Hidetsune at SemEval-2026 Task 11: Adapting Pretrained Reasoning Models with Deep Supervision and Inference Refinement for Content-Independent Validity Classification

Hidetsune Takahashi
Waseda University
takahashi78h@toki.waseda.jp

Abstract

This paper presents a system that adopts several training and inference strategies for SemEval-2026 Task 11 Subtask 1, which focuses on binary classification for content-independent validity reasoning in syllogistic inference. Building on fine-tuning of standard language models, additional approaches were explored, including layer-wise deep supervision and in-context learning. Furthermore, models that had been previously trained on datasets related to logical reasoning were adapted to the task through additional fine-tuning. Finally, refinement was performed at the inference stage by adjusting the softmax-based decision threshold of the selected model. The experimental results illustrate how model selection, training strategies, and threshold adjustment affect not only validity accuracy but also robustness against plausibility-driven bias, thereby contributing to improved logical integrity.

1 Introduction

This paper investigates SemEval-2026 Task 11 (Valentino et al., 2026), which focuses on understanding how language models can acquire content-independent reasoning mechanisms and reduce content bias that affects logical reasoning across languages. In particular, this work addresses Subtask 1, which evaluates models' ability to determine the validity of syllogisms in English. In this subtask, the evaluation syllogisms are structurally similar to those in the training data, with relatively limited noise such as irrelevant premises.

This study begins by fine-tuning several Transformer models that have not been additionally adapted for closely related reasoning tasks, using the provided training dataset to examine their baseline performance. Then, extensions are explored including deep supervision, which assigns different importance to selected layers during training, and in-context learning, which is used to evaluate few-shot inference capability through prompting. Following these experiments with the general-purpose Transformer models, models are examined that have been previously pre-trained on datasets related to logical reasoning tasks. Each of these models is further fine-tuned on the task-specific training data. Finally, decision threshold adjustment is applied to the best-performing configuration identified in our experiments.

The results reveal how different experimental conditions influence both the models' ability to recognize logical validity and their robustness against content bias, thereby affecting overall logical integrity. The final system achieved over 95% validity accuracy by fine-tuning an open-source model and further refining its predictions through softmax-based decision threshold adjustment. The code is available on GitHub[1].

2 Background and Previous Work

It has long been debated to what extent language models can perform logical reasoning in natural language, motivating a growing body of prior research. Ozeki et al. (2024) investigated the performance of large language models (LLMs) on syllogistic reasoning tasks using English and Japanese datasets originally designed for psychological experiments assessing human reasoning. Their findings showed that LLMs exhibit reasoning biases similar to those observed in humans, along with other types of systematic errors. Comparable observations were reported by Dasgupta et al. (2024), who interpret such behavior as a reflection of aspects of human intelligence that LLMs are expected to capture. In contrast, there are domains in which LLMs are required to handle syllogistic structures more rigorously, such as legal text summarization, where domain-specific terminology frequently ap-

[1] https://github.com/Hidetsune/SemEval2026_Task11.git

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 193–199

pears and logical consistency is critical (Song et al., 2025).

Previous studies have also highlighted the content effect, wherein the plausibility of a text influences a model's inference of logical validity (Bertolazzi et al., 2026). They demonstrated that plausibility-related representations can causally bias validity judgments through the use of steering vectors, and vice versa. Furthermore, activation-level interventions have been shown to mitigate such content biases, thereby improving the separation between plausibility and logical validity in model predictions (Valentino et al., 2025).

3 Task Description

SemEval-2026 Task 11 Subtask 1 (Valentino et al., 2026) is closely aligned with the research directions discussed in the previous section, as it aims to investigate the ability of language models to perform content-independent logical reasoning.

Regarding the task dataset setup, the organizers provided two datasets: a training set and a test set. The training data was used for model fine-tuning, while the test data was used during the development phase for experimentation and system selection. Participants were further instructed to use the same test data for the final evaluation phase, as no separate dataset was provided.

The training dataset was provided in JSON format, with each instance consisting of an *id*, a *syllogism*, a *validity* label, and a *plausibility* label. The dataset was balanced with respect to validity, containing 480 samples for each of the *true* and *false* classes. For the test dataset, only the *id* and *syllogism* fields were provided, and participants were required to infer the corresponding validity label for each syllogism. The task evaluates model performance using the following metrics.

Overall Accuracy (ACC). Overall Accuracy, which is referred to as validity accuracy later in this paper, measures the proportion of syllogisms for which the model correctly predicts logical validity. This metric reflects the model's basic capability to perform logical reasoning.

Total Content Effect (TCE). Total Content Effect quantifies the extent to which plausibility influences validity judgments. Specifically, it is defined as the average difference in accuracy across four logical–plausibility conditions in English. Lower TCE values indicate higher logical integrity, meaning that the model is less susceptible to content-based bias.

Combined Score. The Combined Score, which is used for official ranking, is defined as:

$$\textit{Combined Score} = \frac{ACC}{\ln(1 + TCE)} \quad (1)$$

This metric rewards high validity accuracy while smoothly penalizing content bias, thereby favoring models that are both accurate and robust to plausibility-driven effects.

4 System Description

The system is organized as a multi-stage pipeline. First, several general-purpose models are adapted to the task by fine-tuning them on the provided training data, establishing baseline systems without prior specialization for closely related reasoning problems. Next, two extensions are incorporated: deep supervision, which assigns differentiated importance to selected layers during training, and in-context learning, which enables few-shot inference through prompt-based conditioning. In addition, models that have been previously trained on datasets related to logical reasoning are included and further fine-tuned using the task-specific data. Based on performance comparisons across these configurations, the final stage applies softmax-based decision threshold adjustment to the best-performing model to refine its prediction behavior.

5 Experimental Setup

5.1 Comparison of Models

In the initial stage of our experiments, the following models were fine-tuned and evaluated to identify an appropriate baseline: BERT (Devlin et al., 2019), TwHIN-BERT (Zhang et al., 2022), and GPT-2 (Radford et al., 2019). Three model families were used to compare the effects of model architecture and pre-training domain on this task. BERT was chosen as the main benchmark because it represents a standard Transformer-based encoder model (Devlin et al., 2019). GPT-2 was included as a decoder-only alternative to examine how an autoregressive model performs in comparison with encoder-based models (Qiu et al., 2020).

The influence of pre-training domain was further examined through TwHIN-BERT, which had been trained mainly on social media text (Zhang et al., 2022). Although syllogistic reasoning is a formal

logical task, each sample in this dataset still consists of short premises and a brief conclusion. In this respect, the overall input style remains short and compact, which is not entirely unrelated to the sentence patterns often found in online posts. This makes TwHIN-BERT a meaningful comparison point for observing whether logical relations can still be captured by a model trained on informal text.

All models were fine-tuned on the original training data, with hyperparameters individually set for each model and variant. Early stopping was applied, with the number of training epochs ranging from one to five and a patience of two epochs. Table 1 presents the development set results for all models and variants, together with the corresponding hyperparameter settings.

As shown in Table 1, BERT-base-uncased achieved the highest performance among all evaluated models. In particular, its validity accuracy was approximately 4 % higher than that of TwHIN-BERT-base and nearly 10 % higher than that of GPT-2-medium. These results indicate that relatively simple models are better suited to this task, where the main objective is to determine the validity of syllogisms rather than to process broad semantic knowledge.

This tendency can also be seen clearly in the lower performance of BERT-large compared with BERT-base. One likely explanation is overfitting, given the relatively limited size of the training data. Larger models also tend to absorb a wider range of factual and contextual knowledge during pretraining, which may cause them to rely more on whether a statement appears plausible in everyday terms than on its actual logical validity (Lampinen et al., 2024). In other words, increasing model size alone does not necessarily improve performance on tasks that require strict formal reasoning. The same pattern was observed in the comparisons between TwHIN-BERT-base and TwHIN-BERT-large, as well as between GPT-2 and GPT-2-medium, suggesting that this tendency is not specific to BERT. This trend is also in line with previous studies reporting that relatively lightweight models can still perform reasonably well in specialized linguistic classification tasks (Takahashi et al., 2024, 2025).

5.2 Deep Supervision

To examine whether emphasizing specific layers could be associated with performance differences, deep supervision was applied to the BERT-base-uncased model, which achieved the highest performance in the initial experiments. Auxiliary losses were introduced in deep supervision, in which different coefficients are assigned to selected layers, thereby encouraging these layers to play a more prominent role during fine-tuning.

Let L_{CE} denote the cross-entropy loss function, and let $\mathbf{z}^{(l)}$ be the output logits of the l-th Transformer layer, with $l = 1, \ldots, 12$. The loss computed from the final layer is defined as:

$$L_{\mathrm{main}} = L_{\mathrm{CE}}\left(\mathbf{z}^{(12)}, y\right) \tag{2}$$

where y denotes the ground-truth label.

To incorporate deep supervision, auxiliary losses are additionally computed from a subset of intermediate and upper layers. Let $\mathcal{S}$ denote the set of supervised layers, and let w_l be the coefficient assigned to layer $l \in \mathcal{S}$. The auxiliary loss is defined as:

$$L_{\mathrm{aux}} = \sum_{l \in \mathcal{S}} w_l \, L_{\mathrm{CE}}\left(\mathbf{z}^{(l)}, y\right) \tag{3}$$

The overall training objective is given by

$$L = L_{\mathrm{main}} + \alpha \, L_{\mathrm{aux}} \tag{4}$$

where α is a scalar hyperparameter controlling the contribution of the auxiliary losses.

Layer-wise coefficients w_l were introduced to control the contribution of auxiliary losses across Transformer layers during fine-tuning, such that layers from the middle to the upper part of the network were given progressively larger roles. No auxiliary loss was applied to lower layers ($l = 1, \ldots, 5$), while progressively larger weights were assigned to higher layers to emphasize task-relevant representations.

Specifically, Layers 6–12 were weighted as $(0.2, 0.3, 0.4, 0.6, 0.8, 1.0, 1.0)$, and all preceding layers were assigned zero. The scalar hyperparameter α was varied between 0.1 and 0.3, while all other hyperparameters and the early stopping strategy followed those used for BERT-base-uncased. Results are reported in Table 2.

As shown in Table 2, higher validity accuracy was observed for all tested values of α, with the highest value obtained at $\alpha = 0.1$, which was approximately 2.6% higher than that of the baseline setting. However, the content effect metric also increased by approximately 3 to 4 points, resulting in a lower combined score overall. These results suggest that while deep supervision can be associated

Model	LR	Epoch	Batch Size	Weight Decay	Warmup Steps	Combined Score (↑)	Validity Acc. (↑)	Content Effect (↓)
BERT-base-uncased	2e-5	5	16	0.01	17	23.1	75.4	8.64
BERT-large-uncased	1e-5	4	8	0.01	34	16.7	72.3	26.6
TwHIN-BERT-base	2e-5	2	16	0.01	17	19.8	71.2	12.5
TwHIN-BERT-large	1e-5	1	8	0.01	34	16.3	66.0	20.3
GPT2	5e-6	3	8	0.01	34	18.2	67.0	13.5
GPT2-medium	3e-6	4	4	0.01	68	16.1	66.0	21.4

Table 1: Performance comparison across different model choices

α	Epoch	Combined (↑)	Validity Acc. (↑)	Content Effect (↓)
0.1	2	21.5	78.0	12.9
0.2	2	21.5	76.4	11.8
0.3	2	21.3	77.5	12.9

Table 2: Results across auxiliary loss weights α

Combined Score (↑)	Validity Accuracy (↑)	Content Effect (↓)
16.2	72.8	31.6

Table 3: Results of in-context learning using Qwen3-4B

with higher task-specific accuracy, it may simultaneously reduce the model's ability to maintain logical consistency.

5.3 In-context Learning

In-context learning was explored as a non-fine-tuning approach to examine the performance of few-shot inference. Specifically, Qwen3-4B (Yang et al., 2025) was employed, and one quarter of the training data provided by the organizers was randomly sampled and incorporated into the prompt as in-context examples.

The prompt was formulated as follows:

> Determine the validity (true or false) of the syllogism "{syllogism}".
> Refer to the examples provided below and respond with "true" or "false".
> {in_context_text}

Here, *syllogism* denotes the input text to be inferred, and *in_context_text* represents a JSON-like formatted set of sampled examples with their corresponding labels (true or false). The results are summarized in Table 3.

Although the validity accuracy of 72.8% is not markedly inferior to that of most models reported in the initial experiments (Table 1), the content effect exhibits higher values than those of any of these models, indicating lower logical integrity.

5.4 Use of Task-specific Models

In addition to the baseline models, models were investigated that had been pre-trained or fine-tuned on logical reasoning–related tasks and were therefore expected to be better aligned with the present task. Specifically, this experiment employed DeBERTa-base-long-nli (Sileo, 2024), RoBERTa-large-wanli (Liu et al., 2022), and BART-large-mnli (Lewis et al., 2020; Yin et al., 2019) models with prior fine-tuning on logical reasoning–related tasks. These pretrained models were further fine-tuned on the dataset provided for this task in order to adapt them more effectively to the characteristics of the present task. As in the baseline experiments, the hyperparameters were individually adjusted according to the architecture and scale of each model. Early stopping was applied in all cases. In this experiment, the patience value was increased from two to three epochs to allow for more cautious convergence. In addition, the maximum number of training epochs was set to ten for the pretrained DeBERTa-base model and seven for the pretrained BART-large model. Table 4 summarizes the results obtained from this experiment.

Focusing on validity accuracy, all three approaches achieved higher scores than the baseline fine-tuning of BERT-base-uncased. In particular, fine-tuning DeBERTa-base-long-nli and BART-large-mnli resulted in validity accuracies exceeding 94%. In contrast, RoBERTa-large-wanli exhibited relatively lower performance than the other two approaches across all evaluation metrics. With respect to the content effect, DeBERTa-base-long-nli and BART-large-mnli yielded lower values, suggesting that these models are less influenced by surface plausibility when determining validity. This tendency also contributed to higher combined scores.

5.5 Thresholding

In the final stage of experiments, softmax-based thresholding was applied to the model predictions, using DeBERTa-base-long-nli, which achieved the highest performance in the preceding experiments. Let K represent the total number of classes, and let

$$\mathbf{z} = (z_1, z_2, \ldots, z_K) \in \mathbb{R}^K \tag{5}$$

Model	LR	Epoch	Batch Size	Weight Decay	Warmup Steps	Combined Score (↑)	Validity Acc. (↑)	Content Effect (↓)
DeBERTa-base-long-nli	2e-5	4	16	0.01	16.9	36.2	96.3	4.26
RoBERTa-large-wanli	1e-5	3	8	0.01	33.8	20.6	77.5	15.0
BART-large-mnli	1e-5	3	4	0.01	67.5	31.4	94.2	6.41

Table 4: Performance comparison across different task-specific models

Threshold	Combined Score (↑)	Validity Acc. (↑)	Content Effect (↓)
0.40	31.9	95.8	6.38
0.50	36.2	96.3	4.26
0.60	39.3	95.8	3.21
0.70	39.1	95.3	3.21

Table 5: Performance comparison across thresholds

denote the vector of unnormalized scores produced by the model. The softmax transformation is defined as:

$$\text{softmax}(z_k) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \tag{6}$$

for each class index $k \in \{1, \ldots, K\}$. This transformation converts the raw model outputs into a categorical probability distribution over the K possible classes, ensuring that each value lies in the interval $[0, 1]$ and that the probabilities sum to one.

To examine the effect of thresholding on model behavior, the decision threshold was varied across four settings, ranging from 0.4 to 0.7 in increments of 0.1. The comparative results are presented in Table 5.

While the validity accuracy reached its maximum at a threshold of 0.50, its variation across different threshold values remained limited, with differences within 1.0%. In contrast, the content effect exhibited a clear decreasing trend, reducing from 6.38 at a threshold of 0.40 to 3.21 at 0.60. These results suggest that validity accuracy is relatively insensitive to the choice of decision threshold, whereas the degree to which the model's predictions are influenced by plausibility can be controlled through threshold adjustment, thereby improving the model's logical integrity in this task.

6 Results and Discussion

During the evaluation phase, no separate test dataset was provided, and the same data used for development was employed for final submission. Based on development results, the DeBERTa-base-long-nli model fine-tuned on the task dataset with a decision threshold of 0.60 was submitted, achieving a combined score of 39.3, validity accuracy of 95.8%, and content effect of 3.21. It ranked 28th out of 45 teams.

Overall, the experiments indicate that while techniques such as deep supervision and in-context learning did not yield direct performance improvements, models previously tuned for validity-oriented reasoning tasks can achieve high validity accuracy when further fine-tuned on this dataset. Moreover, decision threshold adjustment was shown to affect the model's reliance on plausibility, highlighting its role in mitigating content bias and improving logical integrity.

7 Limitations and Future Work

One limitation of this study is that deep supervision was explored using only a single configuration of layer-wise coefficients with three values of the scaling factor α, primarily emphasizing later layers. While this design targets task-specific representations, applying auxiliary losses to lower layers could encourage the model to capture more general structural properties of the input, whose potential contributions remain to be investigated.

A second limitation concerns the number of examples included in the prompt for in-context learning. Due to computational constraints, only a quarter of the training data was sampled and incorporated. Future work could examine how varying the number of such examples affects model performance, as well as explore prompt tuning beyond example selection. Additionally, models that explicitly generate reasoning steps, such as chain-of-thought–style approaches, may provide further insights into the model's decision process.

A third limitation of this study lies in the relatively limited exploration of data-level analysis and preprocessing. For instance, simple data cleaning procedures, such as punctuation removal, might have contributed to improved performance. In addition, data augmentation strategies, including back-translation or the incorporation of external data, were left for future work. Further preprocessing techniques, such as lemmatization, could also be considered, particularly given that a basic model achieved relatively reasonable performance among the general-purpose models evaluated.

8 Conclusion

This work systematically examined multiple training and inference strategies for content-independent validity reasoning. Several general-purpose models were first fine-tuned on the task dataset as baselines, followed by experiments with deep supervision and in-context learning. Models previously trained on data related to logical reasoning were then further fine-tuned on the same task dataset, and the best-performing configuration was refined through softmax-based decision threshold adjustment.

The results indicate that the BERT-base-uncased performed reasonably among standard models, while deep supervision improved validity accuracy at the cost of reduced logical integrity. In contrast, fine-tuning a model that had been pre-trained on logical reasoning–related datasets achieved high validity accuracy, and decision threshold adjustment further reduced content effect without affecting accuracy, suggesting improved robustness to plausibility-driven bias. Future work may explore a broader range of experimental configurations, including exploration of data-level analysis and more diverse layer-wise weighting strategies.

References

Leonardo Bertolazzi, Sandro Pezzelle, and Raffaella Bernardi. 2026. How language models conflate logical validity with plausibility: A representational analysis of content effects. *Preprint*, arXiv:2510.06700.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models show human-like content effects on reasoning tasks. *Preprint*, arXiv:2207.07051.

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.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

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.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2024. Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16063–16077, Bangkok, Thailand. Association for Computational Linguistics.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

Yumei Song, Yongbin Qin, Ruizhang Huang, Yanping Chen, and Chuan Lin. 2025. Legal text summarization via judicial syllogism with large language models. *Journal of King Saud University Computer and Information Sciences*, 37(5):111.

Hidetsune Takahashi, Xingru Lu, Sean Ishijima, Deokgyu Seo, Yongju Kim, Sehoon Park, Min Song, Kathylene Marante, Keitaro-luke Iso, Hirotaka Tokura, and Emily Ohman. 2024. OZemi at SemEval-2024 task 1: A simplistic approach to textual relatedness evaluation using transformers and machine translation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 7–12, Mexico City, Mexico. Association for Computational Linguistics.

Hidetsune Takahashi, Sumiko Teng, Jina Lee, Wenxiao Hu, Rio Obe, Chuen Shin Yong, and Emily Ohman.

2025. OZemi at SemEval-2025 task 11: Multilingual emotion detection and intensity. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 109–115, Vienna, Austria. Association for Computational Linguistics.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *Preprint*, arXiv:2505.12189.

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026. Semeval-2026 task 11: Disentangling content and formal reasoning in large language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.