Assessing Logical Coherence of LLMs via Fine-Grained NLI

Jon F. Apaolaza[1], **Begoña Altuna**[2], **Aitor Soroa**[1], **Inigo Lopez-Gazpio**[1]
[1]HiTZ Basque Center for Language Technology - Ixa NLP Group
University of the Basque Country UPV/EHU
[2]GOI institute, Basque Summer University (UEU)
[1]{jonfelix.apaolaza, a.soroa, inigo.lopez}@ehu.eus, [2]begona.altuna@ueu.eus

Abstract

Natural Language Inference (NLI) is a long-standing probe of models' reasoning capabilities, yet it remains unclear how state-of-the-art systems represent and combine logical clauses in a way that supports robust generalization. We study directional effects in deductive NLI and introduce causal coherence, an evaluation paradigm that tests whether predictions remain consistent when the directionality of inference is reversed. Using fine-grained minimal-pair phrase data from PhrasIS, we evaluate encoder, decoder, and encoder–decoder transformers and analyze their behavior under both standard and manipulated settings. Our results show that models frequently fail to maintain logical stability when directionality varies, indicating shallow pattern matching rather than genuine clause composition. We formalize soft and hard causal coherence to disentangle directional consistency from correctness, and we provide an error analysis that highlights systematic failures involving semantic relations. Our findings suggest that deductive causal reasoning and coherence remain missing components in current transformer architectures, and that addressing them is necessary for reliable NLI.

Keywords: Textual Entailment and Paraphrasing, Evaluation Methodologies, Semantics

1. Introduction

Evaluating logical constructs that natural text represent is a challenging task, with no one-size-fits-all solution (Beltagy et al., 2013). Typically, such evaluations aim to compare different model architectures that produce distinct representations for the input, with the broader goal of assessing functional linguistic competence, that is, how models leverage their knowledge constructs to generalize to previously unseen reasoning challenges. While real-world applications or high-level abstraction tasks remain the ultimate objective (Zettlemoyer and Collins, 2007), evaluations conducted on manually annotated, intermediate or fine-grained tasks often prove more practical and informative.

Over the last decade, natural language inference (NLI) datasets have become a popular resource to evaluate and contrast meaning representation systems and logical compositionality. One of the first efforts were the Recognizing Textual Entailment (RTE) task (Dagan et al., 2010; Bentivogli et al., 2009) and the SICK dataset (Bentivogli et al., 2016) which required systems to classify sentence-text pairs as entailment, contradiction, or non-entailment. A closely related line of work is paraphrase detection, where the task is to determine whether two sentences convey identical meaning (Dolan et al., 2004). Although sentence-level inference has become a popular resource for evaluation, fine-grained tasks often offer more precise insights and enable detailed error analysis under controlled conditions as multiple evaluation issues overlap at sentence level, and, also, subtle distinctions and phenomena are obscured (Levy et al., 2013; Lopez-Gazpio, 2024).

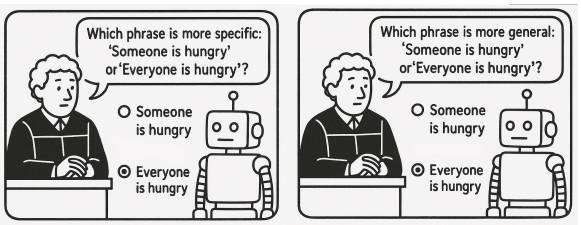

Figure 1: The figure illustrates a failure to maintain logical consistency under manipulated directionality exposing lack of reasoning coherence. The left figure shows the correct inference deduction as the co-ocurring logical propositions are higher in number in the phrase "Everyone is hungry", from which "Someone is hungry" can be derived.

Recently, to better assess the true capabilities of LLMs in NLI, compositional generalization evaluations have been proposed (Chalmers, 1993; Hupkes et al., 2020; Geiger et al., 2020; Fu and Frank, 2024). This paradigm evaluates whether models can predict unseen compositional inferences when their constituent primitive inferences have been observed during training. Extensive research on model performance and techniques to enhance compositionality have also been made in relation to reasoning coherence and the reversal curse.

The Reversal Curse (RC) has emerged as one of the most critical yet underappreciated failures of LLMs in NLI. First documented by Berglund et al. (2024), RC denotes the inability of autoregressive LLMs to infer bidirectional relationships despite being explicitly trained in one directional form. Con-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5431–5444
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

cretely, when a model is trained on a sentence such as *"A relates to B"*, it consistently fails to generalize to the logically equivalent inverse *"B relates to A"*. Although this symmetry may appear trivial from a human perspective, experiments demonstrate that autoregressive models not only fail to resolve the inverse statement but often perform no better than random chance. This logical inconsistency reveals a deep structural limitation of causal transformer architectures and undermines the reliability of LLMs when applied to reasoning tasks that depend on mutual entailment. Moreover, Apaolaza et al. (2025) suggest that RC is not a minor issue but a fundamental limitation in current LLMs as they systematically fail to generalize bidirectional relationships, undermining their logical and causal reasoning. According to the authors, empirical evidence has shown that autoregressive models such as GPT and LLaMA remain vulnerable to this flaw regardless of scale, whereas bidirectional encoders derived from BERT are largely immune.

This investigation claims that RC is a specific instance of a broader deductive incapacity to maintain logical consistency and coherence (See Figure 1), which also arises in scenarios beyond those previously analyzed in Berglund et al. (2024). Although NLI has long been used to assess the reasoning capabilities of models, it remains unclear how state-of-the-art systems represent and combine logical clauses in ways that enable genuine generalization and maintain reasoning coherence, that is, the capacity to maintain logical stability when directionality of the inference varies. In this work, we examine the role of directional causality in shaping compositional generalization within deductive NLI. To this end, we introduce a novel evaluation metric based on causal coherence and evaluate transformer-based state-of-the-art models in a fine-grained minimal pair phrase dataset. **Soft causal coherence** (SoftCoh) measures how often a model gives consistent and coherent predictions between an original pair (e.g. "*Everyone is hungry*" $\Rightarrow$ "*Someone is hungry*") and its reversed pair counterpart (e.g. "*Someone is hungry*" $\Leftarrow$ "*Everyone is hungry*"), without checking whether the prediction is correct according to the gold standard annotation. **Hard causal coherence** (HardCoh) is stricter than SoftCoh as a model is hard-coherent on a pair if it predicts coherent labels for the original and the reversed pairs (consistency) and both label predictions match the gold labels of the original and reversed pairs (correctness).

In line with previous state-of-the-art investigations (Berglund et al., 2024; Lv et al., 2024; Kitouni et al., 2024), our experiments reveal that current models struggle to construct and sustain complex logical clauses when directionality varies, leading to failures in generalization and coherence beyond surface-level patterns. Furthermore, our analysis of causal coherence suggests that deductive causal reasoning constitutes a crucial missing component in transformer-based architectures and that the failure of logical coherence can be generalized to a broader set of logical relations.

This work is structured as follows: Section 1 presents the problem, Section 2 reviews both NLI and the RC. Section 3 defines the experimentation and describes coherence evaluation, Section 4 presents results and Section 5 conducts the data manipulation and coherence evaluation. Finally, Section 6 concludes the work.

2. State-of-The-Art in NLI

The first large NLI dataset was the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015), comprising 570,000 sentence pairs. To address the limited diversity of linguistic structures in SNLI, the Multi-Genre NLI (MultiNLI) dataset (Williams et al., 2018) was introduced, featuring 433,000 sentence pairs. Soon, the need for a multilingual benchmark motivated the creation of the Cross-lingual NLI (XNLI) corpus (Conneau et al., 2018), which initially provided development and test sets for 15 languages and was later expanded to additional languages.

Beyond these general-purpose corpora, with the growing focus on the remarkable capabilities of LLMs, there has been an increasing interest in highly specialized benchmarks (Nie et al., 2020). The cited work attempts to mitigate biases observed in earlier datasets by employing an iterative data collection process that generates increasingly hard examples, thereby exposing model weaknesses and encouraging more robust learning. In fact, it is well established that NLI datasets contain hidden biases (Gururangan et al., 2018; Poliak, 2020; Tsuchiya, 2018; McCoy et al., 2019). Such studies have shown that widely used resources as SNLI and MultiNLI may include superficial cues that allow models to predict labels without truly reasoning over semantic relations, or in some cases even without considering the premise at all.

Fine-grained NLI. Fine-grained evaluation intermediate tasks have proven highly effective in advancing research on logical representation, as they allow isolating specific aspects and concentrating on the phenomenon under investigation (Nie et al., 2020; Apaolaza et al., 2025). Early work on fine-grained inference (MacCartney and Manning, 2007; Angeli and Manning, 2014) defined reasoning within the framework of natural logic, where inference relations are determined by token-level edits. Alternatively, Mitchell and Lapata (2010) provided a notable example of an evaluation bench-

Phrase pair		Entailment
clings	hold on	EQUIVALENT
2 car bombs	Car bombing	FORWARD ent.
22 dead	89 dead	BACKWARD ent.
Doing back flip	Jumping headfirst	OPPOSITION
February 25	April 21	SIMILAR
Brief Gaza truce	Temporary truce	SIMILAR
smiling	are posing	RELATED
says	Syria peace	UNRELATED

Table 1: Example phrase pairs extracted from PhrasIS. The table shows sample phrase pairs along with their annotated entailment relation labels.

mark to explore compositionality at a finer-grained level. The later dataset contains human similarity ratings for three types of word pairs: adjective–noun, noun–noun, and verb–noun. Other related datasets that involve words or phrases include Jurgens et al. (2014), which contains annotated pairs of varying granularity, and Korkontzelos et al. (2013), which focuses on word–phrase similarity pairs mined from dictionary definitions. Also, to evaluate fine-grained reasoning capabilities, subsets of the automatically derived Paraphrase Database (PPDB; Ganitkevitch et al. (2013)) were annotated with similarity scores and inference relations (Pavlick et al., 2015b).

Other efforts include Wieting et al. (2015), who re-annotated two datasets with phrasability judgments, and more recent attempts to annotate fine-grained structure in sentence-level inference include the Interpretable Semantic Textual Similarity task (Agirre et al., 2016). Finally, the PhrasIS benchmark is a dataset that has been designed for the evaluation of fine-grained semantic representations through inference and similarity annotations over phrase pairs (Lopez-Gazpio et al., 2024).

Fine-grained reasoning. As a step towards the evaluation of genuine reasoning capabilities *monotonicity* has been proposed recently (Yanaka et al., 2019a; Geiger et al., 2020) (Yanaka et al., 2019b). Monotonicity aims to determine how the truth of a statement is preserved when parts of a sentence are replaced with more general or more specific phrases, thus, it provides a principled way to reason about entailment. Yanaka et al. (2019b) reveal that NLI models fundamentally struggle with monotonicity reasoning, especially in downward-entailing contexts despite strong performance on standard datasets like SNLI and MultiNLI. Preliminary experiments reveal that models consistently perform better on upward inferences but collapse on downward ones, showing a strong bias learned from existing datasets that rarely include downward reasoning. Overall, their investigation indicates that neural models rely on superficial patterns rather than structural logical reasoning, and that monotonicity remains a significant challenge requiring deeper modeling approaches. Similar conclusions are drawn in Yanaka et al. (2019a) and Geiger et al. (2020), in which the authors address key limitations in current NLI systems by proposing large-scale automatized datasets for monotonicity evaluation. Results indicate some forms of inferences can be learned from data, while others reflect deeper structural limitations. This observation is in line with the results presented in Li et al. (2022) with regard to directionality and inference learning.

RC and inconsistent coherence. The empirical evidence supporting the RC is robust. Berglund et al. (2024) fine-tuned various GPT and LLaMA models on synthetic datasets involving fictitious celebrity relations. While the models performed well on forward queries (e.g., *"Daphne Barrington is the director of A Journey Through Time"*) they failed almost entirely when asked the inverse (*"The director of A Journey Through Time is Daphne Barrington"*). Importantly, this pattern held across different architectures and scales, with GPT-3 (175B), LLaMA-7B, and GPT-3-350M all exhibiting near-zero accuracy on reversed tasks. Similar results were reported by Grosse et al. (2023) using influence functions: reversed sequences consistently showed lower training influence, strengthening the case for an inherent representational asymmetry. Zhu et al. (2024) provided a theoretical explanation: under cross-entropy loss, increasing the conditional weight of B given A does not imply a reciprocal increase in A given B. Thus, the parameter updates in autoregressive models naturally produce directional asymmetries that break symmetry.

Mitigation of RC. Attempts to mitigate RC through prompting have been largely unsuccessful. Chain-of-Thought (CoT) prompting, despite improving other forms of reasoning, does little to alleviate reversal failures (Guo et al., 2024). Even when explicitly provided with demonstrations consistent with training data, models remain unable to reliably reverse relational statements. This suggests that the problem is not superficial but is rooted in the inductive biases of autoregressive objectives and causal attention mechanisms. As a result, researchers have turned to alternative training regimes such as bidirectional causal optimization (BICO) (Lv et al., 2024), data augmentation and permutation training (Guo et al., 2024), autoregressive blank infilling (ABI) (Lv et al., 2024), and factorization agnostic training (PLM) (Kitouni et al., 2024). While these methods improve performance in specific configurations, they do not provide a general solution, and in some cases they fail entirely for tasks requiring the generation of long or complex descriptions.

The persistence of the RC across families and scales of LLMs raises fundamental questions about the nature of reasoning. Apaolaza et al. (2025) pro-

vide a comprehensive synthesis of these issues and highlight four critical findings: (i) RC is robust across model sizes and architectures; (ii) it remains unclear whether RC reflects pure memorization effects or a deeper incapacity for genuine knowledge generation; (iii) the role of pre-training and data augmentation remains inconclusive, and (iv) the scope of the RC is not delimited and it may extend far beyond the narrow cases analyzed, exposing a fundamental flaw in compositional generalization.

3. Problem Formulation

Building on the trajectory of fine-grained evaluation and compositional generalization, our work investigates whether models struggle to generalize in complex deductive reasoning and to what extent their knowledge and common sense reasoning remain coherent when inference directionality varies. This problem closely resembles the RC and subsequent research, yet our experiments reveal that causal incoherence in reasoning extends far beyond the inference labels previously analyzed. To address this task, we evaluate well-known state-of-the-art LLMs in PhrasIS, which is a fine-grained inference benchmark, providing further insights into genuine reasoning capabilities under the standard evaluation scenario (see Section 4) or the manipulated evaluation scenario (see Section 5).

PhrasIS benchmark. PhrasIS is a fine-grained NLI benchmark containing over 10K naturally occurring phrase pairs annotated by human experts with seven labels: EQUIVALENT, FORWARD/BACKWARD entailment, OPPOSITION, SIMILARITY, RELATEDNESS, and UNRELATED to evaluate compositional generalization at the phrase level. Table 1 shows some examples annotated for inference. [1]

The dataset fills a gap between word-level and sentence-level resources, enabling the assessment of compositional models at a finer-than-sentence granularity while extending beyond isolated words. Unlike many prior datasets, PhrasIS is built from naturally occurring text, with phrase pairs extracted from image captions and news headlines. The dual source contributes complementary linguistic properties with news phrases often featuring complex verb groups, and image captions more frequently containing adjectives and descriptive modifiers.

Evaluation is organized into two tracks: Positives track (only related labels; excludes unrelated) and the full track with all labels (positives plus unrelated pairs). Each track is decomposed over three configurations: Images (I) from captions, Headlines (H) from news, and All (A) combining both sources.

[1] Inference labels are further described in Appendix A.1

The inference labels follow the framework of MacCartney and Manning (2007); MacCartney (2009); Pavlick et al. (2015a), where entailment is defined as a semantic containment relation analogous to set inclusion over linguistic expressions of all types, including words and phrases. Consistent with this view of monotonicity, PhrasIS also adopts the principle that entailment can be characterized in terms of the edits required to transform one phrase into another (MacCartney, 2009). In this direction, PhrasIS extends traditional annotations by introducing additional categories that capture finer distinctions of relatedness. These nuanced labels move beyond direct entailment, offering a richer resource for evaluating models' ability to perform complex semantic inference. Also, the controversy surrounding quantification, plurality, and coordination in PhrasIS provides an ideal evaluation ground for testing models' ability to maintain causal coherence across logical inferences with variable directionality.

Causal coherence: SoftCoh and HardCoh In order to test the capacity to maintain deductive causal coherence over the same phrase pairs when directionality varies, we define a new paradigm for evaluation, the so called causal coherence. Whether agnostic to the true label of the phrase pair (SoftCoh) or not (HardCoh), we define the causal coherence as the capacity of a model to maintain the reasoning prediction consistent with what has been predicted in the reversed formulation. SoftCoh is expected to be easier than HardCoh as the correctness of the predictions are dependent on maintaining relative coherence to the model's predictions, but not to annotated true labels. SoftCoh is also set to be more generic than HardCoh as under varying guideline annotations it is defined to yield comparable estimates of relative coherence.

3.1. Soft causal coherence

Let $\mathcal{D} = \{(x_i, x_i^{\text{rev}})\}_{i=1}^{N}$ be the set of original phrase pairs x_i (e.g. *"Someone is hungry"* and *"Everyone is hungry"*) and their reversed counterparts x_i^{rev} (e.g. *"Everyone is hungry"* and *"Someone is hungry"*). Let $f(\cdot)$ denote the model's prediction function and $Rev(\cdot)$[2] denote the reversal function which returns the mutually exclusive counterpart of the original phrase pair's gold standard label. We define a soft

[2] Causal coherence assume that labels are direction-sensitive and therefore mutually exclusive under reversal. This means that when the order of the phrase pair is reversed, the corresponding gold label may also need to change to its directional counterpart (e.g., forward entailment becomes backward entailment), while symmetric labels (e.g., equivalence) remain unchanged. As a result, coherence is evaluated by requiring consistency with the correct reversed form of the label.

Model and Citation	# P
Encoder models	
ModernBERT B / L (Warner et al., 2025)	149M / 395M
ALBERT v2 B / L (Lan et al., 2020)	11.8M / 17.9M
RoBERTa B / L (Liu et al., 2019)	125M / 355M
DeBERTa v3 B / L (He et al., 2020)	184M / 435M
Decoder models (autoregressive)	
OPT (Zhang et al., 2022)	125M / 350M
Pythia (Biderman et al., 2023)	160M / 410M
GPT-2 S / M (Radford et al., 2019)	117M / 345M
Encoder-Decoder models	
BART B / L (Lewis et al., 2020)	140M / 400M
Flan-T5 S / B (Chung et al., 2024)	60M / 220M

Table 2: Size-comparable models selected for evaluation. *B* stands for Base, *L* stands for Large, *M* stands for Medium, and *S* stands for Small.

coherence (SC) indicator for each pair as:

$$SC(x_i, x_i^{\text{rev}}) = \begin{cases} 1, & \text{if } f(x_i) = Rev(f(x_i^{\text{rev}})), \\ 0, & \text{otherwise.} \end{cases}$$

Then, the soft causal coherence of the model is:

$$\text{SoftCoh}(f) = \frac{1}{N} \sum_{i=1}^{N} SC(x_i, x_i^{\text{rev}}).$$

3.2. Hard causal coherence

Let $\mathcal{D} = \{(x_i, x_i^{\text{rev}}, y_i, y_i^{\text{rev}})\}_{i=1}^{N}$ be the set of original phrase pairs x_i, their reversed counterparts x_i^{rev}, and their corresponding gold labels y_i and y_i^{rev}. We define a hard coherence (HC) indicator as:

$$HC(x_i, x_i^{\text{rev}}) = \begin{cases} 1, & \text{if } f(x_i) = Rev(f(x_i^{\text{rev}})) \\ & \text{and } f(x_i) = y_i \text{ and } f(x_i^{\text{rev}}) = y_i^{\text{rev}}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the hard causal coherence of the model is:

$$\text{HardCoh}(f) = \frac{1}{N} \sum_{i=1}^{N} HC(x_i, x_i^{\text{rev}}).$$

4. Experimental Setup

To establish preliminary baselines, we evaluated three families of transformer architectures (Vaswani et al., 2017): bidirectional encoders, autoregressive decoders, and encoder–decoder models (Min et al., 2023) in both tracks defined by Phrasis (Positives and All) taking into account phrase sources (Images, Headlines and All). Implementations[3] were built with the Hugging Face Transformers library (Wolf et al., 2020), using `AutoModelForSequenceClassification` and its task-specific classification head. We conducted hyperparameter search with Optuna (Akiba et al., 2019) in a NVIDIA A100-SXM4-80GB GPU, selecting configurations by validation performance on an 80/20 split of the training set. After selection, each model was retrained on the full training data and evaluated on the corresponding PhrasIS test track. For comparability, we capped each search at 150 trials per model and used stratified batch sampling to preserve label balance. Table 2 lists the open-weight models used to establish these baselines prior to our analysis of causal coherence. The search space comprised: (i) task-head dropout in $[0.0, 0.5]$ with step 0.1; (ii) warm-up ratio in $[0.01, 0.10]$ with step 0.01; (iii) standard learning-rate and weight-decay ranges; (iv) batch size $\{8, 16, \ldots, 64\}$; (v) up to 10 training epochs; (vi) a linear learning-rate scheduler; (vii) AdamW parameters $\epsilon = 10^{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$; and (viii) gradient clipping at 1.0.

Prior evaluation consisted of traditional machine learning methods such as Support Vector Machines (SVM) (Suthaharan, 2016), Random Forest (Breiman, 2001), Gradient Boosting (Friedman, 2001) and other feature-based approaches (Nagarhalli et al., 2021), which yielded best test accuracies of up to 52% in the Positives track and of up to 81% for the full track for SVM models. Our more modern transformer-based baselines therefore provide the first comprehensive comparison across recent architectures.

Evaluation is carried out across both test tracks (see Table 3): (i) a track restricted to positive pairs (Positives), and (ii) a track also containing non-related pairs (All). Furthermore, the dataset is structured for evaluation under three data configurations: (i) phrase pairs derived from image captions (I), (ii) phrase pairs derived from news headlines (H), and (iii) the full release combining both sources (A). Together, PhrasIS produces six distinct evaluation scenarios that will be evaluated independently. We use weighted F-Score for evaluation to take into account the class proportion distribution. The weighted F-Score computes the harmonic mean between precision and recall for each class label (Sokolova and Lapalme, 2009).

4.1. Experiments and results

Table 3 reports weighted F-Score for the *Positives* test track and the *full* test track (*All*). Across all families, scores were consistently higher on *Test All* than on *Test Positives*, indicating that the UNRELATED class in the full track was comparatively easy to discriminate, whereas decisions among the positive track posed the greater challenge.

Encoders performed best overall on the fine-grained *Test Positives* evaluations. In particular, DeBERTa v3 achieved the strongest and most stable results across the base and large architectures. DeBERTa v3-large scored top performance in the Positives combined track, while DeBERTa v3-base achieved best performing results on the standalone Images and Headlines tracks, with ModernBERT-

[3]Code available at: https://github.com/Jonapa/assessing_logical_coherence

Model name	Positives			All		
	I	H	A	I	H	A
Encoder models						
ModernBERT B	0.66	0.63	0.68	0.86	0.85	0.86
ModernBERT L	0.71	0.67	0.72	0.86	0.87	0.87
ALBERT v2 B	0.67	0.65	0.67	0.85	0.86	0.86
ALBERT v2 L	0.68	0.63	0.71	0.86	0.86	0.86
RoBERTa B	0.64	0.54	0.70	0.85	0.87	0.87
RoBERTa L	0.70	0.66	0.71	0.86	0.85	0.88
DeBERTa v3 B	**0.73**	**0.73**	0.75	0.87	**0.88**	0.88
DeBERTa v3 L	0.71	0.72	**0.76**	**0.88**	0.87	**0.89**
Decoder models (autoregressive)						
OPT 125	0.59	0.51	0.61	0.83	0.83	0.84
OPT 350	0.59	0.47	0.61	0.83	0.82	0.84
Pythia 160	0.55	0.42	0.52	0.81	0.77	0.80
Pythia 410	0.62	0.51	0.58	0.83	0.81	0.83
GPT-2 S	0.57	0.46	0.56	0.80	0.78	0.82
GPT-2 M	0.60	0.50	0.62	0.82	0.80	0.83
Encoder-Decoder models						
BART B	0.63	0.44	0.63	0.83	0.83	0.85
BART L	0.67	0.55	0.68	0.86	0.86	0.87
Flan-T5 S	0.46	0.34	0.64	0.85	0.82	0.85
Flan-T5 B	0.66	0.61	0.67	0.86	0.85	0.86

Table 3: Model performance for all configurations. *Test positives* collects evaluation results on the positive track, and *Test All* for the whole track containing the negative unrelated pairs. *I* stands for the images track, *H* stands for the Headlines track and *A* stands for the All track, which combines I and H. Evaluation is given using the weighted F-Score.

large and RoBERTa-large close behind. Autoregressive decoder-only models (OPT, Pythia, GPT-2) underperformed relative to encoders on *Test Positives*, especially on the H track, reinforcing the view that bidirectional context benefits fine-grained inference over short clause pairs. On the encoder model family, the performance in the headlines track has been lower than in the images track, suggesting that more complex relations are contained in news phrases that often feature more complex verb groups. Image captions more frequently contain adjectives and descriptive modifiers, which seem to make the inference easier for models.

Encoder–decoder models showed mixed behavior. While BART-base and Flan-T5-base were competitive on *Test All*, they underperformed encoders on *Test Positives*; Flan-T5-small, for instance, was notably weak on H (0.34). On the full *Test All* track, most encoder–decoder models closed the gap to encoders, again suggesting that the presence of UNRELATED pairs reduces task difficulty. Comparing tracks, the images track generally exceeded the headlines one for most models, and the aggregated full track followed the same pattern, supporting prior observations that the two sources exhibit distinct phenomena and that headlines contains more challenging lexical or pragmatic contrasts. Overall, these results motivate using the positive-pair setting as the primary probe of clause-level reasoning and set a strong encoder-based baseline for the causal coherence analyses that follow.

4.2. Discussion and error analysis

A central observation from our experiments is that transformer-based models were highly sensitive to hyperparameter choices in the fine-grained NLI setting. We explored up to 150 Optuna trials per model and employed stratified batch sampling alongside regularization on the task head (dropout) to mitigate variance. Despite these safeguards, the best validation runs in several scenarios exhibited clear signs of overfitting (e.g., large train–validation gaps and unstable performance across seeds). To reduce selection bias, we replaced evidently overfit trials with alternatives from the Optuna search that offered stronger generalization proxies even when their peak validation scores were slightly higher. Concretely, selection prioritized runs with smaller generalization gaps and stable behavior across images and headlines tracks, rather than maximizing a single held-out score.

Further analysis clarified where errors concentrated. For the top-performing DeBERTa v3-(base and large) models on the *Test Positives* track, we observed that most confusions arose between ENTAILMENT and SIMI, while RELATED was comparatively easy to separate. On the Headlines (H) track, the model showed a tendency to over-predict SIMI in the presence of lexical overlap, suggesting reliance on shallow surface cues; the Images (I) track displayed fewer such confusions, consistent with its more compositional phrasing. These patterns align with the overall gaps between *Test Positives* and *Test All*, reinforcing that the UNRELATED class reduces difficulty in the full track. A more detailed analysis of best performing models' contingency matrices is presented in Appendix A.2.

5. Causal Coherence Evaluation

To perform a coherence exploration based on directionality, we sampled the EQUIVALENT, FORWARD and BACKWARD relations and duplicated them in the reversed order, with the corresponding reversed entailment relation. Note that as for the definition of the forward and backward entailment labels reversing the phrase pairs always produces the opposite-direction label for the new manipulated pair. For the pairs annotated as equivalent, altering the order of the phrase does not change the gold label.

Regarding data partitions we define two splits: (i) the easy or seen scenario split, and (ii) the hard or unseen scenario split. The easy split is composed of the original and manipulated sentences from the validation set of the PhrasIS dataset, which is expected to be easy as the constructs to form the reasoning deduction have already been analyzed by the model during the fine-tuning step. The hard split is composed of the original and manipulated sentences from the evaluation set of the PhrasIS

Model name	Positives			All		
	I	H	A	I	H	A
Encoder models						
ModernBERT B	0.83	0.81	0.87	0.87	0.81	0.89
ModernBERT L	0.89	0.84	0.92	0.91	0.80	0.89
ALBERT v2 B	0.89	0.89	0.95	0.91	0.79	0.89
ALBERT v2 L	0.95	0.83	0.93	0.91	0.87	0.90
RoBERTa B	0.81	0.69	0.90	0.88	0.77	0.89
RoBERTa L	0.87	0.74	0.89	0.91	0.83	0.88
DeBERTa v3 B	**0.96**	0.93	**0.97**	**0.98**	**0.93**	**0.95**
DeBERTa v3 L	**0.96**	**0.97**	0.94	**0.98**	0.90	**0.95**
Decoder models (autoregressive)						
OPT 125	0.72	0.52	0.75	0.80	0.69	0.74
OPT 350	0.69	0.54	0.70	0.73	0.63	0.67
Pythia 160	0.70	0.34	0.59	0.65	0.48	0.64
Pythia 410	0.70	0.52	0.70	0.78	0.60	0.73
GPT-2 S	0.63	0.50	0.67	0.66	0.57	0.70
GPT-2 M	0.73	0.51	0.73	0.73	0.60	0.69
Encoder-Decoder models						
BART B	0.82	0.39	0.82	0.88	0.70	0.89
BART L	0.85	0.56	0.87	0.94	0.75	0.92
Flan-T5 S	0.53	0.24	0.85	0.92	0.69	0.86
Flan-T5 B	0.80	0.70	0.84	0.91	0.77	0.91

Table 4: Performance in PhrasIS for all configurations of the seen soft coherence evaluation.

Model name	Positives			All		
	I	H	A	I	H	A
Encoder models						
ModernBERT B	0.83	0.81	0.85	0.87	0.78	0.89
ModernBERT L	0.89	0.84	0.92	0.91	0.77	0.87
ALBERT v2 B	0.87	0.85	0.92	0.91	0.77	0.87
ALBERT v2 L	0.95	0.80	0.90	0.89	0.85	0.89
RoBERTa B	0.72	0.67	0.89	0.86	0.75	0.87
RoBERTa L	0.84	0.69	0.88	0.90	0.78	0.87
DeBERTa v3 B	**0.96**	0.92	**0.97**	**0.98**	**0.91**	**0.95**
DeBERTa v3 L	0.93	**0.97**	0.94	**0.98**	0.90	0.94
Decoder models (autoregressive)						
OPT 125	0.72	0.52	0.75	0.80	0.69	0.74
OPT 350	0.69	0.54	0.69	0.73	0.63	0.65
Pythia 160	0.70	0.31	0.59	0.65	0.47	0.64
Pythia 410	0.70	0.52	0.67	0.78	0.60	0.73
GPT-2 S	0.63	0.50	0.67	0.65	0.57	0.70
GPT-2 M	0.73	0.51	0.73	0.73	0.60	0.69
Encoder-Decoder models						
BART B	0.80	0.33	0.80	0.85	0.69	0.89
BART L	0.82	0.46	0.83	0.94	0.69	0.90
Flan-T5 S	0.49	0.20	0.81	0.89	0.59	0.83
Flan-T5 B	0.80	0.69	0.82	0.90	0.76	0.90

Table 5: Performance in PhrasIS for all configurations of the seen hard coherence evaluation.

dataset, which is expected to be harder than the seen scenario as the constructs to form the reasoning deduction have not been previously seen by the model.

5.1. Experiments and results

Tables 4–7 summarize SoftCoH and HardCoH across evaluation scenarios and tracks.

SoftCoH vs. HardCoH. Under the seen setting (which includes validation distribution used for model selection, so propositions are partially familiar even after directionality manipulation), scores are uniformly high. In SoftCoH (Table 4), DeBERTa v3-base and large dominate: Base peaks on Positives–I (0.96), Positives–A (0.97), and All–I/All–A (0.98/0.95), while Large tops Positives–H (0.97). HardCoH (Table 5) preserves this ordering with slightly lower values: DeBERTa v3-base again leads Positives–I (0.96), Positives–A (0.97), and All–I/All–A (0.98/0.95), and large remains best on Positives–H (0.97). Encoders overall clearly outperform decoder-only and encoder–decoder families; the latter close the gap on All (easier due to the presence of UNRELATED), but trail on Positives where clause-level distinctions matter most.

Seen vs. Unseen. Moving to the unseen setting introduces a clear performance drop for every family and track. Declines from seen to unseen are on the order of an absolute 0.10–0.20 for encoders on SoftCoH (Tables 4 and 6) and a further few points for HardCoH (Tables 5 and 7), with the H split suffering most. Decoder-only models degrade more sharply, especially on H, indicating poorer generalization to novel clauses when surface cues and familiar templates disappear. Encoder–decoder models sit between encoders and decoders: they remain competitive on All but lose ground on Positives, where fine relations are critical.

Unseen: SoftCoH vs. HardCoH. In unseen–SoftCoH (Table 6), encoders still lead. DeBERTa v3-large is best on Positives–I/H (0.85/0.79), while DeBERTa v3-base is best on Positives–A (0.84) and on All–I/H/A (0.82/0.75/0.80). Notably, Flan-T5-small ties the top score on All–A (0.80), suggesting some benefit from sequence-to-sequence pretraining when the UNRELATED class is present. The consistent I > H pattern persists, reflecting Headlines' denser lexical/pragmatic contrasts. In unseen–HardCoH (Table 7), the stricter criterion amplifies gaps: DeBERTa v3-base again leads Positives–I/Positives–A (0.79/0.79) and All–I/All–A (0.76/0.75), while DeBERTa v3-large edges Positives–H (0.74). Encoder–decoder models (e.g., BART-large at All–A = 0.73) come close on All but remain behind the best encoders on Positives. Decoder-only models lag across the board, their largest deficits occurring on H under HardCoH, where directionality and subtle semantic shifts must be tracked without bidirectional context.

Takeaways. (i) Encoders—especially DeBERTa v3—are consistently the strongest, with seen > unseen and SoftCoH > HardCoH as expected; (ii) H is the hardest split, revealing generalization gaps in prepositional and lexico-semantic phenomena; (iii) All is easier than Positives due to the separable UNRELATED class; and (iv) decoder-only models generalize least well. These patterns motivate a closer look at error types and directional confusions, which we provide next in the manual evaluation section to go beyond automated metrics.

Model name	Positives			All		
	I	H	A	I	H	A
Encoder models						
ModernBERT B	0.73	0.63	0.67	0.71	0.62	0.72
ModernBERT L	0.75	0.67	0.76	0.71	0.66	0.77
ALBERT v2 B	0.78	0.76	0.77	0.75	0.59	0.73
ALBERT v2 L	0.80	0.69	0.78	0.78	0.75	0.78
RoBERTa B	0.71	0.54	0.77	0.74	0.70	0.77
RoBERTa L	0.79	0.65	0.76	0.77	0.70	0.77
DeBERTa v3 B	0.84	0.78	**0.84**	**0.82**	**0.75**	**0.80**
DeBERTa v3 L	**0.85**	**0.79**	0.81	0.77	0.71	0.79
Decoder models (autoregressive)						
OPT 125	0.67	0.40	0.65	0.67	0.52	0.65
OPT 350	0.65	0.37	0.60	0.65	0.47	0.63
Pythia 160	0.53	0.26	0.47	0.53	0.27	0.46
Pythia 410	0.68	0.38	0.58	0.67	0.40	0.56
GPT-2 S	0.58	0.38	0.54	0.56	0.37	0.53
GPT-2 M	0.69	0.34	0.57	0.62	0.41	0.57
Encoder-Decoder models						
BART B	0.72	0.35	0.74	0.76	0.61	0.76
BART L	0.75	0.48	0.80	0.80	0.70	0.79
Flan-T5 S	0.58	0.27	0.77	0.79	0.59	**0.80**
Flan-T5 B	0.80	0.65	0.76	0.79	0.70	0.79

Table 6: Performance in PhrasIS for all configurations of the unseen soft coherence evaluation.

Model name	Test Positives			Test All		
	I	H	A	I	H	A
Encoder models						
ModernBERT B	0.68	0.57	0.63	0.68	0.59	0.68
ModernBERT L	0.71	0.63	0.72	0.68	0.63	0.71
ALBERT v2 B	0.69	0.65	0.70	0.70	0.54	0.66
ALBERT v2 L	0.73	0.61	0.73	0.70	0.65	0.71
RoBERTa B	0.64	0.47	0.72	0.68	0.62	0.71
RoBERTa L	0.71	0.57	0.71	0.72	0.61	0.73
DeBERTa v3 B	**0.79**	0.73	**0.79**	**0.76**	**0.70**	**0.75**
DeBERTa v3 L	0.78	**0.74**	0.77	0.73	0.66	0.74
Decoder models (autoregressive)						
OPT 125	0.62	0.33	0.58	0.61	0.44	0.59
OPT 350	0.60	0.28	0.55	0.61	0.39	0.57
Pythia 160	0.45	0.19	0.38	0.46	0.18	0.40
Pythia 410	0.60	0.29	0.51	0.60	0.32	0.50
GPT-2 S	0.51	0.28	0.47	0.51	0.28	0.47
GPT-2 M	0.60	0.26	0.50	0.54	0.30	0.49
Encoder-Decoder models						
BART B	0.66	0.25	0.68	0.69	0.53	0.71
BART L	0.69	0.37	0.72	0.73	0.63	0.73
Flan-T5 S	0.49	0.18	0.69	0.73	0.44	0.72
Flan-T5 B	0.71	0.54	0.68	0.73	0.60	0.71

Table 7: Performance in PhrasIS for all configurations of the unseen hard coherence evaluation.

5.2. Manual evaluation

We manually inspected the outputs of the best-scoring system on the hard–unseen setting. Particular attention was paid to cases where one directional judgment was correct but the reverse (or an equivalent relation) was mislabeled. Additionally, we initially hypothesized that forward entailment would be easier, since the more specific clause often appears first, but the qualitative analysis remained inconclusive. A consistent source of incoherence involved predictions of the weaker SIMILAR relation in contexts that required a binding relation (ENTAILS or CONTRADICTS). Two main factors contributed:

(i) Prepositional phenomena. The majority of problematic pairs contained prepositions. When each clause was headed by an unrelated preposition, label choices often appeared insensitive to meaning, yielding near-random assignments. By contrast, when only one clause carried a preposition (e.g., *in the water, the water*), predictions were typically correct, reinforcing the idea that the model relied on surface alignment rather than modeling prepositional semantics. However, the effect of unrelated prepositions in models' reasoning seems to be conflicting and opens the path to further analyze how prepositions are processed.

(ii) Lexico–semantic and pragmatic variation. Some errors reflected limited treatment of near-synonymy or taxonomic relations, as in *"for huge democracy rally", "for democracy march"*, where *"rally"* vs. *"march"* triggered SIMILAR instead of the expected entailment/equivalence. We also observed world-knowledge gaps; for instance, *"a black dog", "the German shepherd dog"* was judged SIMILAR rather than the plausible entailment, indicating that the model failed to connect category and attribute knowledge.

On the positive track, the system was robust in trivial cases: adding an adjective to a context; mapping plural to singular when a group entails a single element; and purely syntactic alternations where only one clause is prepositionally marked. With numerals, the model behaved better than expected in simple contrasts (e.g., "many items" $\Rightarrow$ "one item"), but performance degraded with beyond-surface-level patterns, where the model defaulted to SIMILAR failing to construct complex clauses. Finally, we noticed a recurring bias with public figures, e.g., "Pope Francis" $\Rightarrow$ "pope", which the model tended to mark as equivalent. While such pairs may be equivalent at specific times, they are not universally so, which reflects training data temporal ambiguity.

6. Conclusion

Humans construct complex logical clauses grounded in world knowledge and common sense and are able to produce genuine and consistent reasoning over unseen challenges. However, it remains unclear how state-of-the-art reasoning models represent and combine such clauses to enable generalization and maintain coherence. Evaluations in the domain of NLI suggest that, despite being pre-trained on vast amounts of data, LLMs continue to struggle when confronted with complex deductive reasoning. These findings highlight persistent gaps in generalization, particularly in tasks that require complex deductions rather than surface-level shallow pattern recognition.

In this work we have defined a novel coherence-based evaluation metric in order to measure the ability to maintain directional correctness and coherence. Our causal coherence metrics formalize direction-aware paired consistency for NLI: unlike prior swap or minimal-pair tests that report accuracy drops or invariance, we require compatibility

with a label-reversal operator, directly capturing the Reversal Curse in a general NLI setting. Delimiting the true scope of directional reasoning capabilities is essential, also to determine if it indeed extends to a wider set of logical relations.

Takeaways for future work. (i) Hyperparameter sensitivity is pronounced under fine-grained labels; naive model selection by peak validation tended to overfit. (ii) Encoder models, while being the strongest overall, still conflate ENTAILMENT with SIMILAR and EQUIVALENCE, indicating reliance on surface similarity. (iii) Future work should emphasize generalization-aware selection, stronger regularization, and targeted augmentation for phenomena that drive the dominant confusions identified above.

7. Ethics statement

This work relies primarily on the publicly available PhrasIS dataset and other standard NLI resources, all of which are released under permissive licenses. We do not collect or annotate any new data, and no personally identifiable or sensitive information is involved. Models employed for experimentation are open-weight transformer architectures, which promotes transparency and reproducibility. We fine-tune these models only for controlled evaluation purposes and do not deploy them in production. Because NLI datasets are known to contain annotation artifacts and potential biases, we explicitly analyze model errors and reasoning inconsistencies to reveal such limitations rather than to reinforce them. To foster reproducibility and responsible research, we share our code and experimental configurations. Our study focuses on understanding logical and causal coherence in language models and aims to inform the development of more reliable and fair evaluation methodologies.

8. Limitations

Although our study provides novel insights into fine-grained NLI and causal coherence, several limitations must be acknowledged. First, our analysis relies primarily on the PhrasIS dataset, which, while rich in fine-grained semantic relations, may not fully capture the breadth of linguistic phenomena present in real-world reasoning tasks. As with any single benchmark, there is a risk of overgeneralizing conclusions beyond its domain or annotation scheme. In particular, the distribution of inference types and the specific formulation of phrase pairs may bias models toward certain reasoning strategies not representative of broader language use.

Second, our proposed causal coherence metric evaluates consistency in predictions under directionality manipulation, but coherence does not necessarily imply correctness. A model may be consistently wrong in both directions, inflating soft coherence scores. Conversely, hard coherence depends on gold labels, which themselves may contain ambiguity or annotation disagreement, especially in complex cases involving quantification, plurality, or pragmatic inference. Thus, while causal coherence is a useful diagnostic signal, it should not be interpreted as a complete measure of reasoning ability.

Third, our experiments fine-tune open-weight transformer models in controlled settings, but we do not evaluate large proprietary models or multimodal architectures. Findings may therefore not fully reflect the behavior of state-of-the-art systems. Additionally, hyperparameter sensitivity and overfitting observed during fine-tuning suggest that performance differences could partially result from optimization dynamics.

Finally, our evaluation focuses on static benchmarks and does not consider interactive or context-rich reasoning scenarios. Future work should explore whether causal incoherence persists in more naturalistic or task-oriented settings and investigate whether alternative training objectives or architectural modifications can systematically improve both correctness and coherence.

9. Acknowledgements

Jon F. Apaolaza has worked under support of the HiTZ Chair of Artificial Intelligence and Language Technology (TSI100923-2023-1), funded by MTDFP, Secretaría de Estado de Digitalización e Inteligencia Artificial, ENIA, and by the European Union-Next Generation EU / PRTR, and also holds a PhD grant from the Basque Government (PRE_2025_1_0084). The work has also been supported by DeepMinor project (Project CNS2023-144375) funded by MTDFP/ and by the European Union Next GenerationEU/ PRTR. The authors acknowledge the technical and human support provided by the DIPC Supercomputing Center.

10. Bibliographical References

Eneko Agirre, Aitor Gonzalez-Agirre, Inigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. 2016. Semeval-2016 task 2: Interpretable semantic textual similarity. *Proceedings of SemEval*, pages 512–524.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter opti-

mization framework. https://dl.acm.org/doi/abs/10.1145/3292500.3330701.

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *EMNLP*, pages 534–545.

Jon F Apaolaza, Begoña Altuna, Aitor Soroa, and Inigo Lopez-Gazpio. 2025. Exploring the dilemma of causal incoherence: A study on the approaches and limitations of large language models in natural language inference. *Procesamiento del Lenguaje Natural*, 74:207–219.

Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 11–21, Atlanta, Georgia, USA. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

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.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

David J Chalmers. 1993. Connectionism and compositionality: Why fodor and pylyshyn were wrong.

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. https://www.jmlr.org/papers/v25/23-0870.html.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 16:105–105.

B. Dolan, C. Quirk, and C. Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 350.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Xiyan Fu and Anette Frank. 2024. The mystery of compositional generalization in graph-based generative commonsense reasoning. *arXiv preprint arXiv:2410.06272*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. 2024. "Mitigating Reversal Curse in Large Language Models via Semantic-aware Permutation Training". In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11453–11464, Bangkok, Thailand. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. https://arxiv.org/abs/2006.03654.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Ouail Kitouni, Niklas Nolte, Diane Bouchacourt, Adina Williams, Mike Rabbat, and Mark Ibrahim. 2024. The Factorization Curse: Which Tokens You Predict Underlie the Reversal Curse and More. *arXiv preprint arXiv:2406.05183*.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. https://arxiv.org/abs/1909.11942.

Omer Levy, Torsten Zesch, Ido Dagan, and Iryna Gurevych. 2013. Recognizing partial textual entailment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 451–455, Sofia, Bulgaria. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Tianyi Li, Mohammad Javad Hosseini, Sabine Weber, and Mark Steedman. 2022. Language models are poor learners of directional inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 903–921, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

I Lopez-Gazpio, J Gaviria, P García, H Sanjurjo-González, B Sanz, A Zarranz, and Maritxalar Montse. 2024. Phrasis: Phrase inference and similarity benchmark. *Logic Journal of the IGPL*, 32(6):1088–1101.

Inigo Lopez-Gazpio. 2024. Revisiting challenges and hazards in large language model evaluation. *Procesamiento del Lenguaje Natural*, 72:15–30.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2024. An analysis and mitigation of the reversal curse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13603–13615, Miami, Florida, USA. Association for Computational Linguistics.

Bill MacCartney. 2009. *NATURAL LANGUAGE INFERENCE*. Ph.D. thesis, STANFORD UNIVERSITY.

Bill MacCartney and Christopher D Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. https://dl.acm.org/doi/10.1145/3605943.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Tatwadarshi P Nagarhalli, Vinod Vaze, and NK Rana. 2021. Impact of machine learning in natural language processing: A review. In *2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*, pages 1529–1534. IEEE.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015a. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1512–1522. Association for Computational Linguistics.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015b. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. https://insightcivic.s3.us-east-.amazonaws.com/language-models.pdf.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Shan Suthaharan. 2016. Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235. Springer.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. https://arxiv.org/abs/1706.03762.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

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, Remi Louf, Morgan

Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Luke Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687, Prague, Czech Republic. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael Jordan, Jiantao Jiao, Yuandong Tian, and Stuart Russell. 2024. Towards a Theoretical Understanding of the 'Reversal Curse' via Training Dynamics. *arXiv preprint arXiv:2405.04669*.

11. Language Resource References

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. 2016. SICK through the semeval glasses. *Language Resources and Evaluation*, 50(1):95–124.

A. Appendix

A.1. Inference Labels

The inference labels in the dataset build on those defined in natural logic, but are extended to distinguish between similarity and relatedness, following distinctions proposed for word-level semantics (Hill et al., 2015), as well as to account for directionality. In this framework, similarity denotes membership in the same semantic class (excluding equivalence and entailment), while relatedness captures any other meaningful relation between phrases that does not qualify as similarity. The complete set of relations can be summarized as follows: EQUIVALENCE: both phrases express the same meaning; OPPOSITION: the phrases are in an inherently incompatible binary relationship (e.g., antonyms); ENTAILMENT[4]: the meaning of one phrase is subsumed by the other, often involving missing or additional contructs; SIMILARITY: the phrases belong to the same semantic class but are not equivalent, entailed, or opposed; RELATED to denote some kind of weak semantic relation that is not strong enough to be annotated as similar; and, UNRELATED to state there is not any semantic relation involved.

Most of the linguistic phenomena annotated in PhrasIS are straightforward for humans to interpret, with the exception of cases involving quantification. For example, contrary to intuition, *someone* is more general than *everyone* (backward entailment), since the denotation of a quantified noun phrase is defined as the set of predicates it satisfies. This principle also applies to number and coordination phenomena: *women* entails *woman*, but not the reverse.

A.2. Best Performing Models: Contingency Matrices

[4]Note that entailment is decomposed into forward and backward entailment depending on the directionality of the logical relation.

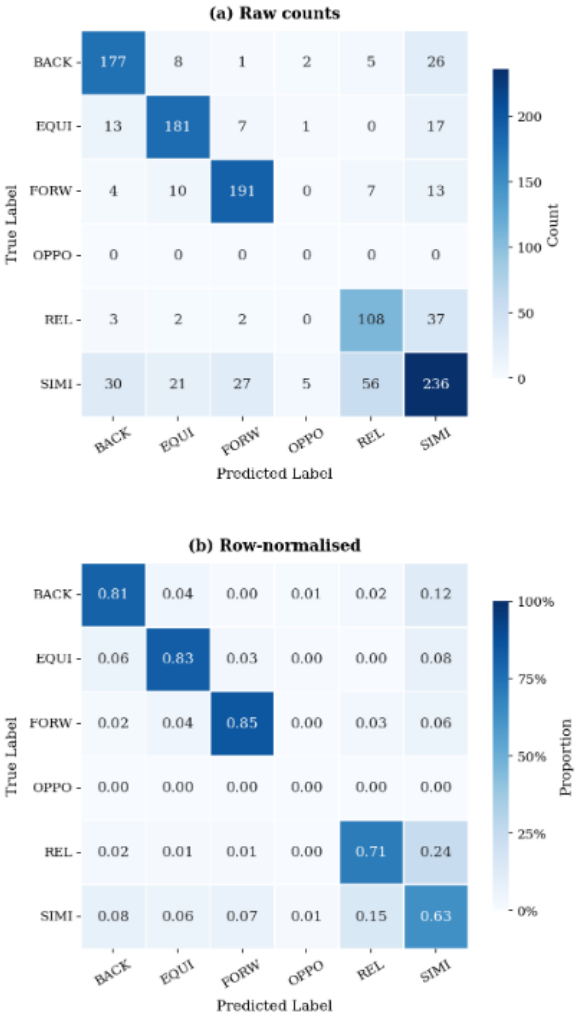

Figure 2: Raw-count and normalized contingency matrices for the DeBERTa v3-base model on the *Test Positives (All)* evaluation track.

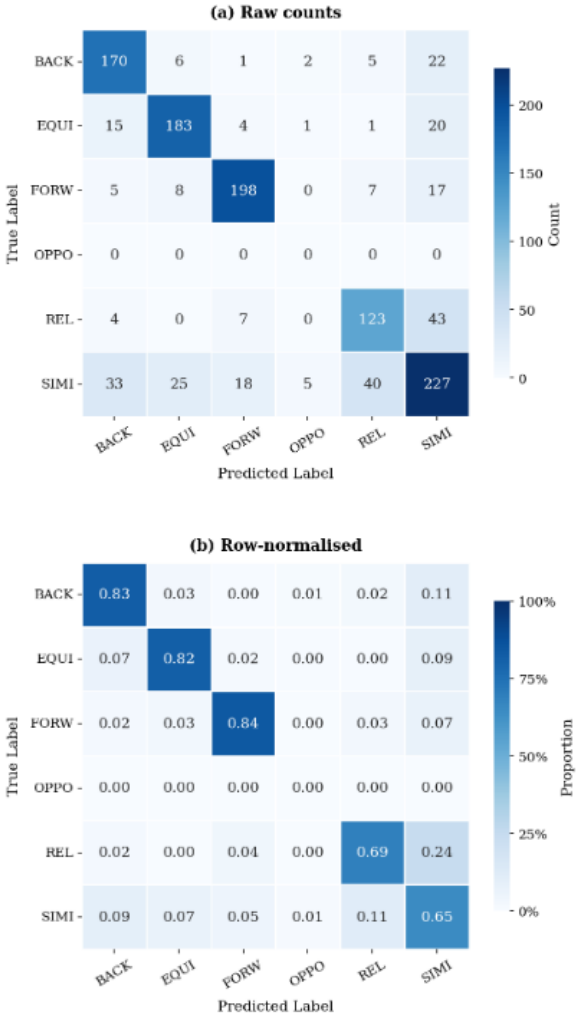

Figure 3: Raw-count and normalized contingency matrices for the DeBERTa v3-large model on the *Test Positives (All)* evaluation track.