When Does Meaning Backfire? Investigating the Role of AMRs in NLI

Junghyun Min[⋆] **Xiulin Yang**[⋆] **Shira Wein**[⋄]
[⋆]Georgetown University [⋄]Amherst College
{jm3743, xy236}@georgetown.edu swein@amherst.edu

Abstract

Natural Language Inference (NLI) relies heavily on adequately parsing the semantic content of the premise and hypothesis. In this work, we investigate whether adding semantic information in the form of an Abstract Meaning Representation (AMR) helps pretrained language models better generalize in NLI. Our experiments[1] integrating AMR into NLI in both fine-tuning and prompting settings show that the presence of AMR in fine-tuning hinders model generalization while prompting with AMR leads to slight gains in GPT-4o. However, an ablation study reveals that the improvement comes from amplifying surface-level differences rather than aiding semantic reasoning. This amplification can mislead models to predict non-entailment even when the core meaning is preserved.

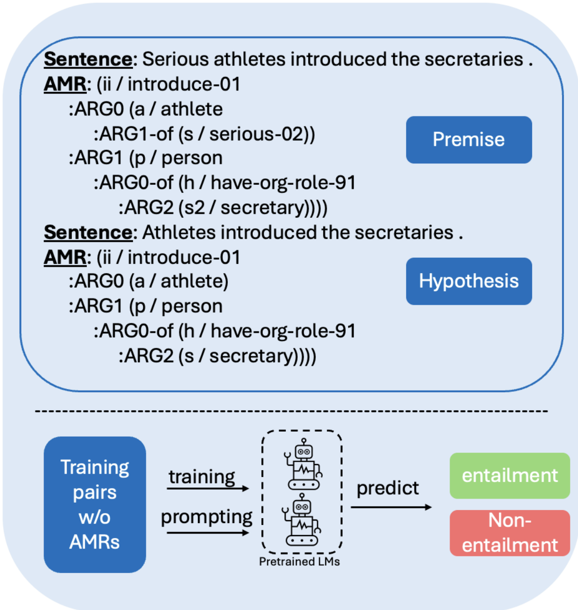

Figure 1: An example of NLI and experiment pipeline of the paper. AMRs are shown in penman notation.

1 Introduction

Since the advent of large language models (LLMs), there has been ongoing debate about the utility of symbolic representations such as Abstract Meaning Representations (AMRs; Banarescu et al., 2013) in (LLM-based) pipelines and existing NLP tasks. While some studies report limited or negative impact of AMRs on mainstream NLP tasks (Jin et al., 2024), recent work has demonstrated their value in specific applications, such as syntactic simplification (Yao et al., 2024) and semantically-controllable text transformation (Li et al., 2025). Perhaps unsurprisingly, incorporating AMR has been particularly well-explored and effective in tasks related to semantics (Wein and Opitz, 2024).

Natural language inference (NLI; Dagan et al., 2010) is a popular task in NLP where the solver is given a *premise* and a *hypothesis*, and asked to determine whether the hypothesis is true if the premise is true. The label space consists of three labels: **entailment** if the hypothesis is true, **contradiction** if the hypothesis is false, and **neutral** if the truth value of the hypothesis cannot be determined; this can also be condensed in two labels: entailment and non-entailment. As shown in Figure 1 "Athletes introduced the secretaries" should be entailed by "Serious athletes introduced the secretaries." Therefore, the label should be **entailment** because the truth of the premise indicates truth of (or *entails*) the hypothesis.

As a meaning-focused task, NLI aligns well with the motivation behind AMRs, i.e., to abstract sentence meaning beyond surface form, given NLI models' tendencies to adopt shallow heuristics rather than understanding the relationship between the premise and the hypothesis, leading to poor generalization to novel data (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019; Serrano et al., 2023). In this paper, we inves-

[1]We publicly release our code at https://github.com/Aatlantise/advarsarial-nli-amr.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 202–211
November 8-9, 2025

tigate whether incorporating AMRs as additional input - either during (a) fine-tuning or (b) prompting - can encourage models to attend more to abstract meaning, thereby improving generalization and overall performance. As illustrated in Figure 1, we add AMRs to either the training data or prompts then evaluate how the addition of AMR affects generalization performance. We find that AMRs generally hinder performance in both fine-tuning and prompting settings, with the exception of prompting on HANS. However, this improvement appears to stem from AMRs amplifying surface-level differences rather than capturing deeper semantic meaning.

2 Related Work

NLI (Dagan et al., 2010) is a hallmark task demonstrating model's ability to understand natural language. Select neural models like `BERT` (Devlin et al., 2019) and `RoBERTa` (Liu et al., 2019) trained on datasets like Multi-genre NLI (MNLI; Williams et al., 2018) and Stanford NLI (SNLI; Bowman et al., 2015) provide test-set performance close to that of humans (Nangia and Bowman, 2019), but the near-human performance on MNLI has been attributed to models optimizing on the spurious correlations between lexical items and labels in the data (Poliak et al., 2018; McCoy et al., 2019; Gururangan et al., 2018; Serrano et al., 2023). The same models that excel in test-set performance suffer from poor generalization to other datasets that represent the same task (Zhou et al., 2020; McCoy et al., 2020; Delbari and Pilehvar, 2025).

Several prior approaches have incorporated logical representations into NLI, for example by combining neural encoders with logical reasoning modules (Chen et al., 2021), training natural logic theorem provers (Abzianidze, 2020), extracting phrase correspondences via natural deduction proofs (Yanaka et al., 2018), or constraining large language models with natural logic inference patterns (Noble et al., 2025). While these works rely on task-specific inference rules or specialized proof systems, our use of AMRs differs in that AMRs provide a broad, task-agnostic semantic abstraction without requiring dedicated engineering. Once an AMR parser is available, AMRs can be used as direct inputs to pretrained models such as BERT (Section 3.2 and ChatGPT (Section 3.3), enabling structured input with minimal task-specific engineering.

LLMs and in-context learning have been used to tackle NLI and generalization in it, with mixed results; Webson and Pavlick (2022) show that the content of prompts do not significantly influence LLMs' performance in NLI tasks, while Kavumba et al. (2023); He et al. (2024) use chain-of-thought reasoning and natural language explanations to improve NLI performance and generalization. However, Zhong et al. (2023) report that its NLI performance is still only comparable to much smaller encoder-only models like `BERT` and `RoBERTa` (Devlin et al., 2019; Liu et al., 2019), leaving adversarial NLI an ongoing area of research.

Recent work on AMRs has set out to utilize AMR graphs for a variety of downstream tasks, including summarization and information extraction (see Wein and Opitz (2024); Sadeddine et al. (2024) for comprehensive overviews). AMRs excel in capturing structure-dependent meaning (Leung et al., 2022) and have shown particular promise in meaning-sensitive tasks such as debiasing translationese (Wein and Schneider, 2024), style transfer (Hua et al., 2023), and sentence-level manipulation (Li et al., 2025), especially when used in conjunction with fine-tuned models.

To the best of our knowledge, Opitz et al. (2023) represents the only prior effort to incorporate AMRs into NLI, and do so for the purpose of interpretable NLI evaluation. They find that metrics based on AMR are robust unsupervised representations of premise-hypothesis relationships when used alongside neural representations like `BERT`.

3 Data & Experiments

3.1 Data & Models

In these experiments, we use two datasets: MNLI (Williams et al., 2018) and HANS (McCoy et al., 2019). MNLI is a crowdsourced dataset, with a test set that is not available to the public. We follow prior work (Wang et al., 2018; Devlin et al., 2019) in taking one of its two developmental splits as the evaluation dataset. Specifically, we take the `matched` developmental set to use as our evaluation dataset. The training dataset includes 297k sentence pairs, while the evaluation set contains around 10k pairs. HANS is a template-based evaluation dataset, with 30k examples. Unlike MNLI and other NLI datasets, its label space consists of only two labels–entailment and non-entailment. We follow prior work (McCoy et al., 2020; Min et al., 2020) in collapsing the model's neutral and contradiction predictions to the single non-entailment label when calculating evaluation metrics, to accommodate the two-class label space of HANS.

We use an off-the-shelf AMR parser from amrlib [2] to parse all the sentences from the two datasets we use. The model is BART-large (Lewis et al., 2019) fine-tuned on AMR 3.0 (Knight et al., 2021). While parsers with higher reported scores exist (e.g. Bevilacqua et al., 2021), we follow Uhrig et al. (2021); Opitz et al. (2023) in selecting an amrlib parser for ease of implementation.

We manually perform a small sanity check over a subset of generated AMRs to verify that AMR parses are acceptable, but do not perform a comprehensive quality check over the entire dataset. We observe that the AMRs produced for sentences in the HANS dataset are generally acceptable, likely benefiting from the sentences' simple structure and short length, though certainly the generated AMRs contain noise; the sentences in MNLI are longer and more complex.

3.2 Experiment 1: Can fine-tuned models benefit from AMR in NLI?

We train three sets of BERT-base models, augmented with AMR information to perform our experiment. We incorporate AMR in three ways: (1) linearized AMR is concatenated to text input (+AMR as text); (2) graph neural network representation of AMR is concatenated to text representation (+AMR as graph); and (3) just the linearized AMR is used as text input (AMR as text only).

We adopt the setup and hyperparameters of previous work in MNLI fine-tuning and HANS evaluation (McCoy et al., 2020; Min et al., 2020). We take the bert-base-uncased model and fine-tune for 3 epochs with a learning rate of 2e-5. While we opt to follow prior work, we note that longer fine-tuning beyond 10 epochs at the same learning rate significantly improves HANS performance in all settings. Each label prediction is made from the [CLS] token's final layer embedding. While the setup is equivalent to those from prior work, we implement the setup in a more modern, current stack. Due to updates in the hardware and software since prior work, slight changes in the resulting model weights are possible. To control for such an effect, we perform a sanity check via baseline in-distribution test set evaluation. Finally, we integrate AMR into the models as text, via *linearization*, removing all newlines and whitespace sequences longer than length two.

3.3 Experiment 2: Can prompt-based models benefit from AMR in NLI?

In this experiment, we evaluate whether incorporating AMRs improves LLMs' performance on NLI, on both the MNLI and HANS dataset, the latter of which remains challenging even after fine-tuning. Jin et al. (2024) find that only instruction-tuned GPT models are capable of reliably processing AMRs. We therefore restrict our evaluation to GPT-4o (Hurst et al., 2024) in zero-shot and 5-shot settings.

We use the following prompt template:

> *You are a helpful assistant trained to determine whether a hypothesis logically follows from a premise. Respond with 'Yes' or 'No'.*
> *Premise: [X].*
> *Hypothesis: [Y].*

Where [X] and [Y] are replaced with the premise and hypothesis in question. The prompt applies to both zero and few-shot settings. We incorporate no additional details or explanations about the task (NLI), the datasets (MNLI and HANS), or the AMRs in our prompt, to best measure the LLM's ability to use representations of meaning for NLI, rather than perform in-context learning. However, it is possible that the model perform better with additional context on the task, dataset, or AMRs.

We test three input conditions: (a) sentence only; (b) AMR only; and (c) sentence + AMR. Label preprocessing follows the same procedure as in the fine-tuning setup for MNLI. In the 5-shot setting, we randomly sampled 5 examples from the training set of each data set. We set the temperature to 0 to ensure deterministic outputs.

4 Results & Discussion

4.1 Experiment 1

We report the accuracies of our fine-tuning models with and without AMRs in Table 1. We report numbers from prior work (McCoy et al., 2020;

[2] https://github.com/bjascob/amrlib-models/releases/tag/parse_xfm_bart_large-v0_1_0

Model	MNLI	HANS
Chance	0.33	0.50
Baseline (McCoy et al., 2020)	0.84	0.57
+Syntactic aug (Min et al., 2020)	0.84	0.65
Ours		
Baseline reproduction (text only)	0.84	0.52
+AMR as text	0.83	0.47
+AMR as graph	0.84	0.49
AMR as text only	0.74	0.51

Table 1: Performance comparison with and without AMR on HANS and MNLI test sets in the fine-tuning setting. Both datasets measure accuracy.

Model	MNLI	HANS
Chance	0.33	0.50
ChatGPT-3.5		
Zhong et al. (2023)	0.89	-
He et al. (2024)	-	0.75
Ours (ChatGPT-4o)		
Text only	**0.91**	0.82
+AMR	0.75	**0.87**
AMR only	0.68	0.70

Table 2: Performance comparison with and without AMR on HANS and MNLI test sets in the LLM zero-shot prompting setting.

Min et al., 2020) in addition to our experiments to serve as comparison baselines and to ensure our setup is correct. Our reported numbers are an average across 10 runs with varying seed.

As shown in Table 1, AMR augmentation does not yield improvements in MNLI performance, nor HANS generalization. Perhaps analogously to previous data-driven attempts at improving generalization (Clark et al., 2019; Min et al., 2020; Yaghoobzadeh et al., 2021), additional AMR information as either text or graph does not affect MNLI performance. Analysis of their confusion matrices reveals AMR adds or subtracts little in terms of MNLI label decision boundary. On HANS performance, We discuss two main findings.

Standalone AMR input for classification intensifies heuristics favoring the entailment label. AMR-only models predict the entailment label for 98.3% of HANS examples, compared to the baseline models at 94.7%. We attribute this to an intensified version of the baseline models' heuristic correlating overlap between the hypothesis and premise to the entailment label, dubbed the lexical overlap heuristic (McCoy et al., 2019). We note this is concurrent with a still competitive MNLI performance, at 84%. We discuss this phenomenon in more detail in Sections A.1 and A.2.

Mixing AMRs and text leads to more (false) negative predictions in novel data. On the other hand, combining AMR information with text strongly affects HANS label decision boundaries in the opposite direction, overriding various shallow heuristics that favor the entailment label observed in McCoy et al. (2020) and in our baseline and AMR-only experiments. Our `+AMR as text` models 86.6% of HANS examples, and `+AMR as graph` models 86.9%, even predicting non-entailment on highly overlapping examples. We attempt to disentangle the effects of AMRs and text in a combined representation in Section A.2, where we find that while AMR can be used to perform NLI, it is less effective than text input and combining the two introduces new artifacts that are more difficult to interpret.

4.2 Experiment 2

The results for prompting with `GPT-4o` are shown in Table 2. We report only the zero-shot results in the main text, as they yield similar overall performance and prediction patterns. Results for the five-shot setting are provided in Section D. Two main observations emerge.

AMRs increase (false) negative predictions. As shown in the table, model performance is consistently lowest when prompted with AMRs alone, while including the original sentence improves results. We find this is because AMRs lead models to make more negative predictions (see Section B). To test this statistically, we fit a logistic regression model predicting non-entailment using SMATCH++ (Opitz, 2023) between hypothesis and premise AMRs and data source (gold vs. predicted). A significant negative interaction (β = -0.042, p < 2e-16) shows that SMATCH similarity influences model predictions more than gold labels.

Further analysis reveals that AMR's sensitivity to surface-level lexical and syntactic variation leads to low structural overlap between semantically equivalent expressions,[3] misleading the model toward non-entailment. This also explains why, on the HANS test set, prompts that include both the sentence and its AMR lead to the highest rate of negative predictions: the AMR repre-

[3] See Section C for an example.

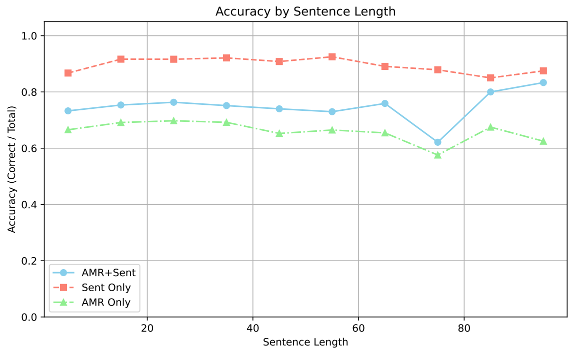

Figure 2: Accuracy of three prompt settings across different sentence lengths on MNLI.

sentation amplifies subtle differences between two otherwise similar strings, making semantic mismatches more salient and pushing the model toward rejecting entailment. Such nuanced contrasts are what HANS is designed to probe in language models, prompting `GPT-4o` to overpredict non-entailment.

AMR does not lead to more robust performance with longer sequence length. Opitz et al. (2023) reported that incorporating AMRs improves robustness in NLI prediction. We investigate whether this finding holds for LLMs. Specifically, we plot accuracy across NLI examples binned by total sequence length (premise + hypothesis). For sequences exceeding 100 words, we group them into a single bin due to their sparsity.

As shown in Figure 2, when `GPT-4o` is prompted with both sentence and AMR inputs, accuracy slightly increases for inputs longer than 80 words. However, this performance remains lower than that of sentence-only prompts across most length bins. We find no evidence that AMR-only prompts enhance robustness to longer sequences.

4.3 Summary

Our fine-tuning experiments suggest that AMR-only models are still susceptible to heuristics. We also observe that combining text with AMR as both graph and text is challenging and results in a strong preference towards the non-entailment label, even for highly overlapping, entailing examples.

Our LLM experiments showcase similar preference towards the non-entailment label. This suggests that AMRs effectively highlight subtle distinctions between minimal pairs, explaining improved **HANS** performance. However, for simpler examples, this heightened contrast can cause the model to overpredict *No*, even for entailing sentence pairs.

5 Conclusion

In this work, we investigate whether AMRs can help PLMs on the task of natural language inference. Specifically, across both fine-tuning and prompting settings, we evaluate whether incorporating AMRs improves entailment classification.

We find that our implementations of AMR integration does not improve performance in fine-tuning, and only lead to slight gains in zero-shot prompting with `GPT-4o`. Importantly, ablation analyses reveal that these gains are not due to deeper semantic understanding, but rather to AMRs exaggerating surface-level differences, which in some cases mislead the model to predict *non-entailment* where entailment holds. Overall, our results suggest that while AMRs offer a promising abstraction mechanism, their integration with LLMs requires careful design to avoid reinforcing shallow heuristics rather than promoting robust reasoning.

Limitations and Future Work

This study focuses on two datasets (MNLI and HANS) and explores a limited set of prompting and fine-tuning configurations. For fine-tuning, we adopt a single AMR linearization strategy; in the prompting setting, we test one prompt template with different conditions. While alternative prompts for zero-shot inference may yield better performance (e.g., Kavumba et al., 2023), our consistent experimental setup enables fair comparisons across conditions. Nonetheless, the findings may not generalize to other inference tasks, domains, or prompting strategies.

Future work could explore more diverse linearization formats, prompt designs, and integration strategies that align AMR structure more directly with model attention or reasoning processes.

Encoder-based models have been shown to be sensitive to minor perturbations in input (Sinha et al., 2021; Jin et al., 2020), and prior work integrating AMR graphs into neural models have used a variety of formats (Wein and Opitz, 2024). Thus, in addition to Python-like and natural language-like representation of AMR's structure (Srivastava et al., 2025; Srivastava and Yao, 2025; Dutt et al., 2025), carefully designing how hierarchical devices in AMRs (e.g. variable names, parentheses, indents, and newlines) could be represented in the embedding space of encoder-only models may be worth further investigations.

Finally, investigating how AMRs interact with LLM decoding beyond surface augmentation may help unlock their full potential in meaning-sensitive tasks.

Responsible Research Statement

We use `ChatGPT-4o` (Hurst et al., 2024) as a coding assistant during the implementation of our experiments, in addition to as a natural language processor.

Acknowledgments

This work was developed on a course project for Advanced Semantic Representation at Georgetown University. We thank Nathan Schneider and our classmates from the course for their helpful comments. A portion of experiments was conducted on the `gcp-gu-hpc-cli` cluster within the High Performance Computing at Georgetown University infrastructure.

References

Lasha Abzianidze. 2020. Learning as abduction: Trainable natural logic theorem prover for natural language inference. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 20–31, Barcelona, Spain (Online). Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI*.

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.

Zeming Chen, Qiyue Gao, and Lawrence S. Moss. 2021. NeuralLog: Natural language inference with joint neural and logical reasoning. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 78–88, Online. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches – erratum. *Natural Language Engineering*, 16(1):105–105.

Zahra Delbari and Mohammad Taher Pilehvar. 2025. Beyond accuracy: Revisiting out-of-distribution generalization in NLI models. In *The SIGNLL Conference on Computational Natural Language Learning*.

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

Ritam Dutt, Shounak Sural, and Carolyn Rose. 2025. Can dependency parses facilitate generalization in language models? a case study of cross-lingual relation extraction. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 338–358, Albuquerque, New Mexic, USA. 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.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Zhijing Jin, Yuen Chen, Fernando Gonzalez Adauto, Jiarui Liu, Jiayi Zhang, Julian Michael, Bernhard Schölkopf, and Mona Diab. 2024. Analyzing the role of semantic representations in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3781–3798, Mexico City, Mexico. Association for Computational Linguistics.

Pride Kavumba, Ana Brassard, Benjamin Heinzerling, and Kentaro Inui. 2023. Prompting for explanations improves adversarial NLI. is this true? {yes} it is {true} because {it weakens superficial cues}. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2165–2180, Dubrovnik, Croatia. Association for Computational Linguistics.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and 1 others. 2021. Abstract meaning representation (amr) annotation release 3.0.

Wai Ching Leung, Shira Wein, and Nathan Schneider. 2022. Semantic similarity as a window into vector- and graph-based metrics. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 106–115, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Hongji Li, Andrianos Michail, Reto Gubelmann, Simon Clematide, and Juri Opitz. 2025. Sentence smith: Formally controllable text transformation and its application to evaluation of text embedding models. *arXiv preprint arXiv:2502.14734*.

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

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. 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.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Bill Noble, Rasmus Blanck, and Gijs Wijnholds. 2025. In the mood for inference: Logic-based natural language inference with large language models. In *Proceedings of the 5th Workshop on Natural Logic Meets Machine Learning (NALOMA)*, pages 33–47, Bochum, Germany. Association for Computational Linguistics.

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.

Juri Opitz, Shira Wein, Julius Steen, Anette Frank, and Nathan Schneider. 2023. AMR4NLI: Interpretable and robust NLI measures from semantic graphs. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 275–283, Nancy, France. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Zacchary Sadeddine, Juri Opitz, and Fabian Suchanek. 2024. A survey of meaning representations – from theory to practical utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2877–2892, Mexico City, Mexico. Association for Computational Linguistics.

Sofia Serrano, Jesse Dodge, and Noah A. Smith. 2023. Stubborn lexical bias in data and models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8131–8146, Toronto, Canada. Association for Computational Linguistics.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. 2021. Perturbing inputs for fragile interpretations in deep natural language processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 420–434, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saurabh Srivastava, Sweta Pati, and Ziyu Yao. 2025. Instruction-tuning llms for event extraction with annotation guidelines. *arXiv preprint arXiv:2502.16377*.

Saurabh Srivastava and Ziyu Yao. 2025. Revisiting prompt optimization with large reasoning models-a case study on event extraction. *arXiv preprint arXiv:2504.07357*.

Sarah Uhrig, Yoalli Garcia, Juri Opitz, and Anette Frank. 2021. Translate, then parse! a strong baseline for cross-lingual AMR parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 58–64, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Shira Wein and Juri Opitz. 2024. A survey of AMR applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6856–6875, Miami, Florida, USA. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? reducing translation effect using Abstract Meaning Representation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–765, St. Julian's, Malta. Association for Computational Linguistics.

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.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. 2018. Acquisition of phrase correspondences using natural deduction proofs. 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 756–766, New Orleans, Louisiana. Association for Computational Linguistics.

Peiran Yao, Kostyantyn Guzhva, and Denilson Barbosa. 2024. Semantic graphs for syntactic simplification: A revisit from the age of LLM. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 105–115, Bangkok, Thailand. Association for Computational Linguistics.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8215–8228, Online. Association for Computational Linguistics.

A Fine-tuning Error Analyses

A.1 Intensified subsequence overlap heuristic with AMR

Compared to text-only MNLI models which are known to incorrectly correlate lexical and sequence overlap to the entailment label (McCoy et al., 2019, 2020; Min et al., 2020), the AMR-only models favor the entailment label even more. The model's preference toward the entailment label results in the AMR-only models consistently predicting non-entailment for around 98% of HANS examples.

It is less likely that the model adopts the subsequence and constituency heuristic, as no text subsequences or constituencies are provided in the training dataset–only AMR parses are provided as input. However, it is possible that a new heuristic had formed. Consider the following two versions of premise-hypothesis pairs:

- **Premise:** *The judge and the president advised the scientist.*

```
(a / advise-01
     :ARG0 (a2 / and
          :op1 (p / person
               :ARG0-of (h / have-org-role
                   -91
                    :ARG3 (j / judge-01)))
          :op2 (p2 / person
               :ARG0-of (h2 / have-org-
                   role-91
                    :ARG2 (p3 / president)
                        )))
     :ARG1 (s / scientist))
```

- **Hypothesis 1, label=non-entailment:** *The scientist advised the judge.*

```
(a / advise-01
     :ARG0 (s / scientist))
     :ARG1 (p / person
          :ARG0-of (h / have-org-role-91
               :ARG3 (j / judge-01)))
```

- **Hypothesis 2, label=entailment:** *The judge advised the scientist.*

```
(a / advise-01
     :ARG0 (p / person
          :ARG0-of (h / have-org-role-91
               :ARG3 (j / judge-01)))
     :ARG1 (s / scientist))
```

The premise and hypothesis AMRs exhibit significant overlap, namely in variables `p`, `h`, `j`, `s`. Given sufficiently many pairs similar to `Premise-Hypothesis 2` in the training set, the model may optimize to correlate variable overlap to the entailment label. Then, when the model predicts on `Premise-Hypothesis 1` pair, instead of considering the semantic structure, it may attend to the significant variable overlap, and predict `entailment`, which is the incorrect answer.

Train setting	Text eval	+AMR eval	AMR only eval
Text	0.84	0.47	
+AMR	0.53	0.83	0.36
AMR only		0.44	0.74

Table 3: MNLI accuracy of our trained models evaluated on each setting. Chance performance is 0.33.

Train setting	Text eval	+AMR eval	AMR only eval
Text	0.96	0.97	
+AMR	0.25	0.13	0.85
AMR only		0.99	0.98

Table 4: Percentage of HANS examples where our trained models evaluated on each setting predict entailment.

A.2 Cross-setting evaluation analysis

To disentangle the effects of text and AMR in `+AMR` models, we evaluate models in not only their own evaluation setting, but in other settings as well. We do not evaluate all models on all settings. Instead, we measure performance on reasonable train-evaluation setting pairs–we do not evaluate `text only` models on `AMR only` settings, and vice versa. `+AMR` models undergo all evaluation settings; all models undergo evaluation in the `+AMR` setting. In this cross-setting evaluation scheme, we do not consider the `+AMR as graph` setting.

First, we observe that cross-evaluation models still perform above chance in MNLI evaluation (0.33), as seen in Table 3, which indicates both text and AMR knowledge can be leveraged despite noise from unseen form. The `+AMR` models' single-mode (text only or AMR only) MNLI accuracies, together with the lower performance of `AMR-only` models compared to `text-only` models indicate that AMR information is more difficult to acquire and use than text input.

Second, we observe that bias towards entailment or non-entailment in MNLI and HANS is strongly correlated, given train-evaluation mismatch (m = 0.6, R^2 = 0.87). Cross-setting evaluation results support the case of a newly developed heuristic for `AMR only` models, as single-

Prompt	MNLI	HANS
Text + AMR	+2,274	+1,759
AMR only	+2,848	+1,393

Table 5: Increase in the number of negative predictions compared to the sentence-only prompt condition.

mode models overwhelmingly predict entailment in HANS examples even when evaluated on +AMR settings, both at above 96%, as seen in Table 4.

On the other hand, it is difficult to pinpoint the cause of the tendency to predict non-entailment in dual-mode models predicting on input containing text. We observe that dual-mode models predict non-entailment for entailing adverbial sentences whose AMRs highly overlap, as shown below:

- **Premise:** *Clearly the bankers waited.*

```
(w / wait-01
    :ARG1 (b / banker)
    :ARG1-of (c / clear-06))
```

- **Hypothesis, label=entailment, pred=non-entailment:** *The bankers waited.*

```
(w / wait-01
    :ARG1 (b / banker)
```

B LLM prediction statistics

The results are reported in Table 5.

C LLM setting error analysis: Example

For example, while the premise *everything you're looking for is available* is semantically equivalent to the hypothesis *everything can be found*, the AMRs for these sentences differ substantially due to lexical choices (e.g., *look for* vs. *find*) and syntactic voice (active vs. passive). The resulting SMATCH++ F-score (Opitz, 2023) between the two graphs is only 27.7.

- **Premise:** *Enter the realm of shopping malls, where everything you're looking for is available without moving your car.*

```
(e / enter-01
      :ARG0 (y / you)
      :ARG1 (r / realm
            :mod (m / mall
                  :mod (s / shop-01)
                  :location-of (a / available
                       -02
                        :ARG2 (e2 / everything
                              :ARG1-of (l /
                                   look-01
                                          :ARG0 y))
                  :manner (m2 / move-01
                        :polarity -
                        :ARG0 y
                        :ARG1 (c / car
                              :poss y))))
                                   ))
```

- **Hypothesis:** *Everything can be found inside a shopping mall.*

```
(p / possible-01
      :ARG1 (f / find-01
            :ARG1 (e / everything)
            :location (ii / inside
                  :op1 (m / mall
                        :purpose (s / shop-01)
                  ))))
```

D 5-Shot Prompting Result

We report the results of our 5-shot prompting experiments in Table 6.

Prompt	MNLI	HANS
Text only	0.89	0.82
+AMR	0.75	0.88
AMR only	0.69	0.67

Table 6: Performance comparison with and without AMR on HANS and MNLI test sets in the LLM five-shot prompting setting.