GeoHard: Towards Measuring Class-wise Hardness through Modelling Class Semantics

Fengyu Cai[1] **Xinran Zhao**[2] **Hongming Zhang**[3] **Iryna Gurevych**[1] **Heinz Koeppl**[1]
[1]Technical University of Darmstadt [2]Carnegie Mellon University [3]Tencent AI Lab
{fengyu.cai, heinz.koeppl}@tu-darmstadt.de

Abstract

Recent advances in measuring hardness-wise properties of data guide language models in sample selection within low-resource scenarios. However, class-specific properties are overlooked for task setup and learning. How will these properties influence model learning and is it generalizable across datasets? To answer this question, this work formally initiates the concept of *class-wise hardness*. Experiments across eight natural language understanding (NLU) datasets demonstrate a consistent hardness distribution across learning paradigms, models, and human judgment. Subsequent experiments unveil a notable challenge in measuring such class-wise hardness with instance-level metrics in previous works. To address this, we propose *GeoHard* for class-wise hardness measurement by modeling class geometry in the semantic embedding space. *GeoHard* surpasses instance-level metrics by over 59 percent on *Pearson*'s correlation on measuring class-wise hardness. Our analysis theoretically and empirically underscores the generality of *GeoHard* as a fresh perspective on data diagnosis. Additionally, we showcase how understanding class-wise hardness can practically aid in improving task learning. The code for *GeoHard* is available [1].

1 Introduction

Data acts as a crucial intermediary proxy for AI systems to understand and tackle real-world tasks (Torralba and Efros, 2011; Vodrahalli et al., 2018). Therefore, evaluating the hardness of individual instances, or instance-level hardness (Kong et al., 2020; Hahn et al., 2021; Ethayarajh et al., 2022; Zhao et al., 2022), relative to the dataset is key for learning and analyzing NLP tasks. This evaluation is increasingly important with the rise of large language models (LLMs; Touvron et al. 2023; Chung et al. 2024). Measuring hardness aids in selecting examples for in-context learning (ICL; Ye et al. 2024) or training samples for fine-tuning models (Zhou et al., 2023; Xie et al., 2023).

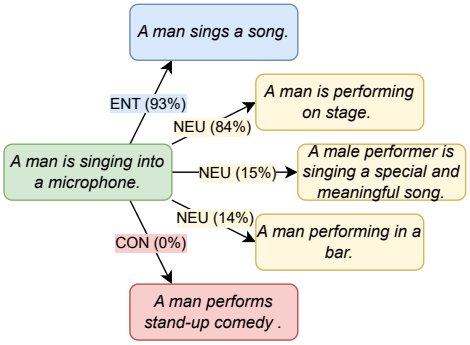


Figure 1: The examples of premise-hypothesis pairs in uncertain NLI (u-NLI; Chen et al. 2020). In u-NLI, the probability of these pairs (in the parentheses) is annotated by crowdworkers. The example showcases NEU's *Middlemost* and *Diverse* semantics, i.e., positioning in the middle between ENT and CON and widely ranging from low (14%) to high probability (84%).

However, another critical yet underexplored component of the dataset is the classes themselves, whose properties, such as ambiguity in their definitions, can also contribute to difficulties. While considerable efforts have been made to address class imbalance in specific datasets (Subramanian et al., 2021; Henning et al., 2023), there remains a lack of comprehensive analysis on class-wise properties that are consistent across different tasks. Conventionally, classes are treated equally, e.g., the demonstrations in In-Context Learning (ICL) typically being evenly sampled among classes (Min et al., 2022). This raises an important question: *How do class-specific properties influence model performance?*

We formally initiate the concept of *class-wise hardness* as the relative difficulty of a class, in analogy to instance-level hardness (Ethayarajh et al., 2022). To make this notion quantifiable, we present the concept of the *empirical* class-wise

[1]https://github.com/TRUMANCFY/geohard

Findings of the Association for Computational Linguistics: ACL 2024, pages 5571–5597

hardness which assesses the class-specific performance given an LM and learning paradigm. Subsequently, the intrinsic class-wise hardness can be approximated by pooling the empirical performances across models and learning paradigms. Our analysis across eight Natural Language Inference (NLI) or Sentiment Classification (SC) tasks reveals the consistent challenge of *Neutral* across a spectrum of tasks, learning paradigms, and models together with human annotation disagreement (Nie et al., 2020). These findings verify the concept and establish the estimation of inherent class hardness.

Then, we study how to measure these class-specific properties leading to consistent class-wise hardness. We first show that naively aggregating Sensitivity Analysis (SA, Hahn et al. 2021) and two similarity-based methods (Zhao et al., 2022, 2023a) fails in measuring class-wise hardness across datasets. This stimulates us to propose a specific metric for class-wise hardness measurement beyond the instance-level measurement. We propose an effective, lightweight, and training-free metric, *GeoHard*, which analyzes data distribution from the geometrical space of semantic embeddings. *GeoHard* utilizes both *inter-* and *intra*-class properties, e.g., *Neutral*'s *MiddleMost* and *Diverse* semantics shown in Figure 1, respectively. Our experiments show that *GeoHard* demonstrates its exceptional capacity in measuring class-wise hardness, outperforming the instance-level aggregation by over 59 percent on *Pearson*'s correlation between measurement and reference. Our theoretical and experimental analysis validates its generalization to other tasks without further adaptation.

As for the practical perspective, we show how to use *GeoHard* to improve *task learning* with class reorganization (Nighojkar et al., 2023). Class reorganization targets a balanced class performance, e.g., by splitting one hard class into two sub-classes (Potts et al., 2021). *GeoHard* is shown to be able to well interpret the heuristic-based reorganization proposed in the previous work (Potts et al., 2021). We demonstrate that class-aware demonstration selection guided by *GeoHard* also benefits ICL.

Our contribution is three-fold:

1. We initiate the concept of class-wise hardness (Section 2) and show that the direct aggregation of the current instance-level hardness metrics fails to correlate with class-wise hardness on 8 NLI/SC datasets (Section 4);
2. We instead target class semantics and put forward a geometry-based method, *GeoHard*, which outperforms the baselines by 59% (Section 3). We theoretically and empirically show *GeoHard*'s promising generalization to other tasks (Section 5);
3. We demonstrate the potential application of class-wise hardness measured by *GeoHard* to interpret class reorganization and improve task learning (Section 6).

2 Formulation of Class-wise Hardness

Here, we define class-wise hardness as the difficulty of the class across all the classes, akin to instance-level hardness (Ethayarajh et al., 2022). Formally, given the classes $\mathcal{C} = \{c_1, ..., c_K\}$ for a classification task where c_k is a class, c_k's class-wise hardness can be denoted as $\mathrm{H}(c_k \mid \mathcal{C})$. We denote $\mathbf{H}(\mathcal{C}) = [\mathrm{H}(c_1 \mid \mathcal{C})), ..., \mathrm{H}(c_K \mid \mathcal{C})]$.

As H is intractable, we can empirically obtain class-wise hardness by assessing the performance of c_k given the LM $m \in \mathcal{M}$, e.g., Flan-T5-Large (Raffel et al., 2020) or LLaMA-2-13B (Touvron et al., 2023), and learning paradigms $l \in \mathcal{L}$, e.g., fine-tuning or ICL. We denote this empirical class-wise hardness conditioned on LMs and learning paradigms as $\tilde{\mathrm{H}}(c_k \mid \mathcal{C}, m, l)$. Therefore, class-wise hardness H can be approximated by marginalizing $\tilde{\mathrm{H}}$ on the pairs of models and learning paradigms $\mathcal{P} = \{(m, l) \mid m \in \mathcal{M}, l \in \mathcal{L}\}$:

$$\mathrm{H}(c_k \mid \mathcal{C}) = \mathbb{E}_{(m,l)\in\mathcal{P}}[\tilde{\mathrm{H}}(c_k \mid \mathcal{C}, m, l)] \quad (1)$$

$$\approx \frac{\sum_{(m,l)\in\mathcal{P}} \tilde{\mathrm{H}}(c_k \mid \mathcal{C}, m, l)}{|\mathcal{P}|} \quad (2)$$

In the rest of this section, we calculate empirical class-wise hardness on eight NLI/SC datasets. We observe the consistency of $\tilde{\mathbf{H}}$ among LMs, learning paradigms, and human annotation, which stimulates us to simplify the approximation of H.

2.1 Datasets

We initiate class-wise hardness with 8 NLU datasets, comprising 3 NLI datasets and 5 SC datasets, as shown in Table 6 and Table 7 in Appendix A.1. We chose these datasets based on their popularity and their similar format for comparison. We normalize the label format of the SC datasets to *Positive*, *Neutral*, and *Negative*, as described in Appendix A.2. Lastly, we balance the number of instances within each class[2]. Class imbalance

[2]We eliminate the potential influence of class imbalance by randomly sampling the same number of instances belonging

%	Roberta-Large	OPT-350M	Flan-T5-Large
Amazon	87.6|**71.0**|80.6	87.0|**68.7**|79.3	88.6|**71.6**|81.3
APP	74.2|**60.1**|73.4	73.6|**56.1**|72.6	74.3|**59.0**|73.9
MNLI	91.0|**87.2**|92.9	86.1|**80.5**|85.8	91.3|**87.5**|92.9
SICK-E	92.9|**86.8**|92.4	85.8|**79.1**|89.1	92.9|**85.7**|92.4
SNLI	92.6|**89.2**|95.3	91.0|**86.5**|92.3	92.8|**89.7**|95.5
SST-5	83.1|**53.1**|75.8	82.3|**55.6**|71.5	83.4|**51.7**|76.1
TFNS	93.0|**86.1**|92.2	88.0|**81.1**|88.7	87.3|**77.3**|88.0
Yelp	87.9|**75.4**|86.6	86.4|**73.5**|85.0	88.3|**76.4**|87.0

Table 1: Class-wise performance by the finetuned model **Roberta-Large**, **OPT-350M**, and **Flan-T5-Large**. Each entry presents the F1 score of *Positive*/*Entailment*, *Neutral*, and *Negative*/*Contradiction* concatenated with |. **Bold** indicates the lowest F1 score among classes. The results are averaged by 3 runs with different seeds, as shown in Appendix A.3.

is shown to negatively affect the performance of minority classes (Henning et al., 2023).

2.2 Calculation of Empirical Hardness $\tilde{\mathrm{H}}$

To achieve a precise and complete approximation on H, we encompass various pairs of LMs and learning paradigms for the calculation of $\tilde{\mathrm{H}}(c_k \mid \mathcal{C}, m, l)$, as outlined in Equation 2.

Inter-annotator disagreement can reflect the difficulty of the instance, namely that the higher human disagreement implies more hardness on data (Nie et al., 2020; Basile et al., 2021). Hence, we calculate class-wise human disagreement as the average entropy of the annotation distribution of the instances labeled on MNLI and SNLI [3]. Referring to Table 11 in the Appendix A.3.1, *Neutral*'s class-wise human disagreement is the highest, indicating its exceptional hardness w.r.t. human.

Fine-tuning To generalize empirical class-wise hardness, models from diverse architectures are chosen: we use Roberta-Large (Liu et al., 2019), OPT-350M (Zhang et al., 2022), and Flan-T5-Large (Chung et al., 2024). These models belong to encoder-only, decoder-only, and encoder-decoder structures, respectively. We train these three models separately on eight datasets following the training setups presented in Appendix A.3. We select the checkpoint with the best F1 score on the validation dataset to evaluate the test set. Table 1 shows that *Neutral* performs poorest among classes with all three models on all the datasets, verifying *Neutral*'s consistent hardness w.r.t. fine-tuned LMs.

to each class in training, validation, and test sets, respectively.

[3]Only the inter-annotator agreement of MNLI and SNLI is evaluated due to data accessibility.

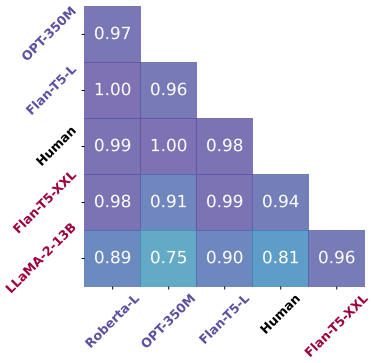


Figure 2: Correlation matrix among class-wise F1 scores of three finetuned models together with two ICLs and class-wise human disagreement on SNLI, where the high consistency is noted. Figure 8 presents MNLI's correlation matrix in Appendix A.3.2.

In-context Learning Beyond model fine-tuning, we also explore another paradigm using large language models (LLMs), where the answer is elicited from LLMs by injecting a cue or instruction (Ye et al., 2023). Specifically, we conduct experiments on MNLI and SNLI using Flan-T5-XXL and LLaMA-2-13B. The templates employed are shown in Appendix A.3.2, and *Neutral*'s relative hardness stands referring to Table 12.

Figure 2 demonstrates *Neutral*'s consistent hardness, e.g., in SNLI, across various LMs m, learning paradigms l, and human annotation, revealing that its class-wise hardness is *intrinsic*. Given this observation, we further relax the approximation of $\tilde{\mathrm{H}}$ in Equation 2 that if the correlation among $\tilde{\mathrm{H}}$ with (m, l) pairs is higher than a specific threshold, we can approximate H with $\tilde{\mathrm{H}}$ with arbitrary m and l:

$$\mathrm{H}(c_k \mid \mathcal{C}) \approx \tilde{\mathrm{H}}(c_k \mid \mathcal{C}, m, l) \tag{3}$$

3 *GeoHard* for class-wise hardness measurement

Regarding the intrinsic class-wise hardness shown in Section 2, we quantitatively measure the corresponding empirical hardness motivated by its semantic properties, e.g., *Diverse* and *Middlemost* semantics of *Neutral*. Specifically, as the name suggests, *GeoHard* measures class-wise hardness by computing the geometrical metrics in the semantic embedding space.

3.1 Notations

The set of K classes is denoted as $\mathcal{C} = \{c_1, ..., c_K\}$. The dataset with N instances is denoted as $\mathcal{D} =$

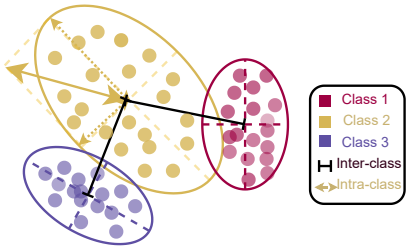


Figure 3: The illustration of *GeoHard* in semantic embeddings space. The ellipses approximate class-wise data distribution. Class 2 is speculated to be difficult due to its large variance and middlemost location.

$\{(X, y)_{1:N}\}$, where X is the input and $y \in \mathcal{C}$ is the corresponding label. And θ signifies model parameters. $\|\cdot\|_{1(2)}$ presents L1(2)-norm. The input instances of the label c_k is denoted by X^{c_k}, i.e., $X^{c_k} = \{X_i | \forall (X_i, y_i) \in D_{train}, y_i = c_k\}$.

3.2 *GeoHard*

Semantic representation As *GeoHard* aims to measure class-wise hardness through modeling semantics, a sentence encoder is therefore required, which maps a sentence to a vector with a dimension E. We denote this mapping function as $f(\cdot)$.

Semantics-guided metrics *GeoHard* consists of *intra-* and *inter*-class metrics modeling two semantics properties, as illustrated in Figure 3. The *intra*-class metric, corresponding to *Diverse* semantics, quantifies the distributional variance within one class, formulated as:

$$\mathrm{H}_{intra}(c_k) = \|\sigma(f(X^{c_k}))\|_2 \tag{4}$$

where σ denotes the element-wise variance across the instances, i.e., $\sigma : \mathbb{R}^{N \times E} \rightarrow \mathbb{R}^E$.

Middlemost semantics indicate one class is located closer to other classes in the representation space. Hence, the *inter*-class metric calculates the average distance from one class center to the other classes. The opposition aims to unite m_{inter} with the m_{intra} regarding the overall hardness tendency:

$$\mathrm{H}_{inter}(c_k) = \frac{-\sum_{\substack{i=1 \\ i \neq k}}^{K} ||\mu(f(X^{c_k})) - \mu(f(X^{c_i}))||_1}{K-1}$$

where $\mu(\cdot)$ presents the element-wise mean operation across the input set, that is $\mu : \mathbb{R}^{N \times E} \rightarrow \mathbb{R}^E$.

To this end, *GeoHard* of one specific class is the amalgamation of the class-wise *intra-* and *inter-class* metrics, i.e., $\mathrm{H}_{GeoHard}(c_k) = \mathrm{H}_{intra}(c_k) + \mathrm{H}_{inter}(c_k)$. And the higher *GeoHard* indicates more class-wise hardness, i.e., a poorer performance.

3.3 Implementation

According to the open reference [4], we apply E5-large-v2 (Wang et al., 2022) to project sentences to a high dimensional space. Jha and Mihata (2021) point out that the nonlinear dimension reduction on contextualized representation benefits downstream tasks. Therefore, we apply Uniform Manifold Approximation and Projection (UMAP; McInnes et al. 2018) to compress sentence representation to E [5]. The complete encoder consists of E5-large-v2 and UMAP [6].

4 Experiments

4.1 Baseline: *Instance* Hardness Aggregation

Sensitivity Analysis (Hahn et al., 2021) measures data hardness by assessing how perturbations in the input affect a model's prediction. It calculates the model's prediction confidence for an instance and its perturbed neighbor on the golden label. A larger derivative between these confidences, i.e., higher sensitivity, signifies greater hardness.

As for the class-wise hardness, we average the sensitivity values of the samples in each class. The higher class-wise sensitivity suggests more difficulty in the class, in consistency with Hahn et al. (2021). We take the finetuned Roberta-Large in Section 2.2 as the reference model. More implementation details can be found in Appendix B.2.

Spread (Zhao et al., 2022) & Thrust (Zhao et al., 2023b) measure the instance-level hardness by estimating the similarity between test instances and training samples. Concretely, Spread calculates the semantic similarity between test instances and a few-shot closest training samples, using the sentence encoders. E5-large-v2 (Wang et al., 2022) is also applied by Spread in line with *GeoHard*, and the number of training selections is 8. Thrust calculates the distance of the decoded instance representation by LLMs between training and test sets. We apply the identical LLM as the original work, i.e.,

[4] E5-large-v2 led Massive Text Embeddings Benchmark leaderboard (Muennighoff et al., 2023) at the time of the work.

[5] We set $E = 2$ for visualization in the experiment.

[6] As E5-large-v2 is trained to capture uni-sentence semantics, we concatenate premise and hypothesis in NLI tasks with six conjunctive words or phrases shown in Appendix C.1.1 referring to the templates applied in Gao et al. 2021.

Metric \ Dataset	SC					NLI			Macro Avg. ↑
	Amazon	**APP**	**SST-5**	**TFNS**	**Yelp**	**MNLI**	**SNLI**	**SICK-E**	(Absolute)
SA	.4730	-.9620	-.2244	-.9047	.8980	-.7219	.3930	-.9962	.2556±.4961
Thrust	-.9780	.9012	.0833	.5157	.9952	.0000	.8311	.0000	.2936±.3792
Spread	-.6350	.4550	.8369	.9944	-.2148	.4292	-.3934	.8471	.2899±.3412
GeoHard-Intra	.5857	.9539	.9892	-.1574	.9947	.9784	.8805	.6647	.7362±.1354
GeoHard-Inter	.9997	.9964	.9908	.9722	.9978	.1500	.9042	.3663	.7972±.1006
GeoHard	.9998	.9958	.9909	.8852	.9977	.8384	.8882	.4871	**.8854**±**.0262**

Table 2: Pearson's correlation coefficients between class-wise hardness measurement and class-wise F1 scores, i.e., the approximation of H. All the metrics have been adjusted so that the higher correlation indicates better measurement. Red indicates that the value is opponent to the original design. **Bold** indicates the best performance among the methods with the highest average correlation and lowest variance across the datasets. ↑ indicates the higher values present better results. *GeoHard*'s results are averaged on 3 runs (and 6 conjunctions for NLI). Please refer to Appendix C for the detailed values.

Flan-T5-Large fine-tuned on UnifiedQA dataset [7] (Khashabi et al., 2020).

As both methods are similarity-based, the smaller similarity indicates more hardness. To this end, we average Spread scores in each class as the class-wise metrics. For Thrust, we select the bottom 25 percentile of Thrust scores in each class as the aggregation [8]. Appendix B.3 and B.4 present their detailed implementation.

4.2 Quantification of class-wise hardness

We benchmark the instance-aggregating methods (SA, Spread, and Thrust) as well as *GeoHard*, including its *intra-* and *inter*-class metrics, on the eight NLI/SC datasets in Section 2.

Section 2 illustrates the consistency between LMs and humans regarding class-wise hardness, and this allows us to select an arbitrary empirical class-wise hardness $\tilde{\mathrm{H}}$ as a *close* approximation of H. Consequently, we apply the class-wise F1 scores from fine-tuned Roberta-Large as the hardness reference. Following the previous work (Zhao et al., 2022), we determine the effectiveness of various metrics by calculating *Pearson*'s correlation coefficient between metrics and the hardness reference (Table 2). Considering negative correlation, we take the **absolute** value of average correlations as shown in the right-most column, namely that higher values indicate better measurement.

[7] https://huggingface.co/allenai/unifiedqa-t5-large

[8] The reason we do not average Thrust for class-wise hardness here is that this metric is inversely proportional to the distance. Therefore, Thrust values will come to infinity when the test sample is extremely close to the training set.

4.3 Analysis on Experimental Results

Table 2 presents the correlation between class-wise hardness measurement and the reference hardness on these eight NLI/SC datasets. The class-wise SA, Spread, and Thrust are shown to be **poorly** correlated to the reference, with average correlations of 0.2556, 0.2936, and 0.2899, respectively. Their large variance in correlation across tasks indicates their incompetence in class-wise hardness measurement. Meanwhile, *GeoHard* significantly outperforms these instance-level methods, exhibiting the lowest variance across the tasks. In addition to these metrics, *GeoHard* surpasses its components, namely the *intra-* and *inter*-class metrics, highlighting their complementarity and underscoring *GeoHard*'s comprehension of class-specific properties.

5 Generalization of *GeoHard*

The previous section showcased the exceptional performance of *GeoHard* in measuring hardness in NLI and SC tasks by leveraging class-wise semantic properties. In this section, we explore *GeoHard*'s robustness and generalization both theoretically and empirically. We conduct the experiments to demonstrate *GeoHard*'s generalization capabilities across various sentence encoders and other types of tasks, further substantiating the connection between class-wise hardness and semantics. Furthermore, we highlight *GeoHard*'s robustness in low resource scenarios, showcasing its advantage as a training-free metric.

5.1 Theoretical proof on generalization

GeoHard's robustness is evident in its ability to effectively elucidate factors contributing to class-wise hardness, such as overfitting depicted in Fig-

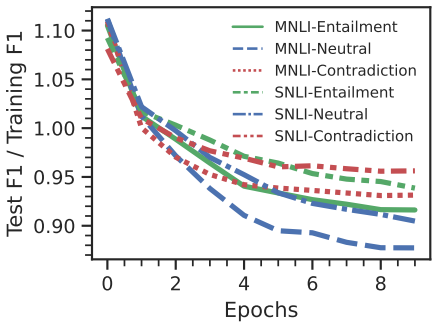


Figure 4: The ratio between F1 scores on the test and training sets for training epochs on NLI tasks. *Neutral* in blue suffers from overfitting most. Figure 10 in Appendix D.2 presents a similar issue in NLI tasks.

ure 4. The *intra-class* metric within *GeoHard* serves to gauge the extent of overfitting, namely, the divergence between training and test data, as elaborated in the following Theorem 1.

Theorem 1 *Assuming a Gaussian distribution for instances within* c_k, $D \sim \mathcal{N}(\mu_{c_k}, \sigma^2_{c_k})$, *the means of the training and test data can be represented as* $\hat{\mu}^{tr}_{c_k} \sim \mathcal{N}(\mu_{c_k}, \sigma^2_{c_k}/n_{tr})$ *and* $\hat{\mu}^{te}_{c_k} \sim \mathcal{N}(\mu_{c_k}, \sigma^2_{c_k}/n_{te})$, *where* n_{tr} *and* n_{te} *are the sizes of the training and test sets within* c_k, *respectively. Note that, conditioned on the class* c_k, *the data* D *is i.i.d. as mentioned above. By applying Chebyshev's inequality (Mitrinovic et al., 2013), the following inequality holds for any arbitrary* $k \in \mathbb{R}_+$ *(see mathematical derivation in Appendix D.1):*

$$\frac{2}{k^2} \geq P\left(|\hat{\mu}^{tr}_{c_k} - \hat{\mu}^{te}_{c_k}| \geq \frac{2k\sigma_{c_k}}{\sqrt{n_{te}}}\right) \quad (5)$$

Hence, for any arbitrary k, the data variance σ_{c_k} reflected by $\mathrm{H}_{intra}(c_k)$, as depicted in Equation 4, serves as an estimation for the distributional gap $|\hat{\mu}^{tr}_{c_k} - \hat{\mu}^{te}_{c_k}|$, indicating the overfitting degree of c_k.

5.2 Cross-embeddings generalization

We incorporate two other architectures of sentence embeddings, i.e., GTE-large (Li et al., 2023) and BGE-large-en-v1.5 (Xiao et al., 2023) into *GeoHard*, substituting E5-large-v2. Observing Figure 5, we find a consistent trend of *GeoHard*'s measurement across different sentence embeddings. The significant gap between *GeoHard* and the instance-level aggregation underscores the robustness of *GeoHard* as a semantic-guided metric.

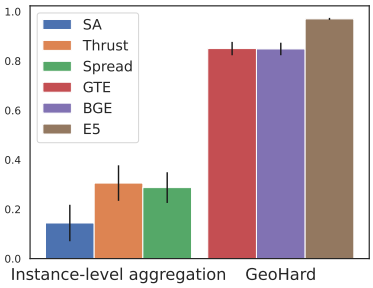


Figure 5: Average *Pearson*'s coefficient between various metrics and hardness reference on five SC tasks. *GeoHard* with different embeddings consistently and significantly outperform instance-level aggregation, demonstrating the robustness of *GeoHard*.

	AG News	Yahoo	Emo	CARAR
GeoHard	-.980±.0	-.838±.0	-.798±.1	-.817±.1

Table 3: Pearson's correlation coefficients, averaged on 3 seeds, between class-wise hardness measured by *GeoHard* and class-wise F1 scores, i.e., the reference of hardness, on topic classification and emotion detection.

5.3 Cross-task generalization

Complementary to the theoretical generalization, we further validate *GeoHard* on other tasks beyond SC and NLI, i.e., topic classification and emotion detection. We include AG News (Zhang et al., 2015), Yahoo Answer Topic (Yahoo; Zhang et al. 2015) for the former and Emo2019 (Emo; Chatterjee et al. 2019), Contextualized Affect Affect Representations for Emotion Recognition (CARAR; Saravia et al. 2018) for the latter.

We fine-tune Roberta-Large on these four datasets to obtain the reference empirical hardness, i.e., class-wise F1 scores, and also conduct *GeoHard*, referring to Table 16-19 in Appendix D.3. According to Table 3, the consistency between the measurement and reference on the tasks other than NLI and SC empirically exhibits the generalization of *GeoHard* for class-wise hardness measurement.

5.4 Robustness in low-resource scenarios

In this section, we will demonstrate the robustness of our method in low-resource scenarios. We have randomly selected 1%, 10%, and 100% of the instances from the training corpus across five SC datasets included in Section 4. As illustrated in Figure 6, *GeoHard* exhibits notably less performance degradation in low-resource settings compared to `PVI` (Ethayarajh et al., 2022), underscoring its robustness as a training-free method.

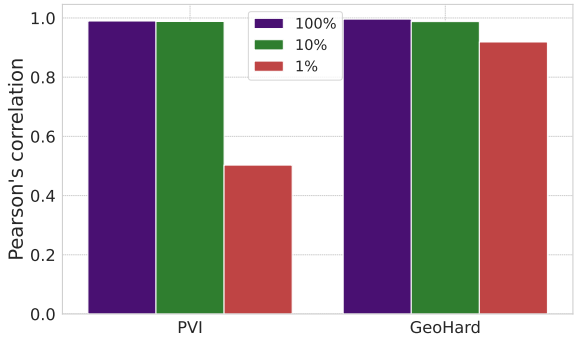
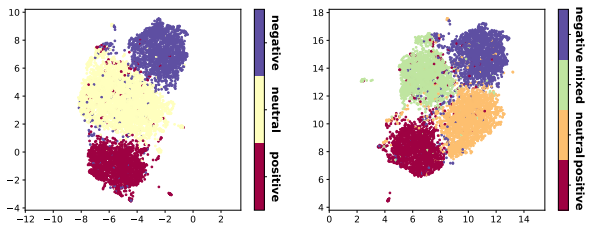


Figure 6: Performance comparison of *GeoHard* and PVI in low-resource scenarios: *GeoHard* experiences less degradation on the average Pearson's correlation (absolute values) across five SC datasets with 1% of the training data compared with the full training data.



(a) 3-way (b) 4-way

Figure 7: Illustration of the class-wise geometrical distribution of 3-way and 4-way Dynasent by splitting *Neutral* to *Mixed* and *Neutral* while maintaining *Positive* and *Negative*. The reorganized class *Mixed* and *Neutral* are highly separable in the representation space.

6 Why class-wise hardness measurement?

In the previous sections, we establish the concept of class-wise hardness, which can be well and robustly measured by *GeoHard*. One of the most relevant literature for the application of *GeoHard* is class reorganization.

Class reorganization has received relatively limited attention compared to research focusing on addressing class imbalance (Subramanian et al., 2021; Henning et al., 2023), primarily due to extra annotation. However, initial task formulations are rarely perfect, and as research progresses, class reorganization becomes necessary for a more comprehensive understanding and effective modeling of the task. For example, in NLI, the task evolved from a 2-way classification (Dagan et al., 2005) to a 3-way classification by separating *Non-Entailment* into *Neutral* and *Contradiction*. Recently, Nighojkar et al. (2023) further subdivided *Neutral* into two distinct classes based on human disagreement.

As for the practical perspective, class reorganization can balance the model performance among the classes (Potts et al., 2021). To resolve the severe imbalance between *Neutral* and other classes in SC dataset, as shown in Table 1, Dynasent (Potts et al., 2021) opted to split *Neutral* to *Mixed* (a mixture of positive and negative sentiment) and *Neutral* (conveying nothing regarding sentiment). This approach aims to achieve a **coherent** categorization, which narrows the performance gap among classes.

In this section, we use Dynasant (Potts et al., 2021) as an example to demonstrate how measuring class-wise hardness can aid in interpreting class reorganization and facilitate the learning process.

6.1 *GeoHard* interprets class reorganization

GeoHard can provide insights into two crucial questions regarding class reorganization: *what* and *how* to reorganize classes. Firstly, even without training a model, *GeoHard* can directly provide hardness estimates across the original classes to locate the operating target. Secondly, *GeoHard* can assess the effectiveness of the formulation strategy and hence guide the class reorganization.

We conduct *GeoHard* on classes before and after Dynasent's class reorganization, which splits *Neutral* to *Mixed* and *Neutral*. For comparison, we randomly split *Neutral* into two classes, labeled *Rand1* and *Rand2*. As illustrated in Figure 7, the *Neutral* and *Mixed* are shown to be highly separable, indicating their distinction in semantics.

As shown in Table 4, the new classes *Mixed* and *Neutral* exhibit lower class-wise hardness compared to the original *Neutral*. The standard derivation among class-wise hardness on the newly organized Dynasent drops by 40.9% (from 1.15 to 0.68). This clearly explains the **coherent** class formation by reorganizing *Neutral* into two sub-classes. However, not all class reorganizations yield beneficial outcomes: a random split may result in high overall class-wise hardness and a severe imbalance of class-wise hardness.

6.2 *GeoHard* propels task learning

We have demonstrated that *GeoHard* can interpret and validate the reorganization of labels. Next, we further investigate how to leverage the class-wise hardness knowledge and its induced class reorganization to enhance task learning, with methods such as ICL. Typically, ICL samples the demonstrations uniformly across classes (Min et al., 2022). Here,

Positive	*Neutral*		*Negative*	**Std**
-5.58	-3.08		-5.48	1.15
Positive	***Rand1***	***Rand2***	***Negative***	**Std**
-3.89	-0.71	-0.71	-3.62	1.53
Positive	***Mixed***	***Neutral***	***Negative***	**Std**
-5.24	-4.26	-3.41	-4.78	**0.68**

Table 4: Class-wise *GeoHard* on the 3-way and reorganized Dynasent (randomly and semantic-guided splits). The results are averaged on 3 seeds. Note that larger *GeoHard* indicates more hardness on the class. **Bold** signifies the smallest standard derivation among class hardness. Underline indicates the lowest hardness on *Neutral* and its splits.

we demonstrate the benefits of splitting the **hardest** class into two **easier** ones in ICL, elucidating the significance of class-wise hardness.

We divide each class into two sub-classes and select instances from these newly formed classes. For classes *Positive* and *Negative*, which lack prepared sub-classes, we employ KMeans on the embeddings to separate instances within each class into two sub-classes (Zhang et al., 2023; Yang et al., 2023). Then, we select the center of each cluster as a demonstration. For example, if *Positive* is selected for reorganization, the demonstrations consist of 2 *Positive* instances, 1 *Neutral* instance, and 1 *Negative* instance. For convenience, we abbreviate the selection as 2P+1NEU+1N.

We randomly sample 1,000 instances from each class from 3-way Dynasent (Potts et al., 2021), wherein *Neutral* class contains 500 *Mixed* instances and 500 new-formed *Neutral* instances. We conduct ICL on two popular LLMs, i.e., OPT-6.7B (Zhang et al., 2022) and Llama-2-7B-32K-Instruct (Touvron et al., 2023) with different setups of demonstrations: (1) even sampling: 1P+1NEU+1N; (2) sampling based on class reorganization: 2P+1NEU+1N, 1P+2NEU+1N, and 1P+1NEU+2N. For both setups, we select the centroid instance from each cluster. The examples of demonstrations can be found in Appendix D.5.

As shown in Table 5, both models utilizing the setup 1P+2NEU+1N attain the best performance, which is the advocated action from *GeoHard* since *Neutral* is measured as the hardest class and the new classes *Mixed* and *Neutral* are relatively easier. However, reconstructing other classes may not lead to such benefits in learning and can even lead to significant degradation (e.g., 2P with LLama-7B).

Demonstration %	OPT-6.7B	LLama-7B
1P+1NEU+1N	61.7±3.54	61.4±1.82
2P+1NEU+1N	61.1±0.02	39.6±10.71
1P+2NEU+1N	**64.3**±1.24	**69.9**±1.74
1P+1NEU+2N	60.4±1.69	34.9±3.36

Table 5: Comparison of different compositions of demonstrations on Dynasent, with each entry presenting the prediction accuracy. **Bold** indicates the highest accuracy on one specific model. The results are averaged on three seeds for random initialization in KMean.

7 Related works

Hardness in NLP datasets Instance-level hardness indicates the difficulty of an instance given a distribution (Ethayarajh et al., 2022), and the taxonomy is summarized in Figure 9 in Appendix B.1. Without training, the reference model or embedding is usually needed. With a model as the reference, Sensitivity Analysis (Hahn et al., 2021; Chen et al., 2023) assesses hardness by perturbing input features and observing the resulting changes in model predictions. Additionally, Thrust (Zhao et al., 2023a) approximates instance hardness based on the external knowledge required by an LLM. In parallel with the model reference, Spread (Zhao et al., 2022) leverages the similarity between test and training samples in the space of semantic embeddings for hardness measurement. Alternatively, information theory-based methods, such as pointwise $\mathcal{V}$-usable information (PVI; Ethayarajh et al. 2022) and Rissanen Data Analysis (RDA; Perez et al. 2021), offer insights into data hardness using training outcomes. Moreover, other methods measure data hardness from training dynamics, including dataset cartography (Swayamdipta et al., 2020), forgetting scores (Toneva et al., 2019), and Error L2-Norm (Paul et al., 2021), etc [9]. This work primarily focuses on the training-free methods, which are more practical and scalable for gauging hardness with LLMs.

Although instance-level hardness is well studied, class-wise hardness is under-explored. Therefore, our work explores the class-wise measurement by aggregating the existing instance-level methods first and then specifically designs *GeoHard*, which requires no additional data or training.

[9]We include the class-wise hardness measurement with some training-based methods and training-dynamics methods in Appendix D.4

Geometrical view of classification complexity In the context of general machine learning, prior research (Ho and Basu, 2002; Lorena et al., 2019) assesses the difficulty of a classification problem through the analysis of data geometry and inter- and intra-class distribution. Various metrics of quantification, such as Fisher's discriminant ratio (Cummins, 2013), overlapping regions (Seijo-Pardo et al., 2019) and network measures (Garcia et al., 2015), have been proposed to qualify class-wise complexity based on geometric features.

Sentences encoders (Reimers and Gurevych, 2019) excel at generating high-dimension sentence embeddings based on semantics. We explore class-wise hardness by leveraging geometrical features within and among the classes, inspired by *Neutral*'s specific semantics.

Neutral* in NLU** *Neutral* depicts undetermined or middlemost semantics while ruling out other classes, and widely exists in NLU tasks such as NLI (Williams et al., 2018; Bowman et al., 2015) and SC (Sun et al., 2019). Generally, the class with the prefix *Non-* also delivers similar semantics with *Neutral*, i.e., excluding other classes. For instance, the Microsoft Research Paraphrase (MRPC; Dolan and Brockett 2005) dataset aims to determine whether a pair of questions are semantically equivalent, i.e., to classify sentence pairs to *Equivalent* and ***Non-*equivalent*. In GLUE (Wang et al., 2019)), six of nine tasks contain a *Neutral* or *Non-* class, indicating the wide existence of classes with undetermined semantics in NLU.

Due to *Neutral*'s semantic prevalence, we initiate class-wise hardness from exploring the tasks containing *Neutral* and then extend to general classes.

8 Conclusion

In this work, to study how class-specific properties influence model learning, we initiate the notion of class-wise hardness analogous to instance-level hardness. The consistent pattern observed across various LMs, learning paradigms, and human annotations on eight NLU datasets affirms the presence of class-wise hardness as an inherent property. In addressing the challenge of estimating class-wise hardness, conventional instance-level metrics fall short, necessitating a tailored approach to measure hardness specifically at the class level. Thus, we introduce *GeoHard*, which models both *inter-* and *intra-*class semantics, surpassing instance-level aggregation by 59%. Moreover, *GeoHard*, formulated with a foundation in semantics, demonstrates robust generalization properties, as validated both theoretically and empirically. Lastly, we showcase the potential of *GeoHard* in reorganizing classes and enhancing task-learning methodologies. We recommend more attention to class-wise hardness and exploring its potential across a broader range of scenarios.

Limitations

Our work, introducing the concept of class-wise hardness and proposing a practical metric, does come with specific limitations that justify further exploration. Firstly, as an initiative work, we only cover limited types of classification tasks in NLU due to the space constraint. Some classification problems such as sequence labeling (He et al., 2020) are not covered in our scope. Class-wise hardness for other formats of classification tasks is still obscure and needs further exploration. Secondly, as our proposed method *GeoHard* is built upon the pre-trained sentence encoders, they inherit their corresponding limitations. For example, it will be intricate to measure class-wise hardness in the complex semantics or long inputs. These cases can not be well modeled by a single-sentence encoder yet. Combining the two problems above leads to a new issue. Hypothetically, given the assumption that the class-wise hardness for other formats of NLP problems still exists, how to model them will be the potential concern, as it is beyond the capacity of sentence encoders. Regarding the application of *GeoHard* and the class-wise hardness that it measures, we have not gone deeper into this problem. A larger-scale study is expected to further explore this topic. In conclusion, further efforts are expected to overcome the limitations of this work.

Ethical Statements

We foresee no major ethical concerns in our work. The datasets we used in this work are all publicly available. As far as we see, there is no sensitive information included. For the language models we applied, the outputs, i.e., the class labels, are not sensitive either.

Acknowledgement

We thank Sheng Lu, Kexin Wang, Indraneil Paul, Hendrik Schuff, and Sherry Tongshuang Wu for their feedback on an early draft of this work.

Fengyu Cai is funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science, and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Xinran Zhao is funded by the ONR Award N000142312840.

References

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 155–167, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Lisa Cummins. 2013. *Combining and choosing case base maintenance algorithms*. Ph.D. thesis, University College Cork, Republic of Ireland.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Luís Paulo F. Garcia, André C. P. L. F. de Carvalho, and Ana Carolina Lorena. 2015. Effect of label noise in the complexity of classification problems. *Neurocomputing*, 160:108–119.

Giovanni Grano, Andrea Di Sorbo, Francesco Mercaldo, Corrado Aaron Visaggio, Gerardo Canfora, and Sebastiano Panichella. 2017. Android apps and user feedback: a dataset for software evolution and quality improvement. In *WAMA@ESEC/SIGSOFT FSE*, pages 8–11. ACM.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2021. Sensitivity as a complexity measure for sequence classification tasks. *Transactions of the Association for Computational Linguistics*, 9:891–908.

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.

Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. 2020. A survey on recent advances in sequence labeling from deep learning models. *ArXiv preprint*, abs/2011.06727.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Tin Kam Ho and Mitra Basu. 2002. Complexity measures of supervised classification problems. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):289–300.

John D Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95.

Rishi Jha and Kai Mihata. 2021. On geodesic distances and contextual embedding compression for text classification. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 144–149, Mexico City, Mexico. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *ArXiv preprint*, abs/2308.03281.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.

Ana C Lorena, Luís PF Garcia, Jens Lehmann, Marcilio CP Souto, and Tin Kam Ho. 2019. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

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.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dragoslav S Mitrinovic, Josip Pecaric, and Arlington M Fink. 2013. *Classical and new inequalities in analysis*, volume 61. Springer Science & Business Media.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Animesh Nighojkar, Antonio Laverghetta Jr., and John Licato. 2023. No strong feelings one way or another: Re-operationalizing neutrality in natural language inference. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 199–210, Toronto, Canada. Association for Computational Linguistics.

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 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.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8500–8513. PMLR.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. 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 2388–2404, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Borja Seijo-Pardo, Verónica Bolón-Canedo, and Amparo Alonso-Betanzos. 2019. On developing an automatic threshold applied to feature selection ensembles. *Information Fusion*, 45:227–245.

Nishit Shrestha and Fatma Nasoz. 2019. Deep learning sentiment analysis of amazon.com reviews and ratings. *ArXiv preprint*, abs/1904.04096.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Fairness-aware class imbalanced learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2045–2051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1521–1528. IEEE Computer Society.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Kailas Vodrahalli, Ke Li, and Jitendra Malik. 2018. Are all training examples created equal? an empirical study. *ArXiv preprint*, abs/1811.12569.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv preprint*, abs/2212.03533.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *ArXiv preprint*, abs/2309.07597.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456, Singapore. Association for Computational Linguistics.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2023. In-context instruction learning. *ArXiv preprint*, abs/2302.14691.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2024. Investigating the effectiveness of task-agnostic prefix prompt for instruction following. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19386–19394.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Xinran Zhao, Shikhar Murty, and Christopher Manning. 2022. On measuring the intrinsic few-shot hardness of datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3955–3963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, and Jianshu Chen. 2023a. Thrust: Adaptively propels large language models with external knowledge. *ArXiv preprint*, abs/2307.10442.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, and Jianshu Chen. 2023b. Thrust: Adaptively propels large language models with external knowledge. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *ArXiv preprint*, abs/2305.11206.

Appendix

A Validation of class-wise hardness

A.1 Dataset

Dataset	Description	Statistics (train/dev/test)
Amazon Review Multi en † (Amazon; Keung et al. 2020)	an Amazon product reviews dataset for multilingual text classification (we only use English part)	120,000 / 3,000 / 3,000
App Reviews † (APP; Grano et al. 2017)	Android app reviews categorized classifying types of user feedback from a software maintenance and evolution perspective	56,151 / 6,804 / 6,633
MultiNLI ‡ (MNLI; Williams et al. 2018)	Multi-Genre Natural Language Inference annotated with textual entailment information	353,408 / 39,270 / 9,369
SICK-E ‡ (Marelli et al., 2014)	A dataset targeting Natural Language Inference	1,920 / 213 / 2,136
SNLI ‡ (Bowman et al., 2015)	Stanford Natural Language Inference Corpus	548,292 / 9,705 / 9,657
SST-5 † (Socher et al., 2013)	Stanford Sentiment Treebank with 5 labels	4,872 / 1,332 / 1,332
Twitter Financial News Sentiment † [10] (TFNS)	A dataset is used to classify finance-related tweets for their sentiment	3,891 / 435 / 1,041
Yelp review † (Yelp; Zhang et al. 2015)	A dataset containing custom reviews from Yelp	351,000 / 39,000 / 30,000

Table 6: The description of the datasets used in the class-wise hardness measurement, together with the statistics of newly-formulated datasets after balancing the number of instances in each class. All the datasets consist of 3 classes, namely *Positive*/*Neutral*/*Neagtive* in SC † and *Entailment*/*Neutral*/*Contradiction* in NLI ‡.

A.2 Dataset Normalization

We follow the setup of the previous work (Shrestha and Nasoz, 2019) to convert 5 degrees of sentiment to 3 classes: for Amazon, APP, and Yelp, we map 1 and 2 to *Negative*, 3 to *Neutral*, and 4 and 5 to *Positive*; for SST-5, we map *very positive* and *positive* to *Positive*, and *negative* and *very negative* to *Negative*. for TFNS, the original class labels *Bearish* and *Bullish* are mapped to *Negative* and *Positive*.

[10] https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment

Dataset	Example	Original labels	Original statistics
Amazon Review Multi en (Amazon; Keung et al. 2020)	Title: bubble Body: went through 3 in one day doesn't fit correct and couldn't get bubbles out (better without)	1, 2, 3, 4, 5	200,000 / 5,000 / 5,000
App Reviews (APP; Grano et al. 2017)	simple and perfect About this software rtl sdr is very useful ... installed done. Thanks.	1, 2, 3, 4, 5	230,452 / 28,806 / 28,807
MultiNLI (MNLI; Williams et al. 2018)	Premise: I burst through a set of cabin doors, and fell to the ground. Hypothesis: I burst through the doors and fell down.	Entailment/Neutral/ Contradiction	353,431 / 39,271 / 9,815
SICK-E (Marelli et al., 2014)	Sentence A: A group of kids is playing in a yard and an old man is standing in the background Sentence B: A group of boys in a yard is playing and a man is standing in the background	Entailment/Neutral / Contradiction	4,439 / 495 / 4,906
SNLI (Bowman et al., 2015)	Text: A soccer game with multiple males playing. Hypothesis: Some men are playing a sport.	Entailment/Neutral/ Contradiction	549,367 / 9,842 / 9,824
SST-5 (Socher et al., 2013)	a metaphor for a modern-day urban china searching for its identity .	very positive/positive/ neutral / negative / very negative	8,544 / 1,101 / 2,210
Twitter Financial News Sentiment [1]	"$BYND - JPMorgan reels in expectations on Beyond Meat https://t.co/bd0xbFGjkT"	Bearish/ Bullish / Neutral	8,587 / 956 / 2,388
Yelp review (Yelp; Zhang et al. 2015)	Tonya is super sweet and the front desk people are very helpful	1, 2, 3, 4, 5	585,000 / 65,000 / 50,000

Table 7: The examples for the given datasets and the original label and statistics before reformatting.

A.3 Experimental Setup

The seeds of training are {1, 10, 100}, and the learning rate is 1e-5. The detailed configuration of **Roberta-Large**, **OPT-350M** and **Flan-T5-Large**, including training are shown at Table 8, 9, and 10. All the experiments are conducted on a single NVIDIA A100.

Datasets	Batch size	Epochs	Seq. length
Amazon	16	10	512
APP	16	10	256
MNLI	64	10	128
SICK-E	16	10	256
SST-5	16	10	128
SNLI	20	5	128
TFNS	16	10	128
Yelp	24	5	256

Table 8: Training configuration of **Roberta-Large**.

Datasets	Batch size	Epochs	Seq. length
Amazon	6	3	256
APP	16	10	256
MNLI	64	5	128
SICK-E	16	10	256
SNLI	64	5	128
SST-5	64	10	128
TFNS	64	10	128
Yelp	16	5	256

Table 9: Training configuration of **OPT-350M**

Datasets	Batch size	Epochs	Seq. length
Amazon	6	3	256
APP	6	5	256
MNLI	12	5	128
SICK-E	6	10	256
SNLI	12	5	128
SST-5	12	10	128
TFNS	12	10	128
Yelp	6	5	256

Table 10: Training configuration of **Flan-T5**.

A.3.1 *Neutral*'s hardness in human disagreement

We present the human variation on the classification as a hardness measurement from human beings. Table 11 presents the distribution of human disagreement of MNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015). The high entropy of *Neutral* reveals its class-wise hardness for humans. For the convenient comparison with other metrics, we take the negative of the entropy to obtain the positive correlation between the knowledge, as shown in Figure 2 and 8.

Dataset \ Class	Entailment	Neutral	Contradiction
MNLI	0.3202	**0.4717**	0.2664
SNLI	0.3515	**0.5175**	0.2781

Table 11: Average entropy of annotation distribution for the instances belonging to the same class in MNLI and SNLI. **Bold** indicates the highest entropy score.

A.3.2 *Neutral*'s hardness in LLMs

Regarding the hardness of *Neutral* w.r.t. LLMs, we conduct two families of LLMs, **Flan-T5-XXL** (Raffel et al., 2020) and **LLaMA-2-13B** (Touvron et al., 2023). The prompting templates for MNLI and SNLI present as follows:

The prompt for **Flan-T5**:
{*premise*}. Does this imply {*hypothesis*}? options:
entailment
contradiction
neutral

The prompt for **LLaMA-2-13B**:
Input: {*premise*} Question: Does this imply that {*hypothesis*}? Please respond with 'Entailment', 'Contradiction', or 'Neutral'. Result:

Models \ Datasets	MNLI			SNLI		
	Entailment	**Neutral**	**Contradiction**	**Entailment**	**Neutral**	**Contradiction**
Flan-T5-XXL	0.90	0.87	0.94	0.90	0.88	0.93
LLaMA-2-13B	0.51	0.28	0.50	0.41	0.39	0.55

Table 12: F1 scores of in-context learning using **Flan-T5-XXL** and **LLaMA-2-13B** on MNLI and SNLI. **Bold** indicates the poorest performance across the class. The results are averaged on the seeds {100, 200, 300}

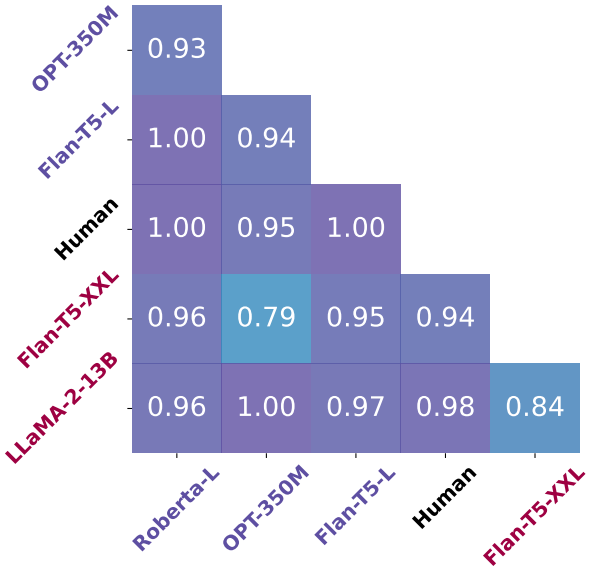


Figure 8: Correlation matrix among class-wise F1 scores of three finetuned models together with two ICLs and class-wise human disagreement on MNLI, where the high consistency is noted.

B Hardness measurement

B.1 Taxonomy of hardness measurement

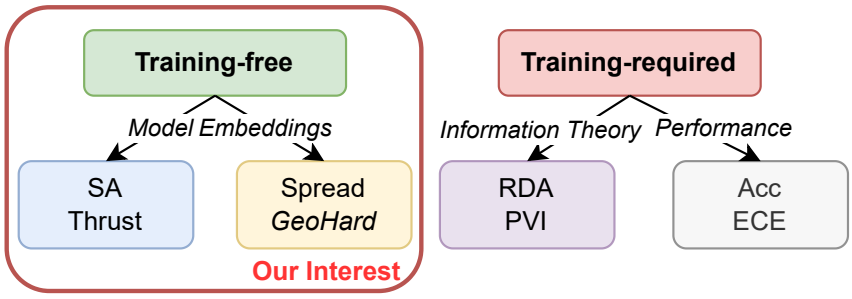


Figure 9: Taxonomy of hardness measurement and the scope of this work.

B.2 SA

The technical steps of SA are as follows:

1) Train a Roberta-Large model θ with $\mathcal{D}_{train}$ and evaluate the model with $D_{test} = \{(X_i^{test}, y_i^{test})\}$;

2) By randomly masking several consecutive words on X_i and reconstructing K samples with LMs, generate perturbed test dataset $D'_{test} = \{(X'_{ij}, y_i)\}$, where $j = 1, ..., K$;

3) Calculate the confidence for each disturbed input X'_{ij} on the golden label y_i with θ;

4) The sensitivity for each input X_i is then defined as the maximum value of the deviation between the confidence values for the original X_i and the corresponding perturbed samples X'_{ij}.

For each dataset, sensitivity values are averaged on the three Roberta-Large models trained in Table 1.

B.3 Thrust

Thrust (Zhao et al., 2023a) measures how likely the query to LLMs can be solved by the internal knowledge of the target model, in other words, how necessary the knowledge is needed to propel the model's inference. There are two essential assumptions of Thrust: (1) LLMs are expected to well study the given tasks. (2) Meanwhile, the particular samples deviate from the output embeddings of LLMs, mainly due to insufficient knowledge of LLMs.

We denote the representation function, namely the decoder of UnifiedQA-Flan-T5-Large, as $f_{thrust}(\cdot)$. We sample a certain number of instances from the datasets, i.e., D^{sample}. Concretely, for the task of sentiment classification, $D^{sample} = 200$, and for the task of natural language inference, $D^{sample} = 600$. Based on the representation obtained, the samples belonging to the identical classes are grouped together as $\mathcal{G}_l = \{(f_{thrust}(x_i), y_i,)|y_i = l\}$, where (x_i, y_i) are the sampled instances, and l is the class index. Then, each G_l is clustered to K clusters by the k-means algorithm, and each cluster and its corresponding centroid are denoted as C_{kl} and m_{kl}, respectively. Regarding the selection of cluster numbers K, we refer to the original setup (Zhao et al., 2023a), i.e., $K = max(ceil(\sqrt[4]{|D^{sample}|}), 3)$. The seeds for sample selection and clustering initialization are both $\{2, 4, 42, 102, 144\}$, and hence the results are averaged on 25 initial setups.

$$s_{thrust}(q) = \left\| \frac{1}{N \cdot K} \sum_{l=1}^{N} \sum_{k=1}^{K} \frac{|\mathcal{C}_{kl}|}{\|d_{kl}(q)\|^2} \cdot \frac{d_{kl}(q)}{\|d_{kl}(q)\|} \right\|$$

where q denotes the query, namely, the test instance, and $d_{kl} = m_{kl} - f(q)$ is a vector pointing from $f(q)$ towards the centroid m_{kl}.

The prompts for NLI and SC tasks used on the model are shown as follows:

The prompt for **NLI** tasks:
{*premise*}. And {*hypothesis*}. What is the relationship between these two sentences? Option: Entailment or Neutral or Contradiction. Answer:

The prompt for **SC** tasks:
{sentence}. Is it a happy review? Answer:

B.4 Spread

Spread aims to measure the instance-level hardness in the few-shot scenario (Zhao et al., 2022). The idea of Spread is to examine the similarity between training and test instances. Concretely, if one test sample is close to train samples semantically, it is taken as an easy instance. We denote the semantic encoder for Spread as f_{Spread}. The distance of one test instance to the training set is defined as the average distance between the test instance to the k-closest training instances. Let $D^{tr} = \{(x_i^{tr}, y_i^{tr})\}$ and $D^{te} = \{(x_i^{te}, y_i^{te})\}$ denote the training and test sets, respectively. x_{ik}^{tr} denotes the k-th closest training instances to the test instance x_i^{te}. K_{shot} is the number of shots to the training sets, and $d(\cdot, \cdot)$ is the measurement between two data points.

$$s_{Spread}(x_i^{te}) = \frac{1}{K_{shot}} \sum_{k=1}^{K_{shot}} d(x_i^{te}, x_{ik}^{tr})$$

B.5 PVI

Algorithm 1 presents the procedure of PVI (Ethayarajh et al., 2022). g' is fine-tuned on the original training dataset $\mathcal{D}$, i.e., $\{(X_i, y_i) | \forall (X_i, y_i) \in \mathcal{D}\}$. Meanwhile, g is fine-tuned on the null-target pairs $\{(\varnothing, y_i) | \forall (X_i, y_i) \in \mathcal{D}\}$, where $\varnothing$ is an empty string.

Algorithm 1 PVI calculation

Input: a dataset $\mathcal{D} = \{(X_{1:N}, y_{1:N})\}$, a model $\mathcal{G}$, and the test instance of (X^{test}, y^{test})

1: $g' \leftarrow$ fine-tune $\mathcal{G}$ on $\mathcal{D}$
2: $g \leftarrow$ fine-tune $\mathcal{G}$ on $\{(\varnothing, y_i) | \forall (X_i, y_i) \in \mathcal{D}\}$
3: PVI(X^{test}, y^{test}) $\leftarrow -\log_2 g[\varnothing](y^{test}) + \log_2 g'[X^{test}](y^{test})$

C Experimental results

As a supplement of Table 2, the following Table 13 presents the fine-grained numerical values of the golden hardness and different metrics.

C.1 *GeoHard*

C.1.1 NLI's fine-grained results with different connecting words or phrases

Table 14 and 15 present the measurement of *Distributional complexity* and *Biased gravity*. Different from the SC datasets, a pair of sentences is in the NLI task. Therefore, a conjunction word is needed to convert a pair of sentences to a natural sentence. As shown in Table 14 and 15, six conjunctive words or phrases are selected, including *And*, *It is true that*, etc. As mentioned above, we average the metrics on different conjunctions to measure the NLI sentence pair.

Datasets	Class	F1(↓)	Sensitivity(↑)	Thrust(↑)	Spread(↑)	Intra-class (↑)	Inter-class (↑)	*GeoHard*
Amazon	Positive	87.6$_{\pm 0.41}$	0.1708$_{\pm 0.0121}$	0.455$_{\pm 0.004}$	0.839	2.837$_{\pm 0.003}$	-11.152$_{\pm 0.096}$	-8.316$_{\pm 0.096}$
	Neutral	71.0$_{\pm 0.96}$	0.2511$_{\pm 0.0222}$	0.575$_{\pm 0.008}$	0.842	2.968$_{\pm 0.007}$	-6.786$_{\pm 0.056}$	-3.818$_{\pm 0.061}$
	Negative	80.5$_{\pm 0.69}$	0.3153$_{\pm 0.0158}$	0.513$_{\pm 0.027}$	0.844	2.712$_{\pm 0.007}$	-9.205$_{\pm 0.071}$	-6.493$_{\pm 0.076}$
APP	Positive	74.2$_{\pm 0.16}$	0.245$_{\pm 0.033}$	0.54$_{\pm 0.024}$	0.876	5.359$_{\pm 0.012}$	-7.28$_{\pm 0.72}$	-1.921$_{\pm 0.709}$
	Neutral	60.1$_{\pm 0.14}$	0.1368$_{\pm 0.0062}$	0.447$_{\pm 0.014}$	0.864	6.251$_{\pm 0.012}$	-4.945$_{\pm 0.342}$	1.306$_{\pm 0.353}$
	Negative	73.4$_{\pm 0.8}$	0.2724$_{\pm 0.0252}$	0.513$_{\pm 0.027}$	0.862	5.668$_{\pm 0.05}$	-7.24$_{\pm 0.72}$	-1.571$_{\pm 0.77}$
MNLI	Entailment	91.1$_{\pm 0.12}$	0.8233$_{\pm 0.01}$	1.503$_{\pm 0.021}$	0.837	4.013$_{\pm 0.008}$	-0.049$_{\pm 0.005}$	3.964$_{\pm 0.012}$
	Neutral	87.2$_{\pm 0.12}$	0.496$_{\pm 0.0085}$	1.503$_{\pm 0.021}$	0.832	4.037$_{\pm 0.009}$	-0.062$_{\pm 0.007}$	3.975$_{\pm 0.015}$
	Contradiction	92.8$_{\pm 0.05}$	0.6828$_{\pm 0.0142}$	1.503$_{\pm 0.021}$	0.833	3.989$_{\pm 0.01}$	-0.07$_{\pm 0.002}$	3.92$_{\pm 0.012}$
SICK-E	Entailment	92.9$_{\pm 0.22}$	0.8968$_{\pm 0.0231}$	1.589$_{\pm 0.07}$	0.859	3.112$_{\pm 0.026}$	-2.205$_{\pm 0.172}$	0.906$_{\pm 0.187}$
	Neutral	86.8$_{\pm 2.11}$	0.3036$_{\pm 0.0431}$	1.589$_{\pm 0.07}$	0.854	3.471$_{\pm 0.013}$	-2.363$_{\pm 0.163}$	1.108$_{\pm 0.171}$
	Contradiction	92.4$_{\pm 0.46}$	0.9049$_{\pm 0.0106}$	1.589$_{\pm 0.07}$	0.863	2.197$_{\pm 0.002}$	-4.135$_{\pm 0.312}$	-1.937$_{\pm 0.312}$
SNLI	Entailment	92.6$_{\pm 0.08}$	0.8712$_{\pm 0.0155}$	1.582$_{\pm 0.011}$	0.877	5.64$_{\pm 0.018}$	-0.069$_{\pm 0.013}$	5.571$_{\pm 0.03}$
	Neutral	89.2$_{\pm 0.17}$	0.6938$_{\pm 0.0205}$	1.582$_{\pm 0.011}$	0.87	5.645$_{\pm 0.02}$	-0.064$_{\pm 0.015}$	5.581$_{\pm 0.035}$
	Contradiction	95.3$_{\pm 0.08}$	0.5447$_{\pm 0.0188}$	1.88$_{\pm 0.179}$	0.864	5.596$_{\pm 0.021}$	-0.104$_{\pm 0.035}$	5.491$_{\pm 0.056}$
SST-5	Positive	83.1$_{\pm 0.73}$	0.2242$_{\pm 0.0389}$	0.821$_{\pm 0.014}$	0.828	1.648$_{\pm 0.022}$	-7.665$_{\pm 0.366}$	-6.017$_{\pm 0.346}$
	Neutral	53.1$_{\pm 1.55}$	0.2347$_{\pm 0.0667}$	0.801$_{\pm 0.013}$	0.824	1.904$_{\pm 0.01}$	-5.014$_{\pm 0.261}$	-3.11$_{\pm 0.27}$
	Negative	75.8$_{\pm 1.7}$	0.364$_{\pm 0.0451}$	0.764$_{\pm 0.022}$	0.83	1.68$_{\pm 0.023}$	-7.376$_{\pm 0.416}$	-5.696$_{\pm 0.418}$
TFNS	Positive	93.0$_{\pm 0.08}$	0.3627$_{\pm 0.0585}$	0.553$_{\pm 0.035}$	0.818	2.25$_{\pm 0.025}$	-7.836$_{\pm 0.176}$	-5.587$_{\pm 0.191}$
	Neutral	86.1$_{\pm 0.29}$	0.1292$_{\pm 0.0133}$	0.523$_{\pm 0.03}$	0.806	2.391$_{\pm 0.02}$	-6.204$_{\pm 0.301}$	-3.813$_{\pm 0.287}$
	Negative	92.2$_{\pm 0.54}$	0.4689$_{\pm 0.0169}$	0.752$_{\pm 0.033}$	0.819	2.769$_{\pm 0.083}$	-7.481$_{\pm 0.163}$	-4.712$_{\pm 0.144}$
Yelp	Positive	87.9$_{\pm 0.14}$	0.0451$_{\pm 0.0021}$	0.455$_{\pm 0.006}$	0.822	4.043$_{\pm 0.011}$	-9.793$_{\pm 0.012}$	-5.75$_{\pm 0.016}$
	Neutral	75.4$_{\pm 0.25}$	0.0832$_{\pm 0.0033}$	0.395$_{\pm 0.006}$	0.819	4.396$_{\pm 0.003}$	-6.328$_{\pm 0.009}$	-1.931$_{\pm 0.006}$
	Negative	86.6$_{\pm 0.12}$	0.0639$_{\pm 0.0036}$	0.455$_{\pm 0.006}$	0.811	4.108$_{\pm 0.015}$	-9.19$_{\pm 0.018}$	-5.082$_{\pm 0.033}$

Table 13: The class hardness measurement on the tasks containing the undetermined class *Neutral* using SA, Spread, and Thrust to class hardness measurement. ↓ indicates that the lower value reflects more hardness, while ↑ indicates that the higher value reflects more hardness.

Datasets	Class	*Maybe*	*And*	*Therefore*	*But*	*On the other hand*	*It is true that*	**Average**
MNLI	Positive	4.078$_{\pm 0.003}$	4.04$_{\pm 0.008}$	3.956$_{\pm 0.024}$	4.017$_{\pm 0.029}$	4.016$_{\pm 0.001}$	3.972$_{\pm 0.05}$	4.013
	Neutral	4.099$_{\pm 0.011}$	4.061$_{\pm 0.008}$	3.976$_{\pm 0.026}$	4.037$_{\pm 0.035}$	4.056$_{\pm 0.001}$	3.993$_{\pm 0.052}$	4.037
	Negative	4.063$_{\pm 0.001}$	4.017$_{\pm 0.009}$	3.906$_{\pm 0.018}$	3.99$_{\pm 0.03}$	4.013$_{\pm 0.001}$	3.946$_{\pm 0.061}$	3.989
SNLI	Positive	6.256$_{\pm 0.058}$	5.738$_{\pm 0.064}$	5.314$_{\pm 0.034}$	5.44$_{\pm 0.069}$	5.513$_{\pm 0.028}$	5.579$_{\pm 0.062}$	5.640
	Neutral	6.261$_{\pm 0.057}$	5.75$_{\pm 0.061}$	5.31$_{\pm 0.043}$	5.445$_{\pm 0.067}$	5.523$_{\pm 0.029}$	5.581$_{\pm 0.065}$	5.645
	Negative	6.26$_{\pm 0.059}$	5.722$_{\pm 0.06}$	5.188$_{\pm 0.019}$	5.382$_{\pm 0.072}$	5.467$_{\pm 0.046}$	5.556$_{\pm 0.055}$	5.596
SICK-E	Entailment	3.521$_{\pm 0.003}$	3.066$_{\pm 0.3}$	2.537$_{\pm 0.086}$	2.705$_{\pm 0.029}$	3.445$_{\pm 0.035}$	3.397$_{\pm 0.289}$	3.112
	Neutral	3.749$_{\pm 0.048}$	3.398$_{\pm 0.184}$	3.192$_{\pm 0.038}$	3.287$_{\pm 0.013}$	3.639$_{\pm 0.024}$	3.561$_{\pm 0.209}$	3.471
	Contradiction	2.181$_{\pm 0.027}$	2.212$_{\pm 0.061}$	2.155$_{\pm 0.029}$	2.252$_{\pm 0.017}$	2.212$_{\pm 0.016}$	2.172$_{\pm 0.038}$	2.197

Table 14: *Intra-clas* metrics of premise and hypothesis concatenated with different conjunctions on three NLI datasets. The results is averaged on three seeds.

Datasets	Class	*Maybe*	*And*	*Therefore*	*But*	*On the other hand*	*It is true that*	**Average**
MNLI	Positive	-0.033$_{\pm 0.019}$	-0.039$_{\pm 0.003}$	-0.103$_{\pm 0.004}$	-0.042$_{\pm 0.019}$	-0.027$_{\pm 0.002}$	-0.05$_{\pm 0.01}$	-0.049
	Neutral	-0.05$_{\pm 0.027}$	-0.052$_{\pm 0.003}$	-0.101$_{\pm 0.007}$	-0.067$_{\pm 0.031}$	-0.038$_{\pm 0.004}$	-0.065$_{\pm 0.027}$	-0.062
	Negative	-0.036$_{\pm 0.01}$	-0.063$_{\pm 0.006}$	-0.168$_{\pm 0.005}$	-0.048$_{\pm 0.013}$	-0.044$_{\pm 0.002}$	-0.058$_{\pm 0.003}$	-0.07
SNLI	Positive	-0.011$_{\pm 0.004}$	-0.035$_{\pm 0.005}$	-0.111$_{\pm 0.018}$	-0.068$_{\pm 0.016}$	-0.123$_{\pm 0.09}$	-0.064$_{\pm 0.009}$	-0.069
	Neutral	-0.012$_{\pm 0.004}$	-0.034$_{\pm 0.008}$	-0.092$_{\pm 0.015}$	-0.064$_{\pm 0.008}$	-0.122$_{\pm 0.084}$	-0.058$_{\pm 0.008}$	-0.064
	Negative	-0.013$_{\pm 0.005}$	-0.048$_{\pm 0.013}$	-0.164$_{\pm 0.034}$	-0.11$_{\pm 0.028}$	-0.203$_{\pm 0.188}$	-0.087$_{\pm 0.029}$	-0.104
SICK-E	Entailment	-1.372$_{\pm 0.199}$	-2.822$_{\pm 0.9}$	-2.992$_{\pm 0.316}$	-2.877$_{\pm 0.317}$	-1.335$_{\pm 0.161}$	-1.834$_{\pm 0.363}$	-2.205
	Neutral	-1.614$_{\pm 0.219}$	-2.811$_{\pm 0.897}$	-3.274$_{\pm 0.368}$	-3.027$_{\pm 0.277}$	-1.364$_{\pm 0.152}$	-2.089$_{\pm 0.385}$	-2.363
	Contradiction	-2.501$_{\pm 0.382}$	-5.259$_{\pm 1.564}$	-5.631$_{\pm 0.615}$	-5.397$_{\pm 0.498}$	-2.609$_{\pm 0.308}$	-3.411$_{\pm 0.711}$	-4.135

Table 15: *Inter-class* metrics of premise and hypothesis concatenated with different conjunctions on three NLI datasets. The results is averaged on three seeds.

D Generalization and Application of *GeoHard*

D.1 Theoretical generalization

Taking the data distribution as one Gaussian distribution $D \sim \mathcal{N}(\mu, \sigma^2)$, the mean of n instances sampled from D follows $\mathcal{N}(\mu, \sigma^2/n)$. Therefore, the means of the training data and the test data, $\hat{\mu}_{tr}$ and $\hat{\mu}_{te}$, follow $\mathcal{N}(\mu, \sigma^2/n_{tr})$ and $\mathcal{N}(\mu, \sigma^2/n_{te})$, where n_{tr} and n_{te} are the size of training and test sets, respectively. According to Chebyshev's inequality (Mitrinovic et al., 2013), the following inequalities stand with arbitrary $k \in R_+$:

$$P(|\hat{\mu}_{tr} - \mu| \geq \frac{k\sigma}{\sqrt{n_{tr}}}) \leq \frac{1}{k^2}$$
$$P(|\hat{\mu}_{te} - \mu| \geq \frac{k\sigma}{\sqrt{n_{te}}}) \leq \frac{1}{k^2}$$

Assuming $n_{tr} \geq n_{te}$ without loss of generality, we combine the two inequalities above and derive:

$$\begin{aligned}
\frac{2}{k^2} &\geq P(|\hat{\mu}_{tr} - \mu| \geq \frac{k\sigma}{\sqrt{n_{tr}}}) + P(|\hat{\mu}_{te} - \mu| \geq \frac{k\sigma}{\sqrt{n_{te}}}) \\
&\geq P(|\hat{\mu}_{tr} - \mu| \geq \frac{k\sigma}{\sqrt{n_{te}}}) + P(|\hat{\mu}_{te} - \mu| \geq \frac{k\sigma}{\sqrt{n_{te}}}) \\
&\geq P(|\hat{\mu}_{tr} - \mu| + |\hat{\mu}_{te} - \mu| \geq \frac{2k\sigma}{\sqrt{n_{te}}}) \\
&\geq P(|\hat{\mu}_{tr} - \hat{\mu}_{te}| \geq \frac{2k\sigma}{\sqrt{n_{te}}})
\end{aligned}$$

D.2 *Neutral*'s overfitting

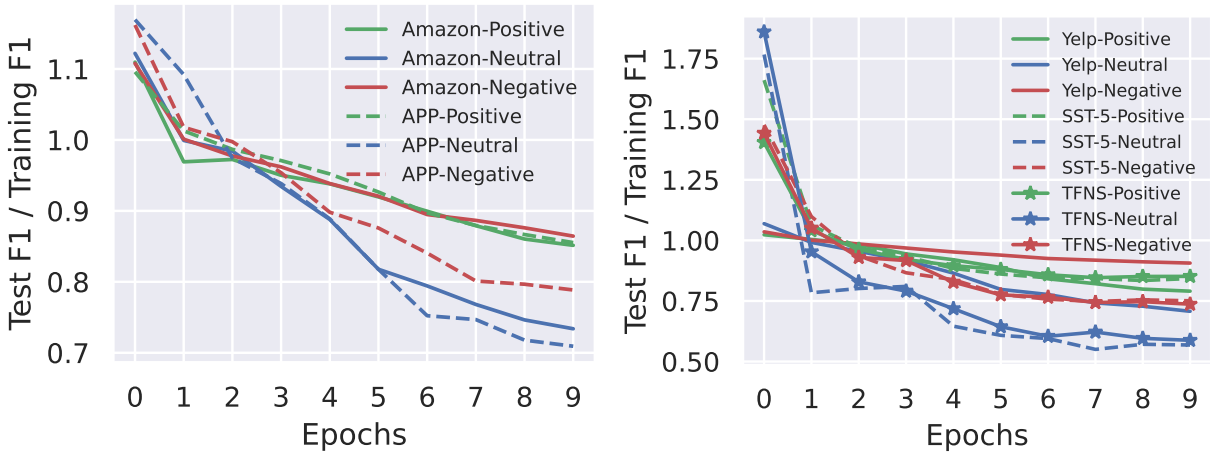


Figure 10: The ratio between F1 scores on the test and training sets for each training epoch on SC tasks (Left: Amazon and APP; Right: Yelp, SST-5 and TFNS).

D.3 Empirical Validation on *GeoHard*'s Generalization

As mentioned in the main part of the paper, we also include four datasets, i.e., AG News, Yahoo, Emo, and CARAR, from the tasks of emotion detection and topic classification. Similarly, we formulate the datasets to balance the number of instances inside each class to achieve class-wise balance. Specifically, we re-sample 10,000 instances from each class in Yahoo to handle the trade-off between computational efficiency and representativeness.

Trained on Roberta-Large with three seeds $\{1, 10, 100\}$, the performance of four datasets present the class-wise F1 scores in Table 16 - 19.

AG News	*World*	*Sports*	*Business*	*Sci/Tech*
F1 score (%)	96.4$_{\pm 0.08}$	99.1$_{\pm 0.07}$	92.7$_{\pm 0.16}$	93.2$_{\pm 0.14}$
Intra-class	4.464$_{\pm .016}$	3.835$_{\pm .034}$	3.775$_{\pm .042}$	3.978$_{\pm .070}$
Inter-class	-15.49$_{\pm .133}$	-17.62$_{\pm .100}$	-13.11$_{\pm .221}$	-13.35$_{\pm .075}$
GeoHard	-11.02$_{\pm .119}$	-13.79$_{\pm .084}$	-9.340$_{\pm .253}$	-9.377$_{\pm .018}$

Table 16: AG News's class-wise F1 scores trained with Roberta-Large and class-wise hardness measured by *GeoHard*.

Emo	*Others*	*Happy*	*Sad*	*Angry*
F1 score (%)	82.4$_{\pm 1.09}$	89.8$_{\pm 0.44}$	90.1$_{\pm 1.37}$	91.5$_{\pm 0.94}$
Intra-class	2.331$_{\pm .025}$	2.286$_{\pm .024}$	2.120$_{\pm .015}$	2.117$_{\pm .006}$
Inter-class	-6.841$_{\pm .081}$	-9.781$_{\pm .147}$	-9.521$_{\pm .186}$	-8.278$_{\pm .428}$
GeoHard	-4.509$_{\pm .063}$	-7.495$_{\pm .170}$	-7.400$_{\pm .177}$	-6.160$_{\pm .425}$

Table 17: Emo's class-wise F1 scores trained on Roberta-Large and class-wise hardness measured by *Geohard*.

CARAR	*Sadness*	*Joy*	*Love*	*Anger*	*Fear*	*Superise*
F1 score (%)	90.9$_{\pm 0.49}$	89.1$_{\pm 0.61}$	90.4$_{\pm 0.68}$	93.2$_{\pm 1.25}$	90.1$_{\pm 0.09}$	95.1$_{\pm 0.83}$
Intra-class	1.817$_{\pm .129}$	2.285$_{\pm .068}$	1.656$_{\pm .013}$	1.641$_{\pm .075}$	1.505$_{\pm .040}$	1.675$_{\pm .007}$
Inter-class	-6.538$_{\pm .144}$	-6.232$_{\pm .216}$	-5.806$_{\pm .079}$	-7.159$_{\pm .122}$	-6.043$_{\pm .123}$	-6.822$_{\pm .116}$
GeoHard	-4.720$_{\pm .263}$	-3.947$_{\pm .199}$	-4.150$_{\pm .066}$	-5.517$_{\pm .196}$	-4.537$_{\pm .107}$	-5.147$_{\pm .111}$

Table 18: CARAR's class-wise F1 scores trained on Roberta-Large and class-wise hardness measured by *GeoHard*.

Yahoo	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
F1 score (%)	64.8±0.72	77.7±0.28	82.0±0.19	59.7±0.50	87.8±0.64	91.9±0.12	59.6±0.17	76.8±0.30	78.8±0.04	81.5±0.23
Intra-class	3.08±.014	2.920±.009	2.726±.013	3.416±.014	2.550±.021	3.156±.004	4.257±.042	4.481±.066	3.195±.020	2.814±.027
Inter-class	-7.836±.033	-8.894±.012	-11.63±.092	-6.872±.006	-12.38±.137	-13.97±.053	-7.303±.053	-8.428±.042	-10.11±.062	-7.294±.066
GeoHard	-4.747±.019	-5.974±.003	-8.910±.080	-3.455±.015	-9.838±.117	-10.81±.056	-3.045±.012	-3.946±.025	-6.922±.043	-4.480±.042

Table 19: Yahoo's class-wise F1 scores trained with Roberta-Large and class-wise hardness measured by *GeoHard*. In Yahoo, the class index from 0-9 denotes the classes (*Society&Culture*, *Science&Mathematics*, *Health*, *Education&Reference*, *Computers&Internet*, *Sports*, *Business&Finance*, *Entertainment&Music*, *Family&Relationships*, *Politics&Government*)

Metric \ Dataset	SC					NLI			Macro Avg.
	Amazon	**APP**	**SST-5**	**TFNS**	**Yelp**	**MNLI**	**SNLI**	**SICK-E**	
PVI	0.985	0.9825	0.9808	1.0000	1.0000	0.9805	0.9628	0.9973	0.9861
Confidence	0.6966	0.9947	0.9959	0.997	0.7183	0.9721	0.9897	0.7478	0.8890
Variablity	-0.8755	-0.5244	-0.1256	0.8683	-0.7595	-0.9991	-0.9949	-0.0117	-0.4278

Table 20: Pearson's correlation coefficients between class-wise hardness measurement and class-wise F1 scores. The metrics include PVI and two metrics from training dynamics, i.e., confidence and variability.

D.4 Other metrics beyond training-free methods

Here, we include two training-based methods to further validate the existence of class-wise hardness. One is PVI (Ethayarajh et al., 2022) and the other is data cartography (Swayamdipta et al., 2020). PVI has been introduced in Appendix B.5. Data cartography focuses on the behavior of the model on data instances during training, referred to as training dynamics. This includes two metrics for each instance: the model's *confidence* in the correct class and the *variability* of this confidence across epochs. Data points characterized by high confidence and low variability are considered easy. In Table 20, we observe that PVI can well model the hardness of the classes. Moreover, we also notice the correlation between class-wise hardness and the metrics from training dynamics. These results further validates the existence of class-wise hardness through a training-based way.

D.5 *GeoHard*'s Application

The following demonstrates two examples of the demonstration applied in the ICL on Dynasent (Potts et al., 2021). Precisely, the upper and the lower demonstrations are 2P+1NEU+1N and 1P+2NEU+1N, respectively.

Sentence: This place is fine.i love this place, the staff is great the food is great and the atmosphere is great.
Sentiment: pos1
#####
Sentence: The casino has some of the lowest house-edge blackjack you will find anywhere.
Sentiment: positive
#####
Sentence: Too bad they only had one available spot that day, it was an appointment at 4:30pm, fortunately for me that is the least busiest time so I was in and out.
Sentiment: neutral
#####
Sentence: I went to the ticket counter. I wasn't going to the ticket counter after the show demanding a refund, but I certainly wouldn't go again.
Sentiment: negative
#####
Sentence: {input}
Sentiment:

Sentence: I tried a new place. I definitely recommend this place if you are looking for some good chinese food, and I definitely will be coming back.
Sentiment: positive
#####
Sentence: It was cool. It is set up like a lounge, but it has a dinky dancefloor, and music that is WAY TOO LOUD for a place that has a lounge setup.
Sentiment: mixed
#####
Sentence: So I'll give this one just one store.
Sentiment: neutral
#####
Sentence: I went to the ticket counter. I wasn't going to the ticket counter after the show demanding a refund, but I certainly wouldn't go again.
Sentiment: negative
#####
Sentence: {input}
Sentiment:

Artifacts/Packages	Citation	Link	License
		Artifacts(datasets/benchmarks).	
Amazon	(Keung et al., 2020)	https://huggingface.co/datasets/amazon_reviews_multi	LICENSE
APP	(Grano et al., 2017)	https://huggingface.co/datasets/app_reviews	Missing
MNLI	(Williams et al., 2018)	https://huggingface.co/datasets/multi_nli	MIT License
SICK-E	(Marelli et al., 2014)	https://huggingface.co/datasets/sick	CC-by-NC-SA-3.0
SNLI	(Bowman et al., 2015)	https://huggingface.co/datasets/snli	CC-by-4.0
SST-5	(Socher et al., 2013)	https://huggingface.co/datasets/SetFit/sst5	Missing
TFNS	N/A	https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment	MIT License
Yelp	(Zhang et al., 2015)	https://huggingface.co/datasets/yelp_review_full	LICENSE
Dynasent	(Potts et al., 2021)	https://github.com/cgpotts/dynasent	Apache License 2.0
		Packages	
PyTorch	(Paszke et al., 2019)	https://pytorch.org/	BSD-3 License
transformers	(Wolf et al., 2019)	https://huggingface.co/datasets/yelp_review_full	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
umap	(McInnes et al., 2018)	https://github.com/lmcinnes/umap	BSD 3-Clause License
		Models	
E5-Large-v2	(Wang et al., 2022)	https://huggingface.co/intfloat/e5-large-v2	MIT License
GTE-Large	(Li et al., 2023)	https://huggingface.co/thenlper/gte-large	MIT License
bge-large-en-v1.5	(Xiao et al., 2023)	https://huggingface.co/BAAI/bge-large-en-v1.5	MIT License
RoBERTa	(Liu et al., 2019)	https://huggingface.co/docs/transformers/model_doc/roberta	MIT License
Flan-T5	(Raffel et al., 2020)	https://huggingface.co/docs/transformers/model_doc/flan-t5	Apache-2.9
OPT	(Zhang et al., 2022)	https://huggingface.co/facebook/opt-2.7b	LICENSE
LLaMA-v2	(Touvron et al., 2023)	https://huggingface.co/docs/transformers/model_doc/llama2	LICENSE

Table 21: Details of datasets, major packages, and existing models we use. The datasets we reconstructed or revised and the code/software we provide are under the MIT License.

E Artifacts and Packages

The details of the datasets, major packages, and existing models are listed in Table 21.