**Key Applications**. By quantifying the *degree of opposition* between context-linked texts through geometric embeddings, our conditional dichotomy quantification task and DoGE framework unlock impactful applications across public policy, media, journalism, and causal analysis. As summarized in Table 3, our work empowers public policy making by summarizing contested opinions during citizen consultations, helps social media moderation via detecting and de-escalating early signs of polarization, supports journalism by retrieving robust counterclaims, and aids causal-risk analysis in evaluating competing causal evidence.

**Societal Implications**. These applications highlight broader societal benefits: facilitating inclusive and evidence-based policy drafting, strengthening digital resilience via earlier polarization detection, enhancing journalism integrity by considering the rebuttal, and strengthening causal accountability in high-stakes scenarios. Furthermore, our benchmarks and embeddings offer a new lens for assessing textual opposition and probing the robustness of LLMs under conflicting inputs. However, this ability to quantify the opposition degree could also be misused to algorithmically rank or amplify divisive content. We therefore strongly advocate for responsible usage, supported by human oversight and regulatory safeguards, especially before large-scale deployment of dichotomy-aware NLP systems.

## 8 Conclusion

In this paper, we formalize the *Conditional Dichotomy Quantification* (ConDQ) task, targeting the need to measure opposition between conditioned outputs. We introduce a suite of benchmark datasets spanning diverse scenarios, including debate, defeasible NLI, and causal reasoning, alongside novel evaluation metrics (DCF and Oppo-Angle) that effectively capture relational consistency and absolute opposition. Our proposed Dichotomy-oriented Geometric Embedding (DoGE) framework leverages complex-valued embeddings and a specialized dichotomous objective to accurately represent dichotomous relationships. Extensive experiments demonstrate that DoGE consistently outperforms existing approaches across diverse backbones and scenarios, highlighting its robustness and versatility. Our work lays a principled foundation for studying conditional dichotomy and supports more nuanced modeling of oppositional perspectives in language.

## Limitations

While our proposed task Conditional Dichotomy Quantification and the embedding framework DoGE demonstrate effectiveness in quantifying conditional dichotomy across various scenarios, there are limitations that are worthy of attention. First, the benchmarks and datasets used in the paper may not encompass all forms of dichotomous relationships found in machine learning and natural language processing tasks. Additionally, models' performance may vary when applied to languages other than English or to areas not represented in the investigated corpus. To address these limitations, future work could (i) expand the scope of benchmark datasets to include a wider array of dichotomous constructs and (ii) investigate cross-lingual and cross-domain performance to broaden the scope of our task and evaluate the generalizability of our framework.

## Ethical Consideration

In developing this novel conditional dichotomy task and our proposed DoGE framework, we have carefully considered the potential ethical implications associated with our work. One primary concern is the potential misuse of DoGE framework in amplifying or detecting polarized content. It is essential to emphasize that our work aims to enhance the understanding and analysis of contrasting perspectives in a responsible manner. We encourage users of our framework to utilize it ethically, particularly in contexts like social media analysis, political discussion, debate, or any other area where polarizing content may have significant impacts. Secondly, semantic opposition is always context-specific and culture-specific. This context-specific property could lead to overgeneralization when the DoGE is applied across different domains or languages. We recommend that users consider particular contextual and domain-specific analysis when using and interpreting the results provided by DoGE.

## References


Denis Apothéloz, Pierre-Yves Brandt, and Gustavo Quiroz. 1993. The function of negation in argumentation. *Journal of Pragmatics*, 19(1):23–38.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Daniel Cer, Yi Yang, Shiyang Kong, Nikhil Hua, Eng Siong Lim, Zhenya Yao, and Rui Zhang. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Shaobo Cui, Lazar Milikic, Yiyang Feng, Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Boi Faltings. 2024. Exploring defeasibility in causal reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6433–6452, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Nisansa de Silva and Dejing Dou. 2019. Semantic oppositeness embedding using an autoencoder-based learning model. In *Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30*, pages 159–174. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference*

*on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Christopher Hidey and Kathy McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1756–1767, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.*, 9(3):90–95.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*. Featured Certification.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. Learning dense representations of phrases at scale. 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)*, pages 6634–6647, Online. Association for Computational Linguistics.

Xianming Li and Jing Li. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand. Association for Computational Linguistics.

Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. 2024. Debatrix: Multi-dimensional debate judge with iterative chronological analysis based on LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14575–14595, Bangkok, Thailand. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Wes McKinney. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference 2010 (SciPy 2010), Austin, Texas, June 28 - July 3, 2010*, pages 56–61. scipy.org.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Jianlin Su. 2022. Cosent (1): A more effective sentence vector scheme than sentence bert.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. 2022. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new SemAntoNeg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

## A Appendix Contents

The appendix is organized as follows:

- **Comparison with Related Tasks** (App. B): situates ConDQ with respect to Semantic Textual Similarity and Contrastive Learning, outlining conceptual and empirical distinctions.
- **Responsible NLP Research** (App. C): enumerates datasets, backbone models, and software artifacts (App. C.1); details content-safety checks (App. C.2).
- **Random Baseline Discussion** (App. D): derives theoretical results for the random baseline.
- **Neutral Argument Acquisition** (App. E): details our differentiated strategy for constructing neutral arguments (using GPT-4o for validation/test sets and cross-scenario sampling for training) to mitigate stylistic overfitting; includes empirical analysis and motivating examples.
- **Dataset Statistics** (App. F): reports corpus-level distributions and verifies the neutrality of constructed neutral argument sets.
- **Complex-Valued Distance Derivation** (App. G): presents a detailed proof of the distance between vectors in the complex-valued space.
- **Experimental Setup** (App. H): specifies backbone architectures, hardware, hyperparameters, and training schedules.
- **Out-of-Domain Evaluation** (App. I): evaluates the generalizability of DoGE across scenarios by training on one scenario and testing on another, with findings highlighting transferability and domain-specific challenges.
- **Role of Context** (App. J): provides illustrative examples (App. J.1), ablation studies without context (App. J.2), and t-SNE visualizations (App. J.3).
- **List of Notations** (App. K): summarizes all symbols and variables for quick reference.

## B Comparison with Related Tasks

We compare ConDQ with related tasks in Table 4. The tasks most related to our conditional dichotomy quantification are *semantic textual similarity (STS)* and *contrastive learning for NLP (CL-NLP)*.

STS aims to quantify the semantic similarity between two texts, often using embeddings or similarity scores. STS primarily focuses on identifying alignments and semantic overlaps between texts, such as paraphrasing or rephrasing. Unlike STS, ConDQ measures the degree of opposition between two outputs conditioned on the same context, making it particularly suited for applications that require understanding oppositional arguments, such as debate, legal reasoning, social media moderation, and public policy drafting.

CL-NLP enhances representation learning by distinguishing between positive and negative pairs, improving model robustness and discriminative capabilities. While CL-NLP emphasizes learning effective representations through contrasting examples, it does not specifically measure the degree of opposition between texts. ConDQ differs by explicitly measuring how two arguments oppose each other conditioned on the same context, emphasizing the dichotomous nature of the arguments rather than merely classifying or contrasting samples. This unique focus on measuring dichotomy fills a gap left by CL-NLP, where the central interest lies in the representation quality of dichotomy, rather than directly evaluating how two conditioned arguments contrast with each other.

## C Responsible NLP Research

### C.1 Artifacts

We provide the datasets and software we use in Table 5. Our use of these artifacts (packages, models) is consistent with their intended use.

### C.2 Content Check

In accordance with ethical considerations mentioned after the limitation section, we carefully examined our collected datasets to ensure that they do not contain personal identifiable information or content that could be deemed offensive. Specifically, the original sources of our benchmark datasets (debate, defeasible NLI, and causal reasoning) are drawn from publicly available resources. Based on these resources, we further collected neutral arguments using the GPT-4o model. During data preparation, we closely inspected samples for

| Aspect | **Conditional Dichotomy Quantification (ConDQ)** | Semantic Textual Similarity (STS) | Contrastive Learning for NLP |
|---|---|---|---|
| Goal | Measuring the dichotomous degree between two outputs conditioned on the same input. | Measuring how semantically similar two pieces of text are. | Learning representations by contrasting positive and negative pairs of samples. |
| Problem definition | The task of Conditional Dichotomy Quantification (ConDQ) aims to measure the dichotomy degree between two outputs, $X$ and $Y$, that are derived from the same context $Z$. Formally, the dichotomy degree is: $\Phi_{(XY\|Z)} = f(\Delta_{(XY\|Z)})$ where $\Phi_{(XY\|Z)}$ represents the dichotomy degree between outputs $X$ and $Y$ given the context $Z$. The term $\Delta_{(XY\|Z)}$ represents the angular distance between the embeddings of $X$ and $Y$ within the embedding space that is influenced by $Z$. The function $f$ maps this angular distance to a dichotomy degree, with larger distances reflecting stronger dichotomy. | Given two sentences $x$ and $y$, STS models learn a similarity function $\text{Sim}(x, y) = \text{sim}(f(x), f(y))$, where $f$ is a sentence encoder and $\text{sim}(\cdot, \cdot)$ measures semantic closeness (e.g., cosine similarity). The model is trained to align $\text{Sim}(x, y)$ with human-annotated similarity scores. | Given an anchor $x$, a positive $x^+$, and negatives $x^-$, an encoder model is trained with various formulations of contrastive learning loss (Khosla et al., 2020), pulling the anchor toward its positive and pushing it away from all negatives. |
| Challenges | (i) Measuring contrast instead of similarity; (ii) Lack of benchmarks and specialized embedding methods. | (i) Accurately gauging semantic similarity; (ii) Dealing with varying levels of semantic overlap. | (i) Creating effective contrastive pairs; (ii) Maintaining meaningful feature representation while distinguishing between classes. |
| Applications | (i) Summarizing opposing arguments in public policy consultations; (ii) Detecting early signs of polarization in social media moderation; (iii) Retrieving counterclaims to improve journalistic balance; (iv) Comparing conflicting causal evidence in scientific or risk analysis. | (i) Identifying semantically equivalent or similar sentences across corpora; (ii) Ranking candidate responses or documents by relevance in retrieval and QA systems; (iii) Evaluating the quality of paraphrase outputs. | (i) Unsupervised pretraining of sentence and document encoders; (ii) Structuring embedding spaces for tasks like clustering, retrieval, and semantic search. |
| Contributions | ConDQ uniquely explores the inherent contrasts between dual outputs, introducing the novel conditional semantic textual dichotomousness task. It enriches research in natural language inference, facilitating dichotomous content measurement across various domains, such as debate, legal reasoning, social media moderation, public policy drafting, and causal analysis. | | |

Table 4: Comparison of ConDQ with Semantic Textual Similarity (STS) and Contrastive Learning for NLP (CL-NLP) across key dimensions: goals, problem formulations, challenges, and applications. Unlike existing tasks that focus on similarity or representation learning, ConDQ introduces a novel perspective by directly quantifying the degree of opposition between conditioned texts, enabling new capabilities in domains such as public policy, journalism, and causal analysis.

any unique identifiers (e.g., names, addresses, or personal attributes that could identify an individual) and removed or masked such information where necessary. Furthermore, we manually reviewed the data to identify and remove potentially hateful, harassing, or otherwise harmful content, using a keyword-based search to flag offensive language. Any instances detected as potentially offensive were removed. This ensures that the curated datasets remain free from offensive content and do not infringe on individuals' privacy.

Based on our efforts in collecting the dataset, we believe that we provide a dataset free from offensive content and suitable for research purposes.

# D Discussion on the Random Baseline Performance

The random baseline performance arises from the inherent structure of the *Conditional Dichotomy Quantification* task and the evaluation metrics used, specifically the Dichotomy Consistency Frequency (DCF) metric.

**Task Setup and Metric Definition**. For a given context $Z$, recall that there are three types of arguments:

- Positive argument ($X$): An argument that supports the context (a claim or hypothesis).
- Negative argument ($Y$): An argument that opposes the context (a claim or hypothesis).

| Artifacts/Packages/Models | Citation | Link | License |
|---|---|---|---|
| | | *Artifacts(datasets/benchmarks).* | |
| $\delta$-SNLI | (Rudinger et al., 2020) | https://github.com/rudinger/defeasible-nli | MIT License |
| *delta*-CAUSAL | (Cui et al., 2024) | https://github.com/cui-shaobo/defeasibility-in-causality | MIT License |
| PERSPECTRUM | (Chen et al., 2019) | https://github.com/CogComp/perspectrum | Missing |
| | | *Packages* | |
| PyTorch | (Paszke et al., 2019) | https://pytorch.org/ | BSD-3 License |
| transformers | (Wolf et al., 2020) | https://huggingface.co/docs/transformers/index | Apache License 2.0 |
| Accelerate | (Gugger et al., 2022) | https://huggingface.co/docs/accelerate/index | Apache License 2.0 |
| nltk | (Bird and Loper, 2004) | https://www.nltk.org/ | Apache License 2.0 |
| numpy | (Harris et al., 2020) | https://numpy.org/ | BSD License |
| pandas | (McKinney, 2010) | https://pandas.pydata.org/ | BSD 3-Clause License |
| matplotlib | (Hunter, 2007) | https://matplotlib.org/ | BSD compatible License |
| seaborn | (Waskom, 2021) | https://seaborn.pydata.org/ | BSD 3-Clause License |
| openai-python | (OpenAI, 2023) | https://pypi.org/project/openai/ | Apache-2.0 license |
| | | *Backbone Models* | |
| BERT | (Devlin et al., 2019) | https://huggingface.co/google-bert/bert-base-uncased | Apache-2.0 license |
| RoBERTa | (Liu et al., 2020) | https://huggingface.co/FacebookAI/roberta-base | MIT License |
| InferSent | (Conneau et al., 2017) | https://github.com/facebookresearch/InferSent | Attribution-NonCommercial 4.0 |
| SBERT | (Reimers and Gurevych, 2019) | www.sbert.net | Apache-2.0 License |
| SimCSE | (Gao et al., 2021) | https://github.com/princeton-nlp/SimCSE | MIT License |
| CoSENT | (Su, 2022) | https://huggingface.co/shibing624/text2vec-base-chinese | Apache license 2.0 |
| USE | (Cer et al., 2018) | https://huggingface.co/Dimitre/universal-sentence-encoder | CC 4.0 License |
| AoE | (Li and Li, 2024) | https://github.com/SeanLee97/AnglE | MIT License |

Table 5: Summary of the datasets, major software packages, and backbone models we use in this paper.

- Neutral argument ($W$): An argument that is relevant to the context (a claim or hypothesis) but does not directly support or oppose it.

The DCF metric evaluates whether the relationships among these arguments in the embedding space adhere to the expected dichotomous structure:

1. The angular distance between the positive and neutral arguments is smaller than the angular distance between the positive and negative arguments:

$$\Delta_{(XW|Z)} < \Delta_{(XY|Z)}. \tag{12}$$

2. The angular distance between the negative and neutral arguments is smaller than the angular distance between the positive and negative arguments:

$$\Delta_{(YW|Z)} < \Delta_{(XY|Z)}. \tag{13}$$

The DCF score reflects the percentage of instances in which both conditions are satisfied.

**Random Guessing and Baseline Performance**. Under a random baseline, the embedding space does not capture meaningful structures, and the angular distances among $X$, $Y$, and $W$ are random. Consequently, there are three distinct relational configurations—defined as events A, B, and C—which would be equally likely to occur by chance:

- Event $A$ represents the case where the angular distance between $X$ and $Y$ is greater than both the angular distance between $Y$ and $W$ and the angular distance between $X$ and $W$.
- Event $B$ represents the case where the angular distance between $Y$ and $W$ is greater than both the angular distance between $X$ and $Y$ and the angular distance between $X$ and $W$.
- Event $C$ represents the case where the angular distance between $X$ and $W$ is greater than both the angular distance between $X$ and $Y$ and the angular distance between $Y$ and $W$.

The probabilities of these events are defined as follows:

$$\begin{aligned} P(A) &= P(\Delta_{(XY|Z)} > \Delta_{(YW|Z)}, \Delta_{(XY|Z)} > \Delta_{(XW|Z)}), \\ P(B) &= P(\Delta_{(YW|Z)} > \Delta_{(XY|Z)}, \Delta_{(YW|Z)} > \Delta_{(XW|Z)}), \\ P(C) &= P(\Delta_{(XW|Z)} > \Delta_{(XY|Z)}, \Delta_{(XW|Z)} > \Delta_{(YW|Z)}) \end{aligned} \tag{14}$$

Since events $A$, $B$, and $C$ are mutually exclusive (i.e., they cannot occur simultaneously), the sum of their probabilities equals 1:

$$P(A) + P(B) + P(C) = 1 \tag{15}$$

Due to symmetry (i.e., equal chance of occurrence), we have:

$$P(A) = P(B) = P(C) = \frac{1}{3} \tag{16}$$

Thus, the probability of satisfying the dichotomous structure purely by chance is:

$$P(A) = \frac{1}{3} \tag{17}$$

It means that the DCF metric for the random baseline performance is 33.33. The random baseline performance establishes a lower bound for

methods on the DCF metric. Models must outperform this baseline to demonstrate their ability to encode and preserve the dichotomous relationships among positive, negative, and neutral arguments. This provides a robust framework for evaluating the quality of embeddings in capturing nuanced oppositional structures.

# E Neutral Argument Acquisition

## E.1 Overview of Our Differentiated Approach

In our dataset construction, we treat training, validation, and test sets differently to ensure robustness and generalization. For the test and validation sets, neutral samples are generated using the GPT-4o model, while for the training set, neutral samples are constructed from the positive and negative examples of other scenarios.

This design aims to prevent models from overfitting by learning shallow patterns of GPT-4o's generation style. By using neutral samples from other scenarios for training, we introduce greater diversity and complexity, encouraging the model to focus on nuanced distinctions among positive, negative, and neutral contexts rather than relying on superficial cues.

## E.2 Motivation of Our Differentiated Approach

**Neutral Arguments for Validation and Testing**. We prompt GPT-4o to generate high-quality and challenging neutral arguments. Specifically, (i) we do not provide contexts in the prompts since it may hinder GPT-4o from generating high-quality neutral arguments. In fact, we observed that when including contexts in the prompts, even with explicit instructions requiring neutrality, the generated arguments tend to contain supportive information for the given contexts (e.g., premise-hypothesis pairs, cause-effect relationships, debate topics) instead of being truly neutral. (ii) Including word chunks from the positive or negative arguments makes the generated arguments resemble positive/negative arguments, thus increasing task complexity. This step ensures that the model cannot rely solely on lexical differences to distinguish neutrality.

**Neutral Arguments for Training**. To justify our dataset construction, we investigate the impact of GPT-4o-generated neutral arguments on model performance. When these arguments are used for both training and testing, the model tends to identify superficial patterns and stylistic consistencies rather than meaningful conditional dichotomous structures. In Table 6, even after shuffling arguments at both the column level (interchanging positive, negative, and neutral positions inside each column) and the word level (shuffling the order of the words), the model's performance remained significantly above the random baseline. [4] This suggests that GPT-4o's neutral outputs share predictable syntactic or stylistic cues that the model can exploit, leading to overfitting. By incorporating neutral samples from other scenarios during training, we introduce diversity that dilutes these spurious patterns, ultimately improving model robustness and ensuring that the model learns genuine conditional dichotomy rather than relying on superficial signals.

| **Setup** | DCF | $\text{DCF}_{\text{positive}}$ | $\text{DCF}_{\text{negative}}$ |
|---|---|---|---|
| Origin | 94.03 | 95.17 | 95.16 |
| Shuffled Column | 74.74 | 88.08 | 78.84 |
| Shuffled Words | 68.58 | 78.88 | 77.03 |
| Random | 33.33 | 50.00 | 50.00 |

Table 6: Evaluation results of DoGE (BERT version) trained on GPT-4o-generated neutral arguments across different settings (original, column-shuffled, word-shuffled). Shuffling reduces performance, but it is still far above random, highlighting the risk of overfitting to superficial patterns of GPT-generated neutral arguments.

## E.3 Prompt Templates for Generating Neutral Arguments in Validation and Test Sets

We employ zero-shot prompting with the GPT-4o (OpenAI, 2023) model to generate neutral arguments. These arguments are constrained to contain word counts similar to those of positive or negative arguments. We randomly retain half of the noun and verb chunks from positive/negative examples and instruct the model to incorporate these into the neutral output. This strategy increases the complexity of the generated text and prevents the model from exploiting simple lexical cues.

## E.4 Neutral Argument Collection from Other Scenarios for Training Set

Neutral samples within the training set are sourced from human-annotated examples in other datasets, selected from diverse scenarios to ensure a balanced representation of neutrality. This approach

[4] the discussion about random baseline is presented in App. D.

aims to improve the semantic diversity in the training data, which helps mitigate the model's tendency to learn superficial patterns that are specific to GPT-generated content.

For example, when constructing the training dataset A, we first determine the length of a positive sample (e.g., 10 words). Next, a sentence of approximately 10 words is randomly selected from the positive or negative samples in datasets B or C. The same procedure is applied to the negative sample. Therefore, for each positive-negative pair in dataset A, two neutral sentences are generated: one with a length close to the positive sample and another with a length close to the negative sample.

Besides, the much larger test set (423,626 instances) compared to the training set (8,462 instances) in *Defeasible NLI* is intentional and is rooted in both data provenance and methodological aims. First, neutral arguments in the *Defeasible NLI* track are created by sampling syntactic chunks from the *Debate* and *Causal Reasoning* collections. Because these source corpora are modest in size, enlarging the training split would force heavy re-use of identical neutral sentences; this repetition would reduce lexical diversity and increase the risk of overfitting to surface patterns. Second, a compact training set incentivizes models to learn generalizable geometric regularities, whereas the expansive test set supplies a statistically robust basis for judging how well a model discriminates fine-grained dichotomies.

## F Statistics of Datasets

Figure 7 illustrates the distribution of angular distance differences between positive-neutral ($\Delta_{(XW|Z)}$) and negative-neutral ($\Delta_{(YW|Z)}$), i.e., $\Delta_{(XW|Z)} - \Delta_{(YW|Z)}$ across three distinct scenarios: debate, defeasible NLI, and causal reasoning. The embeddings are generated using DoGE (BERT) and DoGE (RoBERTa). Across all scenarios and backbone models, the distributions are tightly centered around zero, confirming that neutral embeddings are not systematically closer to either side. This balance ensures that models cannot exploit superficial alignment biases and that the evaluation of dichotomy genuinely reflects semantic contrasts rather than dataset artifacts. The consistency across both BERT and RoBERTa variants of DoGE further supports the robustness of this design.

## G Derivation of Distance in Complex-Valued Space

We derive the distance between two vectors in the complex-valued space following (Li and Li, 2024). We first need to represent the vectors as complex-valued numbers. This allows us to compute the angle between them efficiently by leveraging properties of complex-valued numbers in polar coordinates. Consider two vectors, $\mathbf{X}$ and $\mathbf{W}$, each represented as a complex-valued vector:

$$\mathbf{X} = \mathbf{a} + \mathbf{b}\mathrm{i}, \quad \mathbf{W} = \mathbf{c} + \mathbf{d}\mathrm{i} \tag{18}$$

where:

- $\mathbf{a}$ and $\mathbf{b}$ are the real and imaginary parts of $\mathbf{X}$, respectively.
- $\mathbf{c}$ and $\mathbf{d}$ are the real and imaginary parts of $\mathbf{W}$, respectively.
- $\mathrm{i}$ is the imaginary unit, $\mathrm{i} = \sqrt{-1}$.

In polar coordinates, the magnitude of a complex-valued number is given by:

$$r_{\mathbf{X}} = \sqrt{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2}, \quad r_{\mathbf{W}} = \sqrt{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2} \tag{19}$$

These represent the lengths of the vectors $\mathbf{X}$ and $\mathbf{W}$, respectively, in the complex plane. To compute the angular difference directly, we divide $\mathbf{X}$ by $\mathbf{W}$:

$$\frac{\mathbf{X}}{\mathbf{W}} = \frac{\mathbf{a} + \mathbf{b}\mathrm{i}}{\mathbf{c} + \mathbf{d}\mathrm{i}} \tag{20}$$

Using the division rule for complex-valued numbers, we obtain:

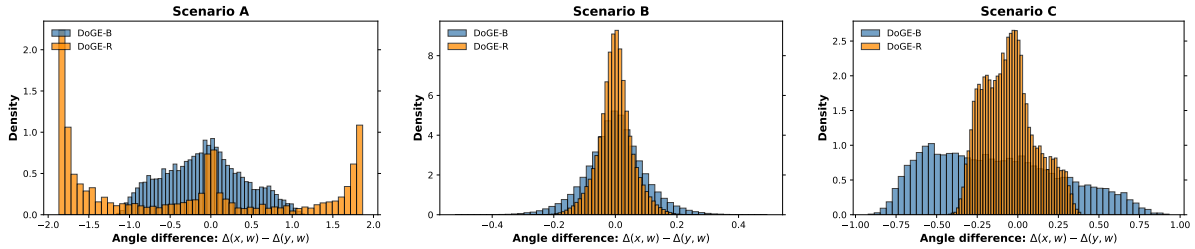
$$\frac{\mathbf{X}}{\mathbf{W}} = \frac{(\mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{d}) + (\mathbf{b} \cdot \mathbf{c} - \mathbf{a} \cdot \mathbf{d})\mathrm{i}}{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2} \tag{21}$$

The result of this division is another complex-valued number. The real part of this complex-valued number corresponds to the cosine of the angle difference, and the imaginary part corresponds to the sine of the angle difference. Next, we normalize the result to remove the influence of the magnitudes of $\mathbf{X}$ and $\mathbf{W}$. To do this, we divide by the magnitudes $r_{\mathbf{X}}$ and $r_{\mathbf{W}}$, which we computed earlier:

$$\frac{\mathbf{X}}{\mathbf{W}} = \gamma \Delta\theta_{XW} \tag{22}$$

where:

$$\gamma = \frac{r_{\mathbf{X}}}{r_{\mathbf{W}}} = \frac{\sqrt{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2}}{\sqrt{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2}} \tag{23}$$




(a) Scenario A (b) Scenario B (c) Scenario C

Figure 7: Distribution of the angular distance differences between positive-neutral and negative-neutral pairs, computed as $\Delta_{(XW|Z)} - \Delta_{(YW|Z)}$ across three conditional dichotomy scenarios: debate (Scenario A), defeasible NLI (Scenario B), and causal reasoning (Scenario C). Results are based on DoGE embeddings using BERT (blue) and RoBERTa (orange) backbones. Distributions centered around zero indicate that, at the dataset level, neutral arguments are not biased toward either the positive or negative side. This supports the neutrality and balance of the constructed neutral set.

Now, we compute the normalized angle difference $\Delta\theta_{XW}$ by normalizing the complex-valued division:

$$\begin{aligned}\Delta\theta_{XW} &= \text{abs}\left[\frac{1}{\gamma}\cdot\frac{(\mathbf{a}\cdot\mathbf{c}+\mathbf{b}\cdot\mathbf{d})+(\mathbf{b}\cdot\mathbf{c}-\mathbf{a}\cdot\mathbf{d})\text{i}}{(\|\mathbf{c}\|^2+\|\mathbf{d}\|^2)}\right] \\ &= \text{abs}\left[\frac{(\mathbf{a}\cdot\mathbf{c}+\mathbf{b}\cdot\mathbf{d})+(\mathbf{b}\cdot\mathbf{c}-\mathbf{a}\cdot\mathbf{d})\text{i}}{\sqrt{(\|\mathbf{c}\|^2+\|\mathbf{d}\|^2)(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2)}}\right]\end{aligned} \tag{24}$$

This gives us the angular distance in the complex-valued form as presented in Equation 6.

## H More Details of Experimental Setup

**Backbone Models and Hardware**. We conduct a comprehensive training of DoGE using two pre-trained backbone models: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2020). For both models, the initial learning rate is set to $2 \times 10^{-5}$. The computing infrastructure is as follows: The CPU model is an AMD EPYC 7543 32-Core Processor. The GPU computing model is NVIDIA A100-SXM4-80GB. For DoGE (BERT), the model has 110 million parameters, and it takes 10 minutes to train for 5 epochs on each dataset. Similarly, for DoGE (RoBERTa), the model has 125 million parameters, and it takes around 10 minutes to train for 5 epochs on each dataset. Therefore, the overall GPU computation time is 1 hour.

**Input Format**. During the training phase, the input consists of combinations of context, positive, neutral, and negative examples. Specifically, for each sample, the input is structured as ["context" + "positive"], ["context" + "neutral"], ["context" + "negative"].

**Training Schedule**. The weights $w_1$ and $w_2$ in Equation 10 are set to 1.0 and 3.0, respectively. These hyperparameters are chosen based on preliminary experiments to optimize model performance. We set the batch size to 256 for training and 32 for testing, conducting 5 training epochs for DoGE and its ablated versions (DoGE w/o DICT and DoGE w/o CL). Evaluation is performed using the model from the final epoch. A larger training batch size facilitates Contrastive Learning for dichotomy, enhancing the model's ability to effectively discern differences between dichotomous and non-dichotomous pairs.

**Random Seeds and Reporting**. To ensure the stability and robustness of the results, we train the model using three distinct random seeds: 42, 1015, and 6900. The average performance across these runs is reported, along with the standard deviation to account for any variability in the results.

## I Out of Domain Evaluation

### I.1 Setup

To further assess the robustness and generalizability of the DoGE framework, we conduct out-of-domain (OOD) evaluations. In these experiments, we train DoGE on one scenario and test it on a different scenario. This setup probes whether the model can capture the essence of dichotomous relationships beyond the domain it was originally trained on, thereby evaluating the model's ability to generalize across diverse contextual settings.

### I.2 Results and Analysis

Table 7 reports OOD performance. We have two observations: (i) When trained on Scenario B and tested on Scenario C, it achieves relatively high DCF and Oppo-Angle scores, indicating that

| Train Scenario | Test Scenario | Model | DCF | $DCF_{positive}$ | $DCF_{negative}$ | Oppo-Angle |
|---|---|---|---|---|---|---|
| Scenario A | Scenario B | DoGE | 9.82 ± 0.135 | 20.55 ± 0.233 | 20.46 ± 0.255 | 0.01 ± 0.008 |
| Scenario A | Scenario C | DoGE | 24.30 ± 0.631 | 41.30 ± 2.029 | 41.86 ± 0.493 | 2.50 ± 0.951 |
| Scenario B | Scenario A | DoGE | 43.49 ± 0.625 | 63.36 ± 1.037 | 62.95 ± 0.882 | 0.04 ± 0.014 |
| Scenario B | Scenario C | DoGE | 51.52 ± 4.675 | 76.13 ± 3.174 | 66.57 ± 3.164 | 0.79 ± 0.215 |
| Scenario C | Scenario A | DoGE | 49.10 ± 0.847 | 71.56 ± 0.249 | 64.54 ± 0.981 | 2.94 ± 0.279 |
| Scenario C | Scenario B | DoGE | 17.86 ± 1.610 | 36.11 ± 2.238 | 33.42 ± 2.001 | 0.92 ± 0.143 |

Table 7: Out-of-domain evaluation results. The metrics include Dichotomy Consistency Frequency (DCF), Positive Consistency ($DCF_{positive}$), Negative Consistency ($DCF_{negative}$), and Opposite Angle Scale (Oppo-Angle). Higher values across these metrics indicate better generalization and the ability to capture dichotomous relationships.

the representations learned for defeasible NLI can transfer effectively to the causal reasoning scenario; (ii) Conversely, certain scenario shifts (e.g., Scenario A to Scenario B) pose greater challenges, as reflected in lower DCF values.

## J Role of Context in Conditional Dichotomy Quantification

Context plays an essential role in transforming otherwise independent statements into dichotomous arguments. Without a shared contextual framework, positive and negative statements may not be inherently oppositional, even if they express divergent viewpoints. In contrast, when contextualized, these same statements can form clearly oppositional positions that either support or challenge a given premise, hypothesis, or causal relationship.

### J.1 Motivation Example for the Role of Context

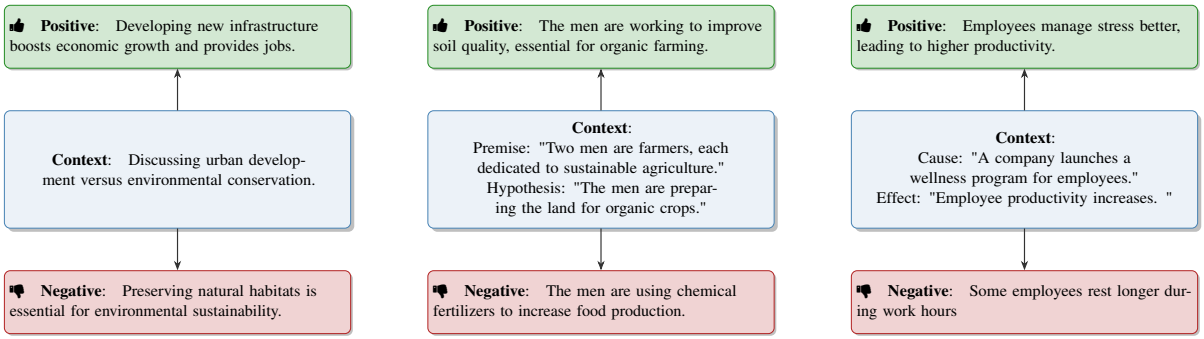
We list different motivational examples describing the role of context in different scenarios.

**Example in Debate**. In debate scenarios, context is crucial for transforming individual statements into oppositional arguments. Without a shared topic, positive and negative statements may not be inherently oppositional. The following are examples of how context influences the perception of opposition in debate, as illustrated in Figure 8a. We could observe that:

- *Without context*: The statements "Developing new infrastructure boosts economic growth and provides jobs" and "Preserving natural habitats is essential for environmental sustainability" are not inherently oppositional. Each statement highlights a positive aspect of a different domain: economic development and environmental conservation, respectively. Without a specific context, these statements are only independent, covering separate domains.
- *With context*: However, when provided with the debate topic, i.e., the context, "Discussing urban development versus environmental conservation," these statements become oppositional. The first statement advocates for urban development by emphasizing economic benefits, aligning with the pro-development side of the debate. The second statement underscores the importance of environmental conservation, aligning with the opposing stand that prioritizes preserving natural habitats over development. The shared context frames these statements as directly oppositional viewpoints on the same issue, illustrating how context turns independent statements into dichotomous arguments.

**Example for Defeasible Natural Language Inference**. In defeasible NLI, the presence of context transforms positive and negative statements that are not inherently oppositional into clearly oppositional positions by framing a specific inference relationship. As the example shown in Figure 8b, we could find that:

- *Without context*: Individually, positive and negative statements are not inherently oppositional. The first could indicate a general farming technique, while the second reflects a choice made for crop yield. Without the specific context, the two statements can be valid independently.
- *With context*: Given the premise-hypothesis inference relationship emphasizing sustainable agriculture and organic crops, these statements become oppositional. Specifically, the




(a) Scenario A: Debate. (b) Scenario B: Defeasible NLI. (c) Scenario C: Causal Reasoning.

Figure 8: Examples demonstrating how context transforms independent statements into dichotomous arguments across three scenarios: debate, defeasible natural language inference (NLI), and causal reasoning. Each subfigure shows the influence of shared context in framing statements as oppositional viewpoints or influences.

positive statement aligns with organic farming, supporting the hypothesis that these men are helping the land remain sustainable. However, the negative statement says that these men introduce chemical fertilizer, which opposes the organic practice. As seen, context justifies these two statements (organic versus chemical fertilizers) are dichotomous and oppositional within a sustainable farming approach.

**Example for Causal Reasoning**. In causal reasoning, context establishes a specific cause-effect relationship that additional statements can either support or refute. Without this context, statements may not appear oppositional. As the example shown in Figure 8c, we could observe that:

- *Without context*: The statements "Employees manage stress better, leading to higher productivity" and "Some employees rest longer during work hours" are not necessarily oppositional on their own. The first statement suggests a positive outcome of stress management, while the second indicates that employees are taking longer breaks. Without a specific cause-effect framework, these statements could describe independent aspects of workplace behaviors.

- *With context*: However, when provided with the contextual information of the cause "A company launches a wellness program for employees" and the effect "Employee productivity increases," these two statements become oppositional regarding their influences on strength of causal-effect relationship. The positive statement supports the causal link by explaining that the wellness program helps employees manage stress better, thereby enhancing productivity. The negative statement weakens or challenges the causal link by suggesting that the wellness program leads some employees to rest longer during work hours, potentially decreasing productivity. Within this context, the two statements offer oppositional perspectives on the effectiveness of the wellness program, transforming them into dichotomous arguments that either strengthen or weaken the perceived causal relationship.

### J.2 Empirical Evaluation of Contextual Influence

To quantify the importance of contextual information in dichotomy assessment, we evaluated model performance with context (Table 2, presented in the main body of this paper) and without context (Table 8). Our observations indicate that while our proposed DoGE achieves the highest scores across all metrics and scenarios when context is available, the removal of context leads to varied changes in Dichotomy Consistency Frequency (DCF) across other models or baselines. Specifically, some models exhibit an increase in DCF, others a decrease, but overall, the DCF scores are comparable or lower than 33.33, the threshold for random guessing. This convergence suggests that without context, the models' ability to accurately quantify dichotomy diminishes, as the arguments become more neutral.

Key observations include:

- **Convergence towards random guessing when removing context**: Removing context leads to Dichotomy Consistency Frequency scores (DCF, $\text{DCF}_{\text{positive}}$, $\text{DCF}_{\text{negative}}$) being

| Scenario | *Scenario A: Debate* | | | | *Scenario B: Defeasible NLI* | | | | *Scenario C: Defeasible Causality* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | DCF | $DCF_{positive}$ | $DCF_{negative}$ | Oppo-Angle | DCF | $DCF_{positive}$ | $DCF_{negative}$ | Oppo-Angle | DCF | $DCF_{positive}$ | $DCF_{negative}$ | Oppo-Angle |
| Open-source models | | | | | | | | | | | | |
| InferSent-*GloVe* | 36.24 | 62.32 | 61.12 | 8.59 | 25.98 | 54.04 | 53.05 | 14.55 | 33.67 | 59.98 | 58.36 | 12.61 |
| InferSent-*fastText* | 41.33 | 66.51 | 65.55 | 17.71 | 33.24 | 60.06 | 58.81 | 23.64 | 38.37 | 66.58 | 60.52 | 20.37 |
| USE | 22.35 | 52.82 | 49.78 | 21.54 | 23.36 | 54.20 | 53.02 | 29.55 | 20.88 | 48.62 | 45.60 | 21.47 |
| BERT baselines | | | | | | | | | | | | |
| BERT | 36.82 | 58.04 | 58.74 | 1.33 | 19.52 | 38.64 | 37.12 | 1.16 | 32.14 | 57.81 | 46.87 | 1.61 |
| CoSENT | 32.14 | 57.81 | 46.87 | 1.61 | 29.14 | 56.14 | 54.47 | 5.21 | 31.74 | 55.73 | 53.19 | 4.21 |
| SBERT | 27.03 | 53.05 | 49.49 | 15.18 | 24.51 | 54.13 | 52.25 | 26.11 | 23.18 | 50.15 | 45.24 | 18.05 |
| SimCSE | 31.57 | 57.95 | 52.61 | 15.87 | 23.68 | 53.26 | 50.80 | 26.33 | 28.36 | 59.61 | 48.05 | 21.95 |
| AoE | 27.71 | 56.87 | 49.67 | 8.04 | 26.35 | 54.81 | 52.20 | 12.33 | 26.60 | 57.88 | 45.46 | 9.65 |
| RoBERTa baselines | | | | | | | | | | | | |
| RoBERTa | 35.59 | 55.10 | 54.09 | 0.00 | 24.30 | 47.40 | 46.98 | 0.00 | 25.93 | 51.91 | 45.64 | 0.00 |
| SimCSE | 32.11 | 57.99 | 52.19 | 16.11 | 24.05 | 53.43 | 50.98 | 26.85 | 30.07 | 60.43 | 48.90 | 22.35 |
| LLaMA baselines | | | | | | | | | | | | |
| $LLaMA\text{-}2_{(7B)}$ | 37.44 | 56.49 | 54.79 | 18.76 | 27.31 | 46.15 | 45.53 | 11.57 | 28.89 | 51.48 | 46.55 | 5.68 |
| $LLaMA\text{-}2_{(13B)}$ | 38.70 | 60.26 | 57.48 | 11.06 | 22.37 | 42.75 | 42.34 | 4.54 | 27.07 | 49.95 | 43.40 | 3.03 |
| $AoE_{(7B)}$ | 35.42 | 54.11 | 53.37 | 21.03 | 29.83 | 49.05 | 48.87 | 11.63 | 31.02 | 51.77 | 48.30 | 4.92 |
| $AoE_{(13B)}$ | 36.45 | 56.63 | 55.03 | 10.66 | 23.36 | 44.25 | 43.95 | 3.60 | 28.90 | 51.34 | 45.64 | 2.20 |
| $LLaMA\text{-}3.1_{(8B)}$ | 28.34 | 49.91 | 43.67 | 5.70 | 26.27 | 43.96 | 43.06 | 9.43 | 26.38 | 45.22 | 39.15 | 2.67 |
| $LLaMA\text{-}3.1_{(70B)}$ | 32.15 | 53.57 | 51.46 | 17.73 | 25.33 | 44.66 | 43.93 | 11.65 | 29.32 | 52.11 | 43.36 | 6.10 |
| Our Method (BERT version) | | | | | | | | | | | | |
| DoGE (ours) | 36.90 ± 1.145 | 62.59 ± 0.948 | 57.75 ± 1.139 | 18.43 ± 5.695 | 31.93 ± 1.388 | 52.06 ± 1.433 | 51.21 ± 1.568 | 2.41 ± 0.694 | 45.67 ± 2.245 | 74.12 ± 3.183 | 58.25 ± 1.519 | 5.06 ± 0.941 |
| DoGE w/o DICT | 31.72 ± 1.187 | 58.63 ± 2.065 | 52.94 ± 0.764 | 64.08 ± 7.420 | 27.15 ± 1.670 | 48.40 ± 1.356 | 49.55 ± 1.525 | 4.44 ± 3.753 | 45.63 ± 1.940 | 75.75 ± 0.827 | 57.95 ± 1.530 | 69.25 ± 2.041 |
| DoGE w/o CL | 34.77 ± 0.823 | 58.67 ± 0.645 | 55.2 ± 0.725 | 9.48 ± 0.677 | 29.28 ± 1.394 | 49.28 ± 2.113 | 48.28 ± 1.717 | 2.07 ± 0.164 | 48.03 ± 1.250 | 75.76 ± 1.413 | 60.09 ± 0.935 | 5.70 ± 0.865 |
| Our Method (RoBERTa version) | | | | | | | | | | | | |
| DoGE (ours) | 47.66 ± 1.304 | 69.81 ± 1.219 | 63.56 ± 0.697 | 42.28 ± 5.149 | 31.96 ± 5.811 | 49.85 ± 6.703 | 49.63 ± 7.370 | 0.00 ± 0.005 | 52.66 ± 1.171 | 80.01 ± 1.292 | 63.38 ± 1.278 | 0.08 ± 0.036 |
| DoGE w/o DICT | 38.58 ± 3.671 | 60.84 ± 4.926 | 56.61 ± 1.762 | 31.66 ± 44.760 | 31.98 ± 0.923 | 44.97 ± 3.192 | 43.72 ± 2.892 | 0.17 ± 0.173 | 34.57 ± 0.925 | 53.97 ± 3.427 | 50.83 ± 1.098 | 0.00 ± 0.000 |
| DoGE w/o CL | 54.28 ± 0.198 | 73.35 ± 0.550 | 67.22 ± 0.209 | 0.61 ± 0.045 | 33.53 ± 3.814 | 50.59 ± 5.825 | 50.25 ± 6.041 | 0.00 ± 0.000 | 48.67 ± 5.394 | 76.77 ± 5.373 | 60.01 ± 4.346 | 0.01 ± 0.012 |

Table 8: Empirical evaluation of various models on the task of conditional dichotomy quantification in scenarios where contextual input is absent. The evaluation spans three distinct scenarios: debate (Scenario A), defeasible natural language inference (Scenario B), and causal reasoning (Scenario C). The results highlight the indispensable role of context in enabling models to accurately quantify dichotomous relationships. While most models suffer performance declines without contextual input (they tend to behave as random guesses in the absence of context), the proposed DoGE demonstrates relative robustness, maintaining higher DCF and Oppo-Angle scores compared to traditional baselines. This robustness is attributed to the integrated Dichotomous Objective and Contrastive Learning mechanisms, which collectively enhance the model's ability to discern and preserve dichotomous relationships even in the absence of explicit contextual guidance.

comparable or lower than random guessing thresholds across most models ($DCF = 33.33$, $DCF_{positive} = 50.00$, $DCF_{negative} = 50.00$) This trend indicates that, in the absence of contextual cues, arguments that once appeared as opposing standpoints often degrade into loosely related statements, underscoring the critical role of context in establishing meaningful opposition.

- **Effect of Context on Oppo-Angle**: For the Oppo-Angle metric, most models exhibit larger angle values when context is removed. This occurs because these models concatenate the context with the positive, negative, and neutral arguments. Without the context, the shared content is lost, resulting in larger angles. Larger angles imply greater absolute opposition, but at a cost of losing relational consistency (much lower DCF values). In contrast, training with contextual information produces slightly smaller Oppo-Angle values but achieves higher DCF values and relational consistency. This suggests that training with contextual information effectively balances relational consistency with absolute opposition, capturing genuine oppositional relationships rather than relying solely on surface-level similarities, unlike other models.

- **DoGE 's Relative Robustness:** While DoGE also shows reduced performance without context, it generally maintains its status as the top-performing model. This result indicates that the joint use of dichotomous (DICT) and contrastive (CL) objectives in DoGE provides a stronger inherent structure, enabling the model to better preserve dichotomous relationships in the absence of explicit contextual cues.

- **Influence of Objectives (Ablation Study):** Ablation studies on DoGE (i.e., w/o DICT and w/o CL) reveal that removing either objective leads to a more pronounced performance decline without context. This finding confirms that both objectives contribute to robust dichotomy quantification, maintaining appropriate geometric and dichotomous relationships

even when contextual guidance is removed.

In summary, while DoGE remains comparatively effective without context, the overall performance drop across all models reinforces that context is a crucial driver of opposition. These results align with earlier analyses and visualizations, demonstrating that context is not merely supplementary but fundamentally integral to reliable dichotomy quantification.

### J.3 Visualization for Role of Context

We plot the visualization of positive, negative, and neutral samples without context as the prefix for BERT, RoBERTa, LLaMA-3.1$_{(8B)}$, LLaMA-3.1$_{(70B)}$, DoGE (BERT), DoGE (RoBERTa) in Figure 9, Figure 10, Figure 11, and Figure 12, Figure 13, and Figure 14, respectively. From these figures, we have the following observations:

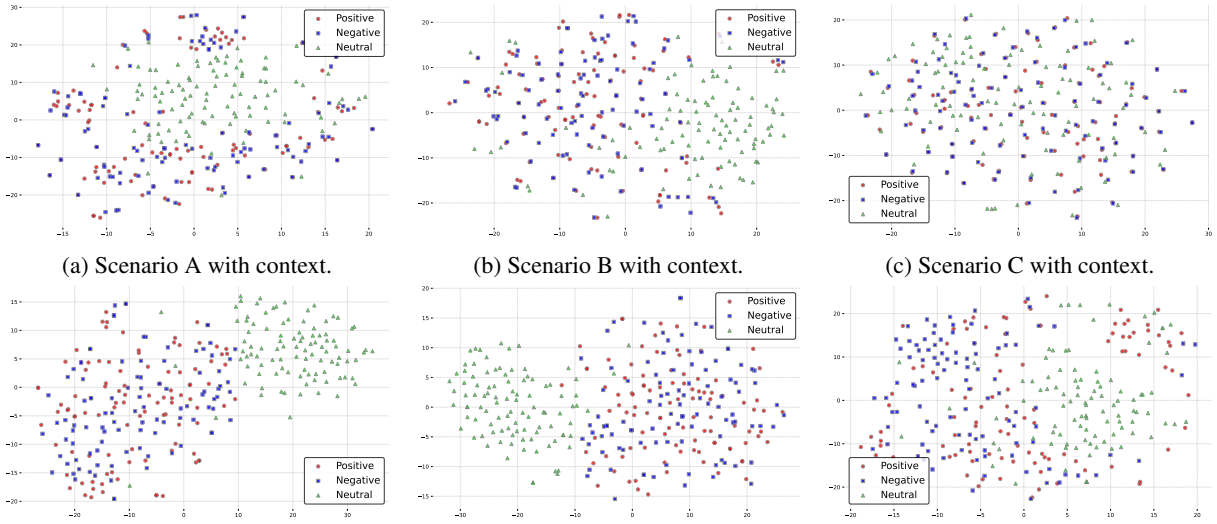
1. When examining the embedding distributions of baseline models such as BERT, RoBERTa, and various LLaMA-based configurations, a consistent pattern emerges. In the absence of contextual information, these baseline models clearly separate neutral samples from positive and negative ones. This initial distinction occurs because neutral arguments originate from data distributions that differ starkly from those of the positive and negative arguments, enabling the models to rely on superficial lexical or stylistic cues. Under these no-context conditions, the baselines effectively use such distributional discrepancies to position neutral instances in distinct regions of the embedding space, clearly isolating them from the other two categories.

2. When contextual input is introduced, these superficial differences are largely diminished. The presence of context tends to align the overall textual profiles of positive, negative, and neutral arguments, thereby obscuring the simple lexical or stylistic markers that the baseline models previously exploited. As a result, baseline models like BERT, RoBERTa, and LLaMA-based models no longer consistently find distributional patterns that readily separate neutral arguments from positive and negative arguments. Instead of the neat clustering observed without context, their embedding spaces become more entangled, reflecting a diminished ability to maintain clear distinctions.

3. In contrast, our proposed method, DoGE, preserves robust separations among positive, negative, and neutral samples regardless of whether context is present or not. Rather than depending solely on external distributional cues, DoGE leverages a specialized dichotomous objective in conjunction with contrastive learning. This combination encourages the model to capture deeper semantic and relational attributes that define the opposition between arguments. Consequently, even when context aligns certain textual properties across all samples, DoGE retains its capacity to represent and differentiate dichotomous relationships. Its embeddings maintain a stable geometric structure that clearly positions neutral arguments between the positive and negative extremes, ensuring that context does not erode the model's ability to recognize and preserve semantic or relational contrasts.

## K List of Notations

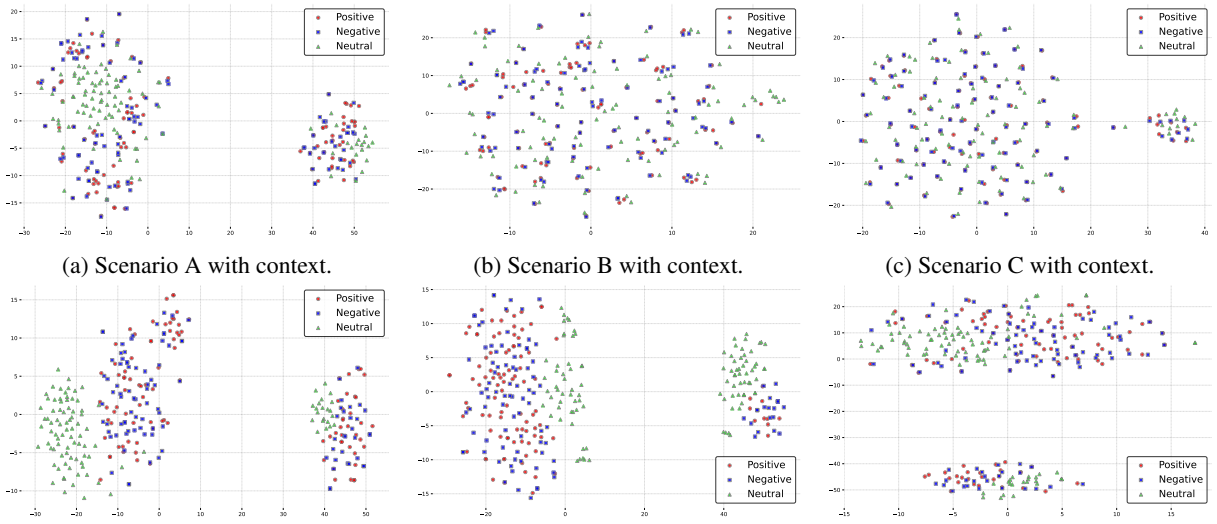
To facilitate a clear understanding of the concepts and methodologies presented in this paper, we provide a comprehensive table of notations in Table 9. This table outlines all the symbols and abbreviations used, along with their precise definitions, ensuring that readers can easily refer to and comprehend the technical aspects of our work.




(a) Scenario A with context. (b) Scenario B with context. (c) Scenario C with context.

(d) Scenario A without context. (e) Scenario B without context. (f) Scenario C without context.

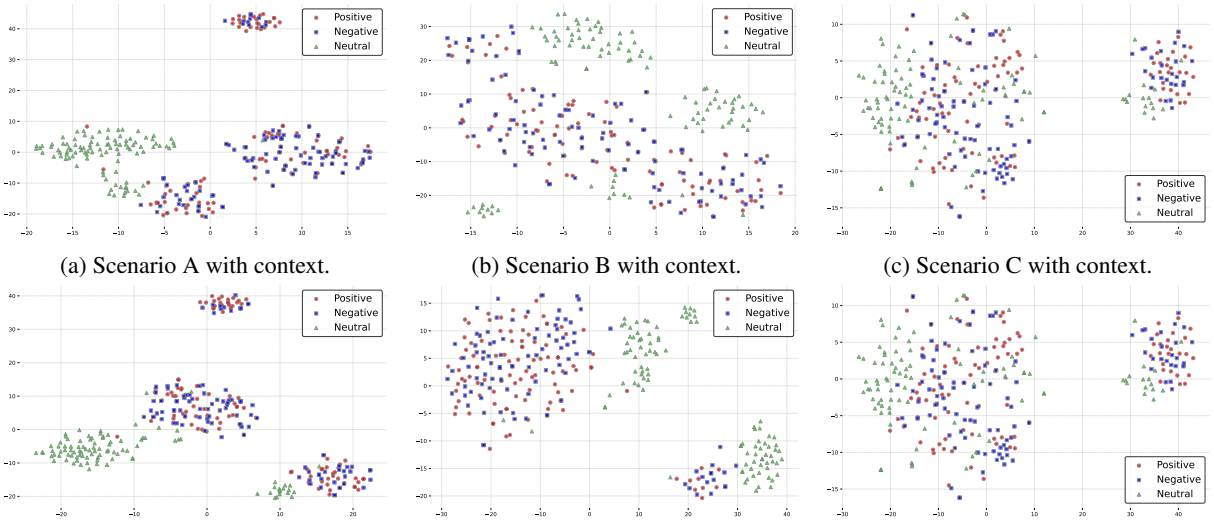
Figure 9: The t-SNE visualization of the embedding space for positive, negative, and neutral argument types using BERT across various scenarios. The figure contrasts the embedding distributions with and without contextual input.




(a) Scenario A with context. (b) Scenario B with context. (c) Scenario C with context.

(d) Scenario A without context. (e) Scenario B without context. (f) Scenario C without context.

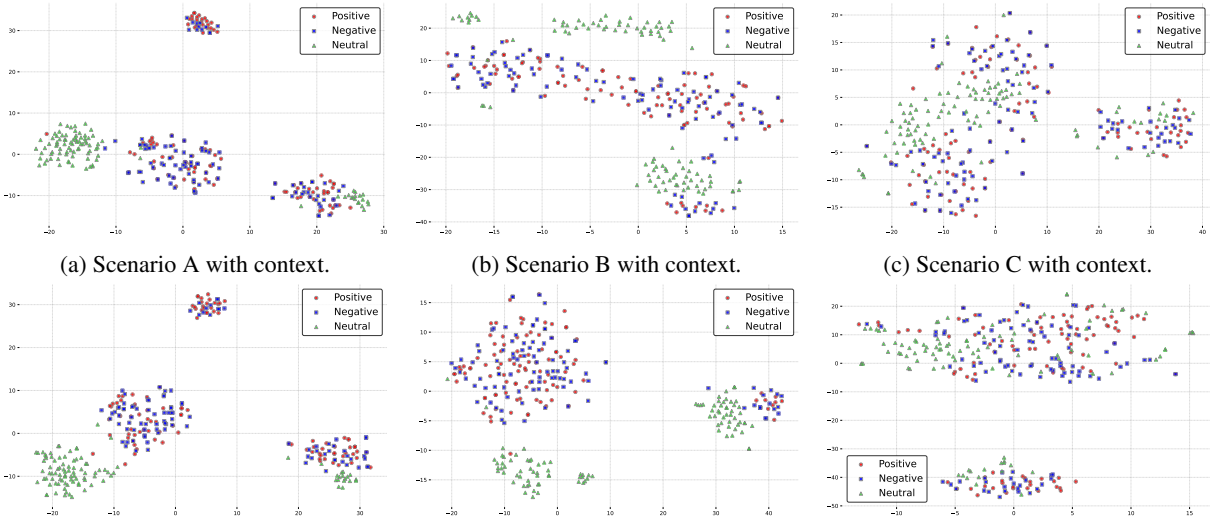
Figure 10: The t-SNE visualization of the embedding space for positive, negative, and neutral argument types using RoBERTa across various scenarios. The figure contrasts the embedding distributions with and without contextual input.




(a) Scenario A with context. (b) Scenario B with context. (c) Scenario C with context.

(d) Scenario A without context. (e) Scenario B without context. (f) Scenario C without context.

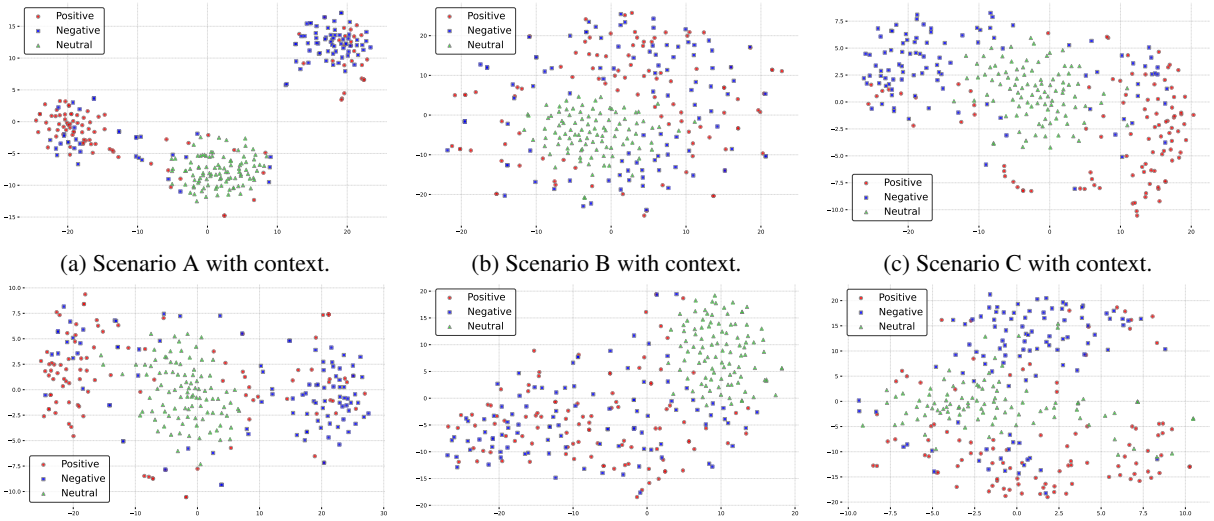
Figure 11: The t-SNE visualization of the embedding space for positive, negative, and neutral argument types using LLaMA-$3_{(8B)}$ across various scenarios. The figure contrasts the embedding distributions with and without contextual input.




(a) Scenario A with context. (b) Scenario B with context. (c) Scenario C with context.

(d) Scenario A without context. (e) Scenario B without context. (f) Scenario C without context.

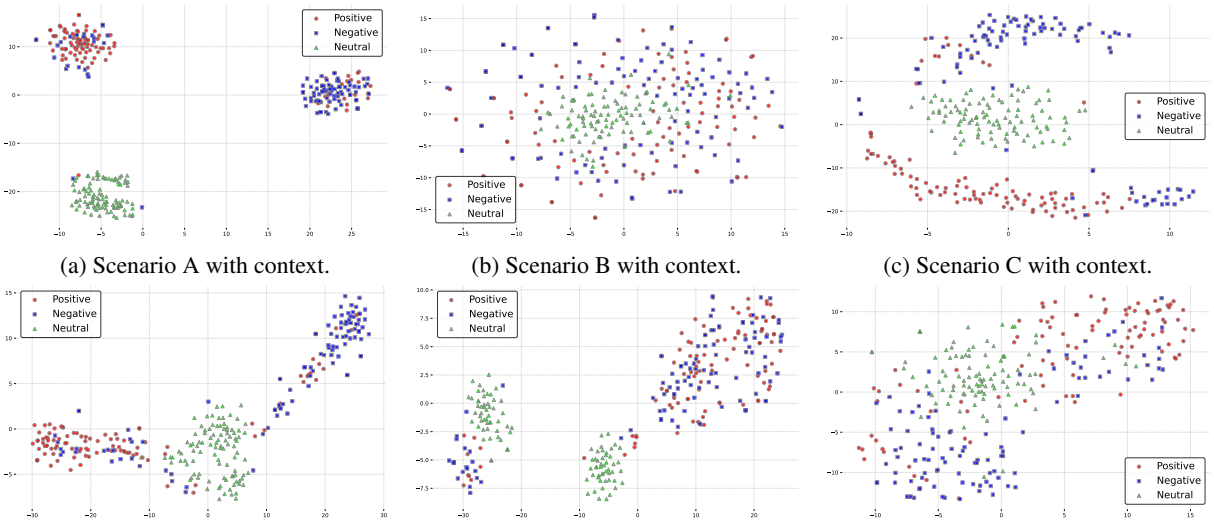
Figure 12: The t-SNE visualization of the embedding space for positive, negative, and neutral argument types using LLaMA-$3_{(70B)}$ across various scenarios. The figure contrasts the embedding distributions with and without contextual input.




(a) Scenario A with context. (b) Scenario B with context. (c) Scenario C with context.

(d) Scenario A without context. (e) Scenario B without context. (f) Scenario C without context.

Figure 13: The t-SNE visualization of the embedding space for positive, negative, and neutral argument types using DoGE (BERT version) across various scenarios. The figure contrasts the embedding distributions with and without contextual input, illustrating DoGE's capability to preserve dichotomous relationships regardless of contextual variations.




(a) Scenario A with context. (b) Scenario B with context. (c) Scenario C with context.

(d) Scenario A without context. (e) Scenario B without context. (f) Scenario C without context.

Figure 14: The t-SNE visualization of the embedding space for positive, negative, and neutral argument types using DoGE (RoBERTa version) across various scenarios. The figure contrasts the embedding distributions with and without contextual input, illustrating DoGE's capability to preserve dichotomous relationships regardless of contextual variations.

| Notation | Meaning |
|---|---|
| $Z$ | A shared conditional context. This is a piece of text or scenario on which multiple arguments (positive, negative, neutral) are based. |
| $X$ | A positive argument conditioned on $Z$. In the dichotomous setting, $X$ supports or strengthens the stance conveyed by $Z$. |
| $Y$ | A negative argument conditioned on $Z$. This argument is the opposite or adversarial counterpart to $X$, providing a contrasting stance. |
| $W$ | A neutral argument conditioned on $Z$. Unlike $X$ or $Y$, the neutral argument does not strongly support or oppose the stance conveyed by $Z$. |
| $\mathbf{E}_{X\|Z}$, $\mathbf{E}_{Y\|Z}$, $\mathbf{E}_{W\|Z}$ | Embeddings of $X$, $Y$, and $W$ when considered under the context $Z$. These embeddings are learned vector representations capturing semantic information as well as the relationship to $Z$. |
| $\mathbf{E} = \mathbf{E}^{\mathrm{re}} + \mathbf{E}^{\mathrm{im}}\mathrm{i}$ | Complex-valued embedding representation of an argument. The embedding vector is decomposed into a real part $\mathbf{E}^{\mathrm{re}}$ and an imaginary part $\mathbf{E}^{\mathrm{im}}$. This complex-valued representation provides richer geometric properties to capture dichotomy. |
| $\Phi_{(XY\|Z)}$ | The dichotomous degree function that quantifies how opposing $X$ and $Y$ are conditioned on the shared context $Z$. Larger values indicate greater dichotomy. |
| $\Delta_{(XY\|Z)}$ | Angular distance between embeddings $\mathbf{E}_{X\|Z}$ and $\mathbf{E}_{Y\|Z}$. This measures the angle between the vectors, reflecting their relational difference rather than just their lexical or semantic similarity. |
| $\Gamma_{(XY\|Z)}$ | The angular distance computed in the complex-valued embedding space. Incorporating the imaginary part allows for a more nuanced representation of opposing stances. |
| DCF | Dichotomy Consistency Frequency. A metric that measures how often the model correctly captures the relational order: neutral arguments lie closer to positive and negative arguments than these arguments lie to each other. A higher DCF means the embedding space consistently reflects the intended dichotomous structure. |
| $\mathrm{DCF}_{\mathrm{positive}}$, $\mathrm{DCF}_{\mathrm{negative}}$ | Sub-metrics derived from DCF. $\mathrm{DCF}_{\mathrm{positive}}$ measures how often the neutral argument is closer to the positive argument than the positive and negative arguments are to each other. $\mathrm{DCF}_{\mathrm{negative}}$ checks the same for the neutral argument and the negative argument. |
| Oppo-Angle | A metric that directly quantifies the absolute opposition between $X$ and $Y$. Unlike DCF, which is relational and involves a neutral argument, Oppo-Angle measures how strongly the positive and negative arguments diverge in the embedding space. Larger Oppo-Angle scores reflect stronger opposition. |
| $\mathcal{L}_{\mathrm{dichotomous}}$ | The dichotomous objective loss term, ensuring that neutral arguments are geometrically positioned between positive and negative arguments, thus enforcing a balanced and interpretable embedding structure. |
| $\mathcal{L}_{\mathrm{cl}}$ | The contrastive learning loss term that pushes positive and negative arguments further apart, enhancing the capture of inherent opposition beyond simple semantic similarity or difference. |

| | |
|---|---|
| $\tau_{\text{dichotomous}}$, $\tau_{\text{cl}}$ | Temperature hyperparameters for two different loss functions: $\tau_{\text{dichotomous}}$ controls the sharpness of the training signal in the dichotomous loss, where smaller values emphasize hard negatives and positives, and larger values smooth the signal; $\tau_{\text{cl}}$ similarly controls the sharpness in the contrastive loss. |
| $w_1$, $w_2$ | Weighting coefficients that balance the dichotomous and contrastive loss terms. They control the relative emphasis on relational consistency vs. absolute opposition during training. |
| $N$ | The number of samples in a dataset or batch of data. Used in averaging metrics over multiple instances. |
| $m$,$b$ | $m$ represents the number of positive pairs in the $b$-th batch. |
| $d$ | The dimension of the embedding vector for each argument. The full embedding in the complex-valued space has dimension $2d$ due to the real and imaginary parts. |

Table 9: A detailed summary of notations and their meanings.