

000 SATS : SCENARIO-ANCHORED TOPOLOGICAL SCOR- 001 002 ING IN FIGURATIVE EXPRESSION 003 004

005 **Anonymous authors**

006 Paper under double-blind review

007 008 ABSTRACT 009

010 Figurative expressions remain challenging for language models, which often de-
011 fault to literal interpretations rather than capturing implicit meaning. This vulne-
012 rability affects the understanding of everyday dialogue and increases the exposure
013 to adversarial prompts that exploit figurative or indirect phrasing. We integrate
014 a topology-based algorithm into encoder-only architectures to strengthen signals
015 relevant to figurative meaning and observe consistent improvements across mul-
016 tiple benchmarks. We further propose SATS, which achieves low latency and
017 matches or exceeds most open-source LLMs while using 9.6 \times fewer parameters
018 (within 0.8% p of Qwen3). Our approach is lightweight and model-agnostic, and
019 complements instruction-tuned LLMs by improving the robustness of detecting
020 and interpreting figurative and implicit meaning.

021 022 1 INTRODUCTION

023 Figurative expressions are linguistic devices that enrich human conversation by conveying implied
024 rather than surface meaning. They appear not only in poetry and fiction, but also in everyday con-
025 versation to evoke humor or convey reproach. For example, the sentence: “This young lady is
026 100% polyester” literally means that the woman is made of polyester, but its intended meaning is
027 metaphorical: she behaves in a way that is unnatural or not humanlike. However, large language
028 models (LLMs) often misinterpret such cases by taking surface cues like polyester literally. In
029 Figure 1, given one scenario and five options, most models select option 1 or option 3 in which
030 polyester appears verbatim. This behavior can be interpreted as a lexical-overlap bias with
031 the scenario text. Figurative expressions affect not only task accuracy but also the safety of LLM
032 services. Recent studies report that metaphorical expressions can be used to avoid hate speech de-
033 tection (Zeng et al., 2025) or to gradually steer models toward harmful behavior via covert framing,
034 commonly called jailbreak (Yan et al., 2025).

035 To mitigate these limitations, we propose a figurative expression detection algorithm that combines
036 an encoder-only model with a topological approach. Our method achieves competitive detection
037 performance with significantly fewer parameters than LLMs. It is designed to determine whether
038

041	Mark asked his mom what she thought about his 042 new girlfriend. She replied: “This young lady is 043 100% polyester.” 044 What does she mean?	045	His girlfriend wore clothes made of polyester. His girlfriend has a beautiful smile. His girlfriend is made of polyester. The girl made a good impression on Mark’s mom. His girlfriend’s behavior was not very natural.	
046		OPTION 1 (X) Llama-3.2-1B / Granite-3.1-1B / SmollLM2-360M / Phi-3-mini-3.8B	OPTION 3 (X) Llama-3-8B / Gemma-2B / Qwen2.5-1.5B	OPTION 5 (✓) Qwen3-4B / Mistral-7B

047 Figure 1: This example illustrates a Metaphor instance from the Pragmatics dataset (Hu et al., 2023).
048 All LLMs are instruction-tuned models, and the temperature was fixed at 0.0 to ensure determin-
049 istic outputs. All models were evaluated with the same prompt, whose details are provided in Ap-
050 pendix E.

054 a sentence is figurative and, when candidate explanations are provided, to select the most plausible
 055 literal explanation. Our contributions are as follows:
 056

- 057 • We introduce topological methods to representative figurative expression benchmarks, to the
 058 best of our knowledge for the first time, presenting a new paradigm beyond existing language-
 059 model approaches. On the Sarcasm benchmark, we observe an average improvement of about
 060 1% p over BERT-based models. In FLUTE, using Representation Topology Divergence (RTD),
 061 we obtain at least 4.8% p accuracy improvement over LLMs with fewer than 3B parameters.
- 062 • We propose a multi-option algorithm, called Scenario-Anchored Topological Scoring (SATS),
 063 which directly compares persistent homology scores, overcoming pairwise-only comparison and
 064 high latency of RTD. In FLUTE, the proposed method reduces latency by approximately 275×
 065 relative to RTD and improves accuracy by 1.9% p . Despite using about 9.6× fewer parameters,
 066 it achieves higher accuracy than all eight LLMs except Qwen3. Even relative to Qwen3, the
 067 performance gap remains small at 0.8% p .

068 2 RELATED WORK

069 2.1 FIGURATIVE EXPRESSION

070 Figurative expressions pose a significant challenge for language models, as they require understanding
 071 beyond literal meaning (Liu et al., 2022). To address this, several detection approaches have
 072 been explored. For instance, Yang et al. (2024) proposed a GPT-based algorithm for verb metaphor
 073 detection, while Li et al. (2023) identified metaphors by contrasting the literal and contextual
 074 meanings of a word. The need for detection has also been emphasized in safety research. Zeng et al.
 075 (2025) and Yan et al. (2025) showed that LLMs fail to block metaphorical implicit hate speech and
 076 are vulnerable to adversarial jailbreaks. Although these studies mainly focus on safety and social
 077 risks, our work is distinct in that it investigates the structural characteristics of figurative expressions
 078 themselves through mathematical methodologies.

079 To support such research, several datasets have been introduced. Representative examples include
 080 the VUA (Leong et al., 2020) and FLUTE (Chakrabarty et al., 2022) datasets, which provides
 081 instances of Sarcasm, Simile, Metaphor, and Idiom in NLI format generated by GPT and human eval-
 082 uators. In the sarcasm domain, SC V2 and MUStARD (Oraby et al., 2016; Castro et al., 2019) are
 083 widely used. Finally, Hu et al. (2023) investigates seven types of expressions, including Irony and
 084 Metaphor by requiring the selection of the most appropriate interpretation among multiple options.

085 2.2 TOPOLOGICAL DATA ANALYSIS

086 Early applications of persistent homology in NLP include Zhu (2013). Building on this direction,
 087 subsequent work constructs graphs from BERT attention and computes persistent homology fea-
 088 tures for spam detection and movie review sentiment classification (Perez & Reinauer, 2022). Other
 089 works reported meaningful performance improvements across benchmarks including Amazon Re-
 090 views, news classification, and grammatical acceptability by leveraging barcodes, mean of zero-
 091 persistent homology ($H_0 M$), and RTD rules derived from attention graphs (Kushnareva et al., 2021;
 092 Cherniavskii et al., 2022). In addition, Proskurina et al. (2023) demonstrated strong performance on
 093 structurally different languages such as English and Russian using chordality.

094 Beyond NLP, Topological Data Analysis (TDA) has also been explored in other domains. Persformer
 095 proposed the first Transformer architecture capable of directly processing persistence diagrams, ad-
 096 vancing representation learning and interpretability of TDA (Reinauer et al., 2022). Topoformer
 097 incorporated topographic organization into Transformer self-attention and demonstrated neurosci-
 098 entific alignment (Binhuraib et al., 2024). More recently, TDA has been shown to enhance knowledge
 099 transfer in distillation (Kim et al., 2024) and provides a promising paradigm for graph and relational
 100 learning (Papamarkou et al., 2024).

101 Although TDA has become a significant research topic across NLP and beyond, no prior work has
 102 directly examined figurative expressions. This study addresses this gap by introducing a TDA-based
 103 approach to understanding figurative expression for the first time.

108

3 METHODOLOGY

109
110 This section provides a background of the topological approach and introduces two methodologies.
111 First, it explains how existing methods are applied to figurative expressions and highlights their
112 limitations. Then, it presents the algorithm proposed in this study to address these limitations.
113114

3.1 BACKGROUND

115
116 In a Transformer, the attention matrix produced by each layer-head pair encodes pairwise token
117 affinities and can be viewed as a weighted directed graph, with tokens as nodes and attention scores
118 as edge weights. For an input of length m , let $W \in [0, 1]^{m \times m}$ denote this matrix. For $\tau \in [0, 1]$,
119 retain only edges whose weights are at least τ to obtain a graph G_τ . As τ increases, the family
120 $\{G_\tau\}$ forms a filtration, a nested family of graphs under increasingly strict thresholds. Topological
121 data analysis (TDA) summarizes how the topology of graph evolves along this filtration through
122 persistent homology, which computes homology across the filtration and records birth-death pairs
123 that mark the appearance and disappearance of features. These summaries are commonly visualized
124 as barcodes, collections of intervals that encode feature lifetimes.
125126 The Betti numbers vary with τ : β_0 counts connected components and β_1 counts cycles. Barcodes
127 record the birth and death of each component and cycle, where longer intervals indicate greater
128 structural stability. In H_0 , each component is born at $\tau = 0$ and dies at its merge time d_i , yielding
129 intervals $[0, d_i]$. Under a standard similarity-to-distance transform and a decreasing-threshold filtration,
130 the sum of H_0 barcode lengths is linearly related to the weight of a minimum spanning tree
(Carlsson & Vejdemo-Johansson, 2022).
131132

3.2 CLASSIFICATION WITH TOPOLOGICAL FEATURES

133 To validate a topological approach to figurative expressions, this study extracts attention matrices
134 from BERT-based models and derives TDA vectors from them. These vectors are combined with the
135 [CLS] token and used as input to a linear classifier for training and prediction. This approach is based
136 on Kushnareva et al. (2021), and an example in our dataset is shown in Figure 2 in Appendix A.
137138 The TDA vectors consist of three groups of features. Topological features include β_0 values, edge
139 counts, and cycle counts across thresholds concatenated into a single vector. Barcode features in-
140 clude the sum and average lengths of the barcodes for H_0 and H_1 , the number of intervals corre-
141 sponding to the birth or death thresholds, and barcodes entropy. Finally, distance-to-patterns mea-
142 sures the normalized distance to predefined pattern graphs, reflecting how attention is distributed
143 across specific tokens.
144145

3.3 REPRESENTATION TOPOLOGY DIVERGENCE

146 Representation Topology Divergence (RTD) was first proposed in (Barannikov et al., 2022) as a
147 metric that quantifies multi-scale topological differences between two point clouds obtained from
148 the same sample set. In the NLP domain, Cherniavskii et al. (2022) converted sentence attention
149 matrices into weighted graphs and applied RTD. In this formulation, RTD is defined according to
150 the appearance of edges across thresholds: an edge that appears in only one graph is marked as birth,
151 and an edge that appears in both graphs is marked as death. This process yields barcodes, with RTD
152 computed as the sum of their lengths. Formally, for each bar α with birth and death interval $[\alpha_i, \alpha_j]$
153 in the barcode, its length is $\alpha_j - \alpha_i$. Thus, RTD is defined as
154

155
$$\text{RTD}(G_a, G_b) := \sum_{\alpha} (\alpha_j - \alpha_i),$$

156 where G_a and G_b are attention graphs. An example of RTD is detailed in Appendix B.
157158 We adapt RTD under the hypothesis that figurative expressions are strongly associated with their
159 correct interpretations. Given a figurative sentence with a literal explanation pair, we construct two
160 inputs by concatenating the sentence with each explanation. We denote the sentence paired with
161 the entailing explanation as E and the sentence paired with the contradicting explanation as C . We
then compute $\text{RTD}(G_E, G_C)$ and $\text{RTD}(G_C, G_E)$. Under our hypothesis, the correct explanation

162 yields more stable barcodes. Therefore, when $\text{RTD}(A, B) < \text{RTD}(B, A)$, we define the combination
 163 with A as correct. To improve RTD performance, we employ a head ensemble strategy that,
 164 without additional parameter tuning, selects and aggregates layer-head combinations according to
 165 their validation performance on auxiliary data. Algorithmic details and procedures are provided in
 166 Appendix C.

167 Despite its utility, RTD has clear limitations. First, it is inherently restricted to pairwise comparisons
 168 and does not generalize to more than two candidates. Second, the computational complexity of RTD
 169 is difficult to analyze explicitly and has been empirically observed to incur a higher cost than $H_0 M$
 170 (Cherniavskii et al., 2022). Similar concerns are noted in more recent work (Tulchinskii et al., 2025),
 171 which further motivates the need for scalable alternatives. To address these limitations, we propose
 172 a novel multi-hop approach in Section 3.4.

174 3.4 SCENARIO-ANCHORED TOPOLOGICAL SCORING (SATS)

175 Where appropriate, our exposition draws on Carlsson & Vejdemo-Johansson (2022) for definitions
 176 and notation. Because persistent homology is defined on simplicial complexes, we lift a graph to
 177 a simplicial complex. Let the thresholded graph be $G_\tau = (V, E_\tau)$ with the superlevel definition
 178 $E_\tau = \{\{a, b\} \in E : w_{ab} \geq \tau\}$, and let its clique (flag) complex be $K_\tau := \text{Cl}(G_\tau)$. Equipping
 179 the set of thresholds $I \subseteq [0, 1]$ with a descending total order \succeq , we obtain a filtration since $\tau_1 \succeq \tau_2$
 180 implies $K_{\tau_1} \subseteq K_{\tau_2}$.

181 For each instance, we concatenate the scenario S with each option O_i to obtain a token sequence of
 182 length T (the length may depend on i , but we keep the notation T). Let $A_{\ell, h}^{(i)} \in \mathbb{R}_{\geq 0}^{T \times T}$ denote the
 183 attention matrix of layer ℓ and head h , and mask special tokens in advance. For notational simplicity,
 184 we subsequently omit (i, ℓ, h) . To remove asymmetry, set

$$186 \quad U = \frac{1}{2} (A + A^\top), \quad \text{diag}(U) = 0$$

187 and self-loops are removed. Because the dataset provides multiple options per scenario, we concatenate
 188 the scenario for each option. To focus on the scenario–option relation, we identify all scenario
 189 vertices into a single node $[S]$ through an equivalence relation \sim and track connectivity on K_τ / \sim .

190 Since we are concerned with connectivity only, we use H_0 . Because H_0 is determined by the 1-
 191 skeleton, K_τ and G_τ are equivalent with respect to the connected components, and $H_0(K_\tau) \cong$
 192 $H_0(G_\tau)$ is true. Consequently, in practice, we compute on the graph G_τ without explicitly con-
 193 structing the complex.

194 Define the death of an option vertex o by

$$197 \quad d(o) = \sup\{\tau \in I : o \text{ and } [S] \text{ are still disconnected in } K_\tau / \sim\},$$

198 which is attained as a maximum in a finite graph. Define the widest path by

$$200 \quad \text{wp}(u, v) := \max_{p: u \rightsquigarrow v} \min_{e \in p} w_e, \quad \text{wp}([S], o) := \max_{s \in S} p(s, o).$$

201 Then the following holds.

$$203 \quad \forall o \in O : d(o) = \text{wp}([S], o) = \max_{s \in S} \max_{p: s \rightsquigarrow o} \min_{e \in p} w_e.$$

205 Instead of directly evaluating the max–min rule of threshold scanning, we use the finite Katz index
 206 (Katz, 1953), which discounts and accumulates contributions of indirect paths:

$$208 \quad \Phi_K(U; \beta) = \sum_{t=1}^K \beta^{t-1} U^t, \quad \beta \in (0, 1), K \in \mathbb{N}.$$

211 We blend the multi-hop term with the direct term via

$$212 \quad U_{\text{blend}} = \lambda U + (1 - \lambda) \Phi_K(U; \beta), \quad \lambda \in [0, 1].$$

214 The scenario–option affinity is

$$215 \quad u(o) := \max_{s \in S} (U_{\text{blend}})_{s, o}, \quad o \in O_i,$$

216 and the SATS score of option i is

$$217 \quad \text{SATS}(i) = \frac{1}{|O_i|} \sum_{o \in O_i} -\log(u(o) + \varepsilon), \quad \varepsilon = 10^{-6}.$$

220 The final prediction is $\arg \min_i \text{SATS}(i)$, and the aggregation over scenario tokens used to compute
 221 $u(o)$ can be chosen as one of max, top- k , or softmax. The default hyperparameters are $K = 3$,
 222 $\beta = 0.6$, and $\lambda = 0.6$. We compute $u(o)$ using top- k aggregation with $k = 3$. Formal proofs
 223 of the equivalence between death and the widest path and the stability of SATS are provided in
 224 Appendix D.

225 Each step costs $O(T^3)$ on a $T \times T$ matrix U over K steps, the total complexity is $O(K T^3)$. How-
 226 ever, because our implementation executes all operations as dense matrix multiplications on the
 227 GPU via PyTorch, the measured latency was observed to be lower than that of RTD. The latency is
 228 primarily determined by K and the token length T , while the hyperparameters β , λ , and the choice
 229 of aggregation (top- k /max/softmax) do not change the leading-order term.

230 4 EXPERIMENT

231 4.1 DATASETS

232 Sarcasm Corpus V2 (SC V2) (Oraby et al., 2016) is a sentence-level sarcasm dataset with three
 233 types: GEN (generic situations), HYP (hyperbolic expressions), and RQ (rhetorical questions). In
 234 our experiments, we split each type into training (70%) and test (30%) sets. FLUTE (Chakrabarty
 235 et al., 2022) is an NLI-format figurative expression dataset constructed using GPT generation, expert
 236 annotation, and crowdsourced labeling. It comprises four categories: Sarcasm, Simile, Metaphor,
 237 and Idiom. In this study, we exclude Sarcasm and use the remaining three categories since the struc-
 238 ture of the Sarcasm differs from the other categories. Because each item contains one hypothesis and
 239 two premises, we treat each hypothesis–premise pair as a single case. Consequently, the effective
 240 number of instances is half of the original. Pragmatics dataset (Hu et al., 2023) requires selecting the
 241 most appropriate sentence from multiple candidates for a given scenario. Each type contains 20–40
 242 instances, and no training split is provided. We use the Irony and Metaphor types, and each scenario
 243 includes four or five candidate options. The overall composition of these datasets is summarized in
 244 Table 1.

245 4.2 MODELS

246 SC V2 employs BERT-based models, including bert-base-uncased (Devlin et al., 2018), roberta
 247 (base, large) (Liu et al., 2019a), and deberta-v3 (base, large) (He et al., 2021). The comparison
 248 models are fine-tuned on the training data, and in the TDA experiments, both the original and the
 249 fine-tuned models are compared.

250 In the FLUTE and Pragmatics experiments, we use models with fewer than 3B parameters and those
 251 with 3B or more. The small models include Granite-3.1-1B (Granite Team, 2024), Llama-3.2-1B
 252 (AI@Meta, 2024), SmolLM2-360M (Allal et al., 2025), Qwen2.5-1.5B (Team, 2024), and Gemma-
 253 2B. The large models include Mistral-7B-V0.3 (Jiang et al., 2023), Llama-3-8B, Qwen3-4B (Team,
 254 2025), and Phi-3-mini-4k (Abdin et al., 2024). The FLUTE experiment additionally includes T5-
 255 Large (Roberts et al., 2022) and T5-FLUTE (Chakrabarty et al., 2022). All generative models, unless
 256 otherwise noted, are instruction-tuned versions.

257 For the RTD and SATS experiments, we use DeBERTa-v3-large, BART-large-MNLI (Lewis et al.,
 258 2020) and RoBERTa-large-MNLI (Liu et al., 2019b).

259 4.3 SETUP

260 We run all experiments on four RTX 5090 GPUs (32 GB each). All evaluations are zero-
 261 shot on the test set except SC V2. All LLMs share the same prompt and decoding settings
 262 (temperature=0.0, top_p=1.0, max_new_tokens=2). Because we request only a numeric
 263 prediction from each LLM (a single integer), we cap max_new_tokens at 2. The full prompt, the
 264 negative-lexicon specification used to suppress tokens (e.g., “answer,” “correct,” “option”), and the
 265 constraint that forces generation of an option index are provided in Appendix E.

270

271 Table 1: Training and test set composition. Each cell reports the original count followed by the ef-
272 fectionate count (original / effective), values in parentheses indicate the number of options per scenario.

	SC V2			FLUTE			Pragmatics		
	GEN	HYP	RQ	Simile	Metaphor	Idiom	Irony	Metaphor	
Train	4564	814	1191	1250 / 625	1250 / 625	1768 / 884	-	-	
Test	1956	350	511	250 / 125	248 / 124	250 / 125	25 (4)	20 (5)	

273

274

275 Table 2: SC V2 results. Each dataset reports Accuracy and F1. Left columns show the Baseline
276 (fine-tuned), and right columns show Topological Features (TDA, fine-tuned). All models used the
277 same training settings batch size : 32 and epochs : 5. The best performance is highlighted in **bold**.

Model	Overall				GEN				HYP				RQ			
	Baseline		TDA		Baseline		TDA		Baseline		TDA		Baseline		TDA	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
bert-base-uncased	75.9	75.0	72.3	72.3	78.4	77.3	77.0	77.0	70.9	70.0	66.0	66.0	78.3	77.8	74.0	74.0
RoBERTa-base	79.2	79.3	78.1	78.7	83.9	83.6	80.4	80.9	73.4	74.4	73.1	74.0	80.4	79.8	80.8	81.3
RoBERTa-large	69.4	75.4	78.6	78.4	50.1	66.8	77.1	77.2	76.6	78.2	76.9	77.1	81.4	81.1	81.8	81.0
DeBERTa-v3-base	79.6	78.9	78.9	78.7	82.2	81.6	81.0	81.0	76.3	75.8	74.3	74.0	80.4	79.2	81.4	81.1
DeBERTa-v3-large	80.1	79.6	81.1	80.6	83.0	82.8	83.3	83.2	75.1	74.5	77.7	76.9	82.2	81.4	82.4	81.8

278

279

280 4.4 RESULTS

281

282 Table 2 reports the SC V2 results. Under identical training settings, models fine-tuned with TDA-
283 based representations (right columns) achieved higher overall accuracy and F1 than models fine-
284 tuned with standard methods only (left columns). Notably, DeBERTa-v3-large achieved the best
285 overall performance, and consistent gains were also observed for RoBERTa-large. These observa-
286 tions suggest that cues related to sarcasm may already be encoded in attention, and that TDA-based
287 representations expose this signal during training, yielding performance gains.

288

289 However, for base-scale models we observed degradations. A plausible explanation is that, com-
290 pared to large models, they contain fewer layers and heads, which makes it harder to capture stable
291 topological features. Although these models underperform relative to baselines counterparts, Ta-
292 ble 6 in Appendix F shows that, even with frozen parameters, the topology-based score alone ex-
293 ceeds the chance level of 50%. Overall, the results indicate that language models internalize aspects
294 of sarcasm and that topological methods can efficiently extract these signals and improve detection
295 performance.

296

297 Table 3 summarizes the FLUTE results. With the exception of Qwen2.5, LLMs with at most 3B pa-
298 rameters perform only modestly above chance, while models with at least 3B parameters generally
299 reach 85% or higher, with Qwen3 leading. The topological baseline (RTD) peaks on the Simile sub-
300 set when a head ensemble is used. It is competitive and even surpasses Mistral v0.3, but its overall
301 accuracy remains below that of several LLMs, including Qwen3. Under SATS, both DeBERTa and
302 RoBERTa obtain the best accuracy on Idiom. DeBERTa-v3-large also attains the highest accuracy
303 on Simile and ranks second overall. Relative to Qwen3, it uses about $9.6 \times$ fewer parameters while
304 trailing by only 0.8% p , and it exceeds all other large LLMs. As shown in Table 4, SATS substan-
305 tially outperforms RTD in speed: RTD is costly due to full-range threshold scanning, whereas SATS
306 executes $O(K T^3)$ operations in parallel on GPUs using PyTorch, producing low latency. In sum-
307 mary, SATS delivers an accuracy comparable to or exceeding that of almost 10 times larger LLMs,
308 while requiring far fewer resources and achieving lower latency.

309

310 We observe the same pattern on the Pragmatics dataset. As with FLUTE, all LLMs are evaluated
311 under identical prompts and settings (details in Appendix E). Pragmatics comprises 25 four-choice
312 questions for Irony (chance level 25%) and 20 five-choice questions for Metaphor (chance level
313 20%). Because the two subsets differ in both the number of questions and the number of options,
314 the overall score is reported as a micro-average. Under micro-averaging, the overall chance level is
315 22.8%. Table 5 shows that most small LLMs operate at or below chance: Gemma is only marginally
316 above chance on all tasks, and aside from Qwen2.5 other small models fall below chance. Qualitative
317 analysis indicates a lexical-overlap bias: Low-performing models often choose options that reuse
318 the salient words of the scenario rather than capturing scenario–option semantics. For instance, in
319 the error cases of Figure 1, Options 1 and 3 repeat the token `polyester` and are selected most

320

321

322

323

324
 325 Table 3: Experimental results (accuracy) for Simile, Metaphor, and Idiom in the FLUTE dataset.
 326 Italic text denotes experimental results cited from Chakrabarty et al. (2022). For each benchmark,
 327 the best performance is shown in **bold**, and the second-best performance is indicated by underlining.

Model	Params	Overall	Simile	Metaphor	Idiom
T5 baselines					
flan-T5-large	780M	27.8	6.4	35.5	41.6
<i>T5-flute</i>	780M	<u>69.7</u>	<u>62.8</u>	73.3	72.9
LLMs \leq 3B params					
Llama3.2-1B-Instruct	1B	51.4	44.0	57.3	52.8
SmolLM2-360M-Instruct	360M	55.1	54.4	55.7	55.2
Gemma-2B-Instruct	2B	56.7	43.2	72.6	54.4
Granite-3.1-1B-Instruct	1B	60.1	48.4	71.8	60.0
Qwen2.5-1.5B-Instruct	1.5B	83.7	70.4	89.5	91.2
LLMs \geq 3B params					
Mistral-7B-Instruct-v0.3	7B	85.6	76.8	90.3	89.6
Llama3-8B-Instruct	8B	89.1	78.4	94.4	94.4
Phi3-mini-3.8B-Instruct	3.8B	89.1	78.4	96.0	92.8
Qwen3-4B-Instruct	4B	91.2	81.6	<u>95.2</u>	96.8
MNLI-pre-trained baselines					
RoBERTa-large-MNLI	356M	23.2	16.8	18.5	34.4
BART-large-MNLI	407M	58.9	45.6	69.4	61.6
RTD Approach					
BART-large-MNLI	407M	64.2	62.4	69.4	60.8
BART-large-MNLI Head Ensemble	407M	76.7	74.4	72.6	81.6
RoBERTa-large-MNLI	356M	76.1	67.2	83.1	84.3
RoBERTa-large-MNLI Head Ensemble	356M	88.5	81.6	91.1	92.8
SATS Approach					
DeBERTa-v3-large	418M	<u>90.4</u>	81.6	91.9	97.6
RoBERTa-large-MNLI	356M	89.8	79.2	92.7	97.6

353
 354 Table 4: FLUTE and Pragmatics latency in seconds by model group.
 355

	FLUTE	Pragmatics
LLMs \leq 3B params		
Llama3.2-1B-Instruct	2.71	0.78
SmolLM2-360M-Instruct	3.33	0.63
Granite-3.1-1B-Instruct	8.19	1.53
Gemma-2B-Instruct	2.87	0.57
Qwen2.5-1.5B-Instruct	3.07	0.70
LLMs \geq 3B params		
Mistral-7B-Instruct-v0.3	5.59	1.30
Llama3-8B-Instruct	6.34	1.42
Phi3-mini-3.8B-Instruct	3.71	0.91
Qwen3-4B-Instruct	4.72	1.04
RTD Method		
BART-large-MNLI	596.12	–
RoBERTa-large-MNLI	1479.15	–
SATS Method		
DeBERTa-v3-large	3.71	2.91
RoBERTa-large-MNLI	5.38	3.31

375 frequently. Among LLMs, Qwen3 achieved the highest performance, consistent with the FLUTE
 376 results. Under SATS, RoBERTa-large-MNLI achieved the best accuracy on Irony, and DeBERTa-v3-
 377 large ranked second on both tasks. In the Overall (micro) metric, RoBERTa-large-MNLI performed
 at a level similar to Qwen3, followed by DeBERTa-v3-large. Because the confidence intervals for

378
 379 Table 5: This table reports accuracy on the Pragmatics dataset (Irony and Metaphor). Because the
 380 dataset is small and the numbers of test cases and options differ across subsets, the overall accuracy
 381 is computed via micro-averaging. We also report 95% confidence intervals. For each benchmark,
 382 the best result is shown in **bold**, and the second-best result is underlined.

Model	Overall (micro)	Irony	Metaphor
LLMs \leq 3B params			
Llama3.2-1B-Instruct	8.9 [2.5, 21.2]	12.0 [2.5, 31.2]	5.0 [0.1, 24.9]
SmolLM2-360M-Instruct	8.9 [2.5, 21.2]	12.0 [2.5, 31.2]	5.0 [0.1, 24.9]
Granite-3.1-1B-Instruct	15.6 [6.5, 29.5]	24.0 [9.4, 45.1]	5.0 [0.1, 24.9]
Gemma-2B-Instruct	26.7 [14.6, 41.9]	28.0 [12.1, 49.4]	25.0 [8.7, 49.1]
Qwen2.5-1.5B-Instruct	53.3 [37.9, 68.3]	60.0 [38.7, 78.9]	45.0 [23.1, 68.5]
LLMs \geq 3B params			
Mistral-7B-Instruct-v0.3	57.8 [42.2, 72.3]	52.0 [31.3, 72.2]	<u>65.0</u> [40.8, 84.6]
Llama3-8B-Instruct	60.0 [44.3, 74.3]	64.0 [42.5, 82.0]	55.0 [31.5, 76.9]
Phi3-mini-3.8B-Instruct	64.4 [48.8, 78.1]	64.0 [42.5, 82.0]	<u>65.0</u> [40.8, 84.6]
Qwen3-4B-Instruct	68.9 [53.4, 81.8]	64.0 [42.5, 82.0]	75.0 [50.9, 91.3]
SATS Method			
DeBERTa-v3-large	<u>66.7</u> [51.0, 80.0]	68.0 [46.5, 85.1]	65.0 [40.8, 84.6]
RoBERTa-large-MNLI	68.9 [53.4, 81.8]	80.0 [59.3, 93.2]	55.0 [31.5, 76.9]

399 large LLMs (including Qwen2.5) overlap substantially, we interpret SATS to provide performance
 400 comparable to that of LLMs. According to Table 4, taking into account the parameter scale and
 401 latency, SATS offers an efficient solution that reaches LLM-level accuracy with lower latency and
 402 much fewer resources, even in multiple choice settings.

5 DISCUSSION

5.1 ANALYSIS OF SELECTED ATTENTION

409 Figure 5 summarizes the distribution of high-accuracy layer-head pairs in RoBERTa-large-MNLI
 410 and DeBERTa-v3-large under SATS. The color legend is as follows: yellow marks the best pair on
 411 FLUTE, green marks the best pair on Pragmatics, and orange marks the top 3 (duplicates allowed)
 412 layer-head pairs per dataset. For visualization, pairs are drawn under the same test conditions in
 413 the experiments. The procedure uses fixed tokenizers and models without stochastic components,
 414 so variability due to random seeds is negligible. For RoBERTa, aside from (1, 11), (3, 15), high-
 415 performing pairs tend to cluster in the mid-to-late layers. DeBERTa shows a similar tendency,
 416 with some early layers also appearing among the top-3. Examining the layer-head distribution
 417 selected by the RTD head ensemble, BART exhibits a relatively even spread across layers, whereas
 418 RoBERTa, while including some early layers, concentrates more frequently in mid-to-late layers.
 419 These observations align with previous findings that semantic information is more prominent in
 420 later layers of RoBERTa models (Li et al., 2021). Details on the selected attention heads in RTD are
 421 reported in Appendix G.

5.2 OPERATION OF SATS METHOD

424 Figure 7 (A) visualizes how SATS selects the correct answer on the FLUTE Simile dataset for
 425 DeBERTa-v3-large. Each token is obtained by encoding the input with the tokenizer of the model.
 426 Special tokens were removed in advance, whereas punctuation marks, including apostrophes, were
 427 retained. Here, $\tau \in [0, 1]$ denotes the threshold on edge weights in the multi-hop Φ . As τ decreases
 428 from 1 to 0, higher-weight edges are activated in descending order, and newly activated edges are
 429 highlighted in red. When connections appear at relatively large values of τ , the $u(o)$ of the corre-
 430 sponding scenario-option pair tends to increase. Consequently, the SATS value decreases, making
 431 the option more likely to be selected as correct. In this example, Option 2 acquires edges earliest
 432 and ultimately attains the smallest SATS value.

432 The bar chart on the right displays $u(o)$ for each option token and larger values indicate stronger
 433 associations with the scenario tokens. For Option 1, s_1 and o_1 , both `gesture`, `connect first`, but
 434 the $u(o)$ for the core token `awkward` is the lowest, suggesting that the meaning of the option does
 435 not align with the scenario. By contrast, in Option 2, s_5 (`ballet`) and o_3 (`elegant`) connect first,
 436 supporting that the implied meaning of the scenario, `ballet` like movements are `elegant`, is
 437 captured. Similarly to Option 1, the s_1 and o_1 , `gesture`, `link forms` at the next threshold. Despite
 438 sharing the same lexical form, it emerges later than the (`ballet`, `elegant`) pair, indicating that
 439 SATS prioritizes contextually coherent connection patterns over simple lexical overlap. (B) and
 440 additional examples are provided in Appendix I.

441 6 CONCLUSION

442 This study shows that a topological approach to figurative expressions can effectively capture
 443 metaphorical meaning within language models. In particular, the proposed SATS extends naturally
 444 from binary labels to multiple choice settings, achieving higher accuracy and lower latency than
 445 RTD. SATS captures contextually coherent semantic connections between scenarios and options
 446 without relying on simple lexical overlap, and it strengthens scenario–option coupling via multi-hop
 447 propagation. Furthermore, SATS, implemented with an encoder-based model that uses far fewer pa-
 448 rameters than large LLMs, achieves accuracy comparable to or exceeding that of LLMs on multiple
 449 benchmarks. Given its reduced resource requirements, the SATS pipeline can serve as a practi-
 450 cal component for integration with instruction-tuned LLMs to assist in detecting and understanding
 451 figurative expressions.

452 7 LIMITATIONS AND FUTURE WORK

453 The SATS approach has several limitations. Although parameter-efficient, its accuracy remains
 454 approximately 2 percentage points lower than that of Qwen3. Because our experiments adopt a
 455 classification setting with an encoder-only model, the method does not produce descriptive outputs
 456 when no literal interpretation is provided. In addition, the available evaluation sets are also small,
 457 FLUTE includes more than 100 test cases and thus supports some degree of generalization, whereas
 458 Fine-Grained Pragmatics provides only 20–25 cases per type, so the resulting estimates carry high
 459 variance.

460 Future research directions to address these limitations are as follows. First, we aim to narrow the
 461 performance gap of SATS by fine-tuning the base encoder to specialize in figurative expressions.
 462 On SC V2, fine-tuning strengthened the topological features and improved accuracy, which pro-
 463 vides empirical support for this direction. As ongoing work, we are investigating the integration of
 464 SATS upstream of instruction-tuned LLMs as a lightweight preprocessing module. When figurative
 465 language is detected, the LLM generates multiple literal-interpretation candidates. SATS then se-
 466 lects the most plausible candidate using its multi-option scoring and supplies it as auxiliary input
 467 to the base model and to the policy filter. A comprehensive evaluation of safety is left to future
 468 work. Second, we plan to broaden benchmarks for the evaluation of generalization. For datasets
 469 without labels of literal interpretation (e.g., HYPO, LCC, TroFi Badathala et al. (2023)), we will
 470 generate plausible candidates of literal interpretation via LLM-based data augmentation and expand
 471 the evaluation set through human verification, enabling more precise measurement across diverse
 472 domains.

473 **LLM Usage** We used large language models only as assistive tools for (i) polishing and stylistic
 474 refinement of the manuscript, and (ii) surfacing candidate papers during literature search. All
 475 substantive content (problem formulation, methods, experiments, analyses, and conclusions) was
 476 written by the authors. Any text or citations suggested by an LLM were reviewed and verified by the
 477 authors, and we cite only sources we personally inspected. No non-public data or reviewer materials
 478 were uploaded to third-party services, and we preserved double-blind anonymity.

479 REFERENCES

480 Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany
 481 Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha

- 486 Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu
 487 Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ro-
 488 nen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng
 489 Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karam-
 490 patziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi
 491 Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi,
 492 Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant,
 493 Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi,
 494 Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia
 495 Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt,
 496 Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang,
 497 Cyril Zhang, Jianwen Zhang, Li Lyra Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren
 498 Zhou. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*,
 499 abs/2404.14219, 2024. URL <https://doi.org/10.48550/arXiv.2404.14219>.
 500
 501 AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- 502 Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo,
 503 Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav,
 504 Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo
 505 Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and
 506 Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model,
 507 2025. URL <https://arxiv.org/abs/2502.02737>.
- 508 Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya.
 509 A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In
 510 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association
 511 for Computational Linguistics: ACL 2023*, pp. 388–401, Toronto, Canada, July 2023. Association
 512 for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.26. URL <https://aclanthology.org/2023.findings-acl.26/>.
 513
- 514 Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology
 515 divergence: A method for comparing neural network representations. In Kamalika Chaudhuri,
 516 Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of
 517 the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine
 518 Learning Research*, pp. 1607–1626. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/barannikov22a.html>.
 519
- 520 Taha Osama A Binhuraib, Greta Tuckute, and Nicholas Blauch. Topoformer: brain-like topographic
 521 organization in transformer language models through spatial querying and reweighting. In *ICLR
 522 2024 Workshop on Representational Alignment*, 2024. URL <https://openreview.net/forum?id=3pLMzgoZSA>.
 523
- 524 Gunnar Carlsson and Mikael Vejdemo-Johansson. *Topological Data Analysis with Applications*.
 525 Cambridge University Press, 2022. ISBN 978-1108838658.
- 526 Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea,
 527 and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In
 528 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume
 529 1: Long Papers)*, Florence, Italy, 7 2019. Association for Computational Linguistics.
 530
- 531 Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figura-
 532 tive language understanding through textual explanations. In Yoav Goldberg, Zornitsa Kozareva,
 533 and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natu-
 534 ral Language Processing*, pp. 7139–7159, Abu Dhabi, United Arab Emirates, December 2022.
 535 Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.481. URL
 536 <https://aclanthology.org/2022.emnlp-main.481/>.
 537
- 538 Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva,
 539 Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontovski, and Evgeny
 540 Burnaev. Acceptability judgements via examining the topology of attention maps. In Yoav

- 540 Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 88–107, Abu Dhabi, United Arab Emirates, December 541 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.7. URL 542 <https://aclanthology.org/2022.findings-emnlp.7/>.
- 543
- 544 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep 545 bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL 546 <http://arxiv.org/abs/1810.04805>.
- 547
- 548 IBM Granite Team. Granite 3.0 language models, 2024.
- 549
- 550 Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style 551 pre-training with gradient-disentangled embedding sharing, 2021.
- 552
- 553 Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine- 554 grained comparison of pragmatic language understanding in humans and language models. In 555 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual 556 Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194– 557 4213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/ 558 2023.acl-long.230. URL <https://aclanthology.org/2023.acl-long.230/>.
- 559 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chap- 560 lot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, 561 Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, 562 Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- 563
- 564 Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 565 1953. doi: 10.1007/BF02289026.
- 566
- 567 Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological char- 568 acteristics help in knowledge distillation? In *Forty-first International Conference on Machine 569 Learning*, 2024. URL <https://openreview.net/forum?id=2dEH0u8w0b>.
- 570
- 571 Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Baran- 572 nikov, Alexander Bernstein, Irina Piontovskaya, Dmitri Piontovski, and Evgeny Burnaev. 573 Artificial text detection via examining the topology of attention maps. In Marie-Francine 574 Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 575 Conference on Empirical Methods in Natural Language Processing*, pp. 635–649, Online and 576 Punta Cana, Dominican Republic, November 2021. Association for Computational Linguis- 577 tics. doi: 10.18653/v1/2021.emnlp-main.50. URL <https://aclanthology.org/2021.emnlp-main.50/>.
- 578
- 579 Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and 580 Xianyang Chen. A report on the 2020 VUA and TOEFL metaphor detection shared task. In 581 Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee 582 Wee, Anna Feldman, and Debanjan Ghosh (eds.), *Proceedings of the Second Workshop on Figu- 583 rative Language Processing*, pp. 18–29, Online, July 2020. Association for Computational Lin- 584 guistics. doi: 10.18653/v1/2020.figlang-1.3. URL <https://aclanthology.org/2020.figlang-1.3/>.
- 585
- 586 Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer 587 Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre- 588 training for natural language generation, translation, and comprehension. In Dan Jurafsky, 589 Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meet- 590 ing of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. As- 591 sociation for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL <https://aclanthology.org/2020.acl-main.703/>.
- 592
- 593 Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. How is BERT surprised? 594 layerwise detection of linguistic anomalies. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto

- 594 Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Lin-*
 595 *guistics and the 11th International Joint Conference on Natural Language Processing (Volume*
 596 *1: Long Papers)*, pp. 4215–4228, Online, August 2021. Association for Computational Lin-
 597 *guistics.* doi: 10.18653/v1/2021.acl-long.325. URL <https://aclanthology.org/2021.acl-long.325/>.
- 598
- 599 Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. Metaphor detection via explicit basic
 600 meanings modelling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Pro-*
 601 *ceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*
 602 *2: Short Papers)*, pp. 91–100, Toronto, Canada, July 2023. Association for Computational Lin-
 603 *guistics.* doi: 10.18653/v1/2023.acl-short.9. URL <https://aclanthology.org/2023.acl-short.9/>.
- 604
- 605 Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of lan-
 606 guage models to interpret figurative language. In Marine Carpuat, Marie-Catherine de Marn-
 607 effe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North*
 608 *American Chapter of the Association for Computational Linguistics: Human Language Technolo-*
 609 *gies*, pp. 4437–4452, Seattle, United States, July 2022. Association for Computational Lin-
 610 *guistics.* doi: 10.18653/v1/2022.naacl-main.330. URL <https://aclanthology.org/2022.naacl-main.330/>.
- 611
- 612 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike
 613 Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining
 614 approach. *CoRR*, abs/1907.11692, 2019a. URL <http://arxiv.org/abs/1907.11692>.
- 615
- 616 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike
 617 Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining
 618 approach. *arXiv preprint arXiv:1907.11692*, 2019b.
- 619
- 620 Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn
 621 Walker. Creating and characterizing a diverse corpus of sarcasm in dialogue. In Raquel Fernandez,
 622 Wolfgang Minker, Giuseppe Carenini, Ryuichiro Higashinaka, Ron Artstein, and Alesia Gainer
 623 (eds.), *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and*
 624 *Dialogue*, pp. 31–41, Los Angeles, September 2016. Association for Computational Linguistics.
 625 doi: 10.18653/v1/W16-3604. URL <https://aclanthology.org/W16-3604/>.
- 626
- 627 Theodore Papamarkou, Tolga Birdal, Michael M. Bronstein, Gunnar E. Carlsson, Justin Curry, Yue
 628 Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, Vasileios Maroulas, Nina Mi-
 629 olane, Farzana Nasrin, Karthikeyan Natesan Ramamurthy, Bastian Rieck, Simone Scardapane,
 630 Michael T Schaub, Petar Veličković, Bei Wang, Yusu Wang, Guowei Wei, and Ghada Zamzmi.
 631 Position: Topological deep learning is the new frontier for relational learning. In Ruslan Salakhut-
 632 dinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix
 633 Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, vol-
 634 *ume 235 of Proceedings of Machine Learning Research*, pp. 39529–39555. PMLR, 21–27 Jul
 635 2024. URL <https://proceedings.mlr.press/v235/papamarkou24a.html>.
- 636
- 637 Ilan Perez and Raphael Reinauer. The topological bert: Transforming attention into topology for
 638 natural language processing, 2022. URL <https://arxiv.org/abs/2206.15195>.
- 639
- 640 Irina Proskurina, Ekaterina Artemova, and Irina Piontkovskaya. Can BERT eat RuCoLA? topo-
 641 logical data analysis to explain. In Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Maciej
 642 Ogrodniczuk, Senja Pollak, Pavel Přibáň, Piotr Rybák, Josef Steinberger, and Roman Yangarber
 643 (eds.), *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (Slavic-*
 644 *NLP 2023)*, pp. 123–137, Dubrovnik, Croatia, May 2023. Association for Computational Lin-
 645 *guistics.* doi: 10.18653/v1/2023.bsnlp-1.15. URL <https://aclanthology.org/2023.bsnlp-1.15/>.
- 646
- 647 Raphael Reinauer, Matteo Caorsi, and Nicolas Berkouk. Persformer: A transformer architecture for
 648 topological machine learning, 2022. URL <https://arxiv.org/abs/2112.15210>.
- 649
- 650 Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel
 651 Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor

- 648 Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini
 649 Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bu-
 650 llian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan
 651 Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten
 652 Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan
 653 Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. Scaling
 654 up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*, 2022. URL
 655 <https://arxiv.org/abs/2203.17189>.
- 656 Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL <https://qwenlm.github.io/blog/qwen2.5/>.
- 657
- 658 Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- 659 Eduard Tulchinskii, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, and Serguei Barannikov.
 660 Rtd-lite: Scalable topological analysis for comparing weighted graphs in learning tasks, 2025.
 661 URL <https://arxiv.org/abs/2503.11910>.
- 662
- 663 Yu Yan, Sheng Sun, Zenghao Duan, Teli Liu, Min Liu, Zhiyi Yin, LeiJingyu LeiJingyu, and
 664 Qi Li. from benign import toxic: Jailbreaking the language model via adversarial metaphors.
 665 In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.),
 666 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4785–4817, Vienna, Austria, July 2025. Association for Computational
 667 Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.238. URL
 668 <https://aclanthology.org/2025.acl-long.238/>.
- 669
- 670 Cheng Yang, Puli Chen, and Qingbao Huang. Can ChatGPT’s performance be improved on verb
 671 metaphor detection tasks? bootstrapping and combining tacit knowledge. In Lun-Wei Ku, Andre
 672 Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association
 673 for Computational Linguistics (Volume 1: Long Papers)*, pp. 1016–1027, Bangkok, Thailand,
 674 August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.57.
 675 URL <https://aclanthology.org/2024.acl-long.57/>.
- 676
- 677 Jingjie Zeng, Liang Yang, Zekun Wang, Yuanyuan Sun, and Hongfei Lin. Sheep’s skin, wolf’s
 678 deeds: Are LLMs ready for metaphorical implicit hate speech? In Wanxiang Che, Joyce Nabende,
 679 Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meet-
 680 ing of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16657–16677,
 681 Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-
 682 251-0. doi: 10.18653/v1/2025.acl-long.814. URL <https://aclanthology.org/2025.acl-long.814/>.
- 683
- 684 Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural lan-
 685 guage processing. pp. 1953–1959, 08 2013.
- 686
- 687
- 688
- 689
- 690
- 691
- 692
- 693
- 694
- 695
- 696
- 697
- 698
- 699
- 700
- 701