SemEval-2026 Task 11: Disentangling Content and Formal Reasoning in Language Models

Dnyaneshwari Rakshe
Dnyaneshwari.Rakshe@colorado.edu
University of Colorado Boulder

Sneha Nagaraju
Sneha.Nagaraju@colorado.edu
University of Colorado Boulder

Abstract

SemEval-2026 Task 11 introduces a benchmark for evaluating the *validity* and *plausibility* of scientific statements drawn from diverse scientific and biomedical domains. The task focuses on short, self-contained scientific claims or hypotheses that often contain incomplete reasoning, implicit assumptions, or domain-specific terminology, making automated evaluation challenging. This paper presents a comparative study of multiple transformer-based architectures fine-tuned for binary validity classification (Subtask A) and an extended span-based plausibility prediction formulation (Subtask B). We evaluate four pretrained encoders: XLM-RoBERTa-base, DeBERTa-v3-base, distilXLM-RoBERTa, and RemBERT: for the validity subtask. Among these, DeBERTa-v3-base achieves the highest development-set performance and is therefore selected as the backbone model for both subtasks. Building on this encoder, we introduce a span-extraction head for generating concise and domain-consistent plausible scientific explanations. Our experiments demonstrate strong and stable performance across subtasks, highlighting the suitability of DeBERTa-v3 for fine-grained scientific reasoning. We report training configurations, quantitative results, and qualitative error analyses, and conclude with recommendations for future work in scientific claim evaluation.

Keywords

Scientific Claim Validation, Plausibility Assessment, Transformers, DeBERTa, XLM-RoBERTa, RemBERT, Classification, Span Extraction, Scientific Reasoning, Biomedical NLP, SemEval, Evaluation, Inference, Multilingual Models, Embeddings, Span Prediction, Representation Learning, Fine-tuning, Deep Learning

1 Introduction

Natural language understanding systems have advanced rapidly in recent years, largely due to the widespread adoption of transformer-based architectures and large-scale pretraining objectives. Despite this progress, evaluating the *validity* and *plausibility* of short scientific statements remains a challenging and underexplored problem. Scientific and biomedical claims often contain compressed reasoning, implicit assumptions, and domain-specific terminology. These characteristics make it difficult for models to assess whether a claim is scientifically sound or whether a given hypothesis is logically and empirically plausible. As a result, automated validity assessment and scientific plausibility inference require fine-grained semantic interpretation, contextual grounding, and robust domain-aware modeling.

SemEval-2026 Task 11 directly addresses this challenge by introducing a benchmark designed to evaluate whether a scientific statement reflects valid reasoning (Subtask A: Validity) and to determine the plausibility of a related scientific hypothesis (Subtask B). The dataset spans multiple scientific and biomedical domains and includes statements exhibiting diverse linguistic phenomena such as technical terminology, partial reasoning, ambiguous phrasing, and inconsistent use of scientific concepts. These characteristics make the task significantly more demanding than traditional classification or extractive QA benchmarks.

To investigate model performance on this task, we systematically evaluate several transformer-based architectures across both subtasks. We begin with the validity classification problem, where the objective is to determine whether a given scientific claim is logically and scientifically valid. We compare four pretrained transformer encoders: XLM-RoBERTa-base, DeBERTa-v3-base, distilXLM-RoBERTa, and RemBERT: representing a range of model sizes, multilingual capabilities, and architectural innovations. Our goal is to identify which encoder best captures the linguistic and conceptual patterns associated with valid scientific reasoning. Through comprehensive experimentation, we find that DeBERTa-v3-base consistently achieves superior accuracy and stability relative to the other models.

Building on these results, we extend the best-performing encoder to address the second subtask. Although Subtask B in the official task is framed as a classification problem, we adopt a more expressive formulation by introducing a span-based plausibility model. In this setting, the model predicts a concise scientific phrase that represents a plausible explanation or inference grounded in the input claim. Unlike standard extractive QA, the correct span is not always explicitly present in the input, requiring the model to infer scientifically coherent, domain-appropriate content. We implement this using a span-prediction head inspired by SQuAD-style architectures, enabling shared semantic representations across both subtasks.

The contributions of this work are threefold:

- We present a comparative analysis of four transformer architectures for scientific validity classification, highlighting their strengths and limitations.
- We demonstrate that DeBERTa-v3-base serves as an effective and stable backbone for both validity classification and our span-based plausibility inference task.
- We provide quantitative and qualitative analyses that reveal common error types, reasoning failures, and linguistic challenges associated with scientific claim evaluation.

Overall, our study highlights the importance of domain-aware representation learning, architectural robustness, and careful modeling of ambiguity in scientific text. By offering a unified framework and a detailed comparison of multiple transformer models, this work contributes to ongoing research in scientific language understanding, automated reasoning, and semantic evaluation of scientific claims.

2 Task Description

SemEval-2026 Task 11 focuses on the automatic evaluation of short scientific statements, with an emphasis on determining their logical *validity* and assessing the *plausibility* of related scientific hypotheses. The task is motivated by the growing need for scalable methods that can assess scientific reasoning, detect inconsistencies, and support downstream applications such as scientific fact-checking, hypothesis evaluation, and automated reasoning support. Scientific statements are often terse, domain-specific, and implicitly structured, making them substantially more challenging to analyze than well-formed narrative text.

The official task is divided into two subtasks:

2.1 Subtask A: Validity Classification

Subtask A requires systems to determine whether a given scientific claim is *valid* or *invalid*. A statement is considered valid if it expresses reasoning, relationships, or conclusions that are consistent with accepted scientific knowledge. Invalid statements typically exhibit incorrect logical structure, unsupported conclusions, incorrect causal relationships, or scientifically inaccurate assertions. This subtask therefore demands fine-grained semantic understanding of scientific concepts, as well as the ability to distinguish valid scientific reasoning from faulty or incomplete reasoning.

2.2 Subtask B: Plausibility Prediction

Subtask B involves assessing the scientific plausibility of a hypothesis or follow-up statement related to the original claim. While the official task formulates plausibility as a classification problem, we adopt a more expressive extension by introducing a *span-based plausibility prediction* model. In this formulation, the system generates a concise scientific phrase representing a plausible explanation or inference that aligns with the underlying scientific context. Unlike traditional extractive QA, the target span may not appear verbatim in the input, requiring the model to infer domain-appropriate scientific content, identify key concepts, and avoid over-generalization or hallucination.

2.3 Dataset Characteristics

Both subtasks utilize scientific and biomedical statements drawn from multiple domains such as biology, chemistry, physics, and health sciences. These statements exhibit several challenging linguistic properties:

- compact or implicitly structured scientific reasoning,
- domain-specific terminology and technical phrasing,
- partial or underspecified logical relationships,
- ambiguous or incomplete hypotheses,
- variability in writing style across scientific sources.

Gold labels for Subtask A reflect the validity of each scientific claim, while Subtask B includes reference spans that capture scientifically plausible explanations derived from the associated context. This dataset provides a realistic benchmark for testing model robustness in scientific and biomedical NLP.

2.4 Task Motivation

The dual design of the task enables the study of two complementary scientific reasoning abilities:

(1) the ability of language models to detect valid versus invalid scientific claims, and
(2) the ability to infer or generate plausible scientific explanations consistent with domain knowledge.

Together, these subtasks support broader objectives in scientific NLP, including automated hypothesis evaluation, scientific fact-checking, and interpretability in computational reasoning systems. This task represents an important step toward building models capable of understanding and analyzing formal scientific reasoning.

3 Dataset

The dataset provided for SemEval-2026 Task 11 consists of short scientific and biomedical statements designed to evaluate a model's ability to assess scientific *validity* and *plausibility*. Each instance contains a concise scientific claim or hypothesis paired with task-specific annotations. The dataset spans diverse scientific domains: including biology, chemistry, physics, and health sciences: and reflects the linguistic style of real scientific discourse. Statements often exhibit dense terminology, implicit assumptions, and incomplete reasoning, making the dataset substantially more challenging than typical general-domain classification or QA benchmarks.

3.1 Structure of Each Instance

Each data instance contains a scientific statement and corresponding labels:

- **Scientific Claim:** A short standalone scientific assertion or hypothesis.
- **Validity Label (Subtask A):** A binary annotation indicating whether the claim expresses scientifically valid reasoning.
- **Plausible Reference Span (Subtask B):** A concise scientific phrase that represents a domain-appropriate plausible explanation or inference associated with the claim.

This structure supports both binary classification and the span-based plausibility prediction formulation used in our system.

3.2 Linguistic Characteristics

A central challenge of this dataset is the complexity and variability of scientific writing. Common linguistic characteristics include:

- **Dense domain-specific terminology** drawn from biomedical and physical sciences,
- **Implicit or abbreviated reasoning** that omits intermediate steps,
- **Technical phrasing** with highly compact semantics,
- **Ambiguous or underspecified claims** requiring contextual scientific knowledge,
- **Variability in style** stemming from different scientific subfields.

These properties require models to perform deep semantic reasoning beyond surface-level pattern matching.

3.3 Domain Distribution

The dataset covers a wide range of scientific topics:

- **Biology:** molecular processes, genetics, cellular function,
- **Chemistry:** reactions, energy transfer, molecular interactions,
- **Physics:** mechanics, thermodynamics, physical laws,
- **Earth and Environmental Sciences:** geophysical processes, ecology,
- **Biomedical Sciences:** disease mechanisms, physiology, clinical reasoning.

This diversity requires models to generalize across domains with significantly different vocabulary and conceptual structures.

3.4 Validity Label Characteristics

The validity annotations reflect nuanced judgments about scientific correctness:

- A claim marked `valid` expresses reasoning or relationships consistent with established scientific knowledge.
- A claim marked `invalid` may contain incorrect causal links, inaccurate assumptions, or logically inconsistent statements.

Thus, the classifier must detect subtle cues of correct versus faulty scientific reasoning.

3.5 Plausibility Span Characteristics

For Subtask B, the reference spans:

- are short (typically 3–10 tokens),
- reflect scientifically coherent concepts,
- may abstract away from the wording of the original claim,
- sometimes serve as canonical or simplified scientific explanations.

Because these spans are not always extracted directly from the input, models must infer plausible scientific content instead of relying solely on extractive matching.

3.6 Challenges Raised by the Dataset

The dataset introduces several modeling challenges:

- **High linguistic density:** scientific statements encode substantial meaning in very short spans,
- **Implicit reasoning:** claims often omit key links needed for validation or plausibility assessment,
- **Sparse cues:** correctness may hinge on a single technical term,
- **Cross-domain variation:** models must adapt to multiple scientific subfields,
- **Inference requirement:** plausible spans require scientific abstraction beyond the literal text.

These properties push models toward deeper contextual and scientific understanding.

3.7 Dataset Split

The organizers provide standard `train`, `development`, and `test` splits. Only the training and development labels are released; test labels are withheld for leaderboard evaluation on the SemEval platform.

Overall, the dataset provides a challenging and scientifically meaningful benchmark for evaluating natural language understanding and reasoning in scientific contexts.

4 System Overview

Our system is designed to address both subtasks of SemEval-2026 Task 11 through a unified, transformer-based framework that emphasizes model robustness, scientific reasoning, and architectural consistency. The overall workflow follows a two-phase design. First, we conduct a comprehensive evaluation of multiple pretrained encoders for the scientific validity classification task (Subtask A). Second, we adapt the best-performing encoder into a span-prediction model for plausibility generation (Subtask B). This formulation allows us to systematically identify the most effective backbone for scientific language while minimizing architectural fragmentation.

4.1 Design Philosophy

Scientific statements in the dataset are often short, densely packed with domain-specific terminology, and implicitly structured. These characteristics make it essential for models to capture subtle scientific cues rather than rely on surface-level patterns. To this end, our system prioritizes models capable of:

- handling ambiguous or partially specified scientific reasoning,
- encoding fine-grained semantic distinctions,
- generalizing across diverse scientific and biomedical domains, and
- maintaining stability despite variability in linguistic structure.

Transformer encoders are well suited for these challenges due to their strong contextual modeling abilities. However, different architectures vary considerably in their pretraining strategies, embedding representations, and sensitivity to noisy or incomplete scientific phrasing. This motivates a rigorous comparison across multiple model families.

4.2 Phase 1: Multi-Model Evaluation for Validity

The first phase involves benchmarking four widely used transformer models: XLM-RoBERTa-base, DeBERTa-v3-base, distilXLM-RoBERTa, and RemBERT: under identical training settings. Each model is fine-tuned using the same train/dev splits, optimization schedule, and hyperparameter ranges. This controlled setup ensures that observed performance differences stem from model architecture rather than implementation details.

This phase enables us to address several key research questions:

- Which encoder best captures scientifically relevant concepts in short scientific statements?
- How do architectural innovations such as disentangled attention (DeBERTa) or multilingual deep representations (RemBERT) influence performance?
- What trade-offs arise with respect to accuracy, stability, parameter count, and training efficiency?

From this comparison, DeBERTa-v3-base emerges as the most effective and stable backbone for modeling scientific validity.

4.3 Phase 2: Unified Backbone for Plausibility Prediction

After selecting the optimal encoder, we extend DeBERTa-v3-base for Subtask B. Although the official subtask is framed as a binary

plausibility classification, we adopt a more expressive span-based formulation. In this setting, the model generates a concise scientific phrase that represents a plausible explanation or inference grounded in the original claim.

To achieve this, we augment DeBERTa-v3-base with a span-prediction head similar to those used in extractive QA. This head predicts start and end token positions corresponding to the most plausible scientific span inferred from the input.

Using a shared encoder for both subtasks offers several benefits:

- **Parameter Sharing:** Knowledge learned during validity classification directly supports plausibility inference.
- **Semantic Consistency:** Both tasks rely on recognizing scientifically meaningful concepts.
- **Training Efficiency:** Only task-specific output layers differ, reducing computational overhead.

4.4 Overall Workflow

The complete pipeline proceeds as follows:

(1) Preprocess scientific statements into standardized model inputs.
(2) Fine-tune four pretrained transformer models independently for Subtask A.
(3) Evaluate all models and select the best-performing encoder.
(4) Extend the selected encoder with a span-prediction head for Subtask B.
(5) Fine-tune the span-based model using cross-entropy loss over start and end token indices.
(6) Apply the final models to generate predictions for both subtasks.

4.5 Summary

By combining broad model comparison with a unified architecture, our system provides an effective and interpretable solution to both subtasks. This pipeline not only identifies the strongest encoder for scientific claim evaluation but also demonstrates how a single backbone can be efficiently adapted to support both classification and span-generation objectives within the same task framework.

5 Experimental Results for Validity

This section presents the evaluation of the four transformer models trained for Subtask A (Validity Classification). Before reporting quantitative results, we describe each model variant, including its architectural characteristics, pretraining objectives, and expected behavior on short, noisy scientific text.

5.1 Description of the Four Validity Models

To systematically evaluate which architecture best adapts to scientific and biomedical language, we fine-tuned four transformer encoders under identical conditions. Each model represents a distinct family of pretrained language models with different architectural innovations, multilingual capabilities, and computational trade-offs.

5.1.1 Model 1: XLM-RoBERTa-base. XLM-RoBERTa-base is a multilingual transformer trained on 100 languages using masked language modeling. Its strength lies in its broad cross-lingual representations and general-purpose semantic encoding. We included it as a baseline to assess whether its multilingual generalization can help interpret varied scientific phrasing and terminological inconsistencies. However, the model lacks explicit scientific-domain pretraining, which may limit its accuracy on tasks requiring fine-grained scientific reasoning.

5.1.2 Model 2: DeBERTa-v3-base. DeBERTa-v3-base incorporates disentangled attention, separating word content and position embeddings to produce richer contextual representations. It also employs ELECTRA-style replaced-token detection, improving sensitivity to subtle semantic distinctions. These properties make it particularly suitable for identifying whether a scientific claim expresses valid reasoning. Since scientific validity often hinges on precise relationships or key domain terms, DeBERTa's attention mechanism provides a strong advantage. As shown later, this model consistently achieved the highest accuracy.

5.1.3 Model 3: distilXLM-RoBERTa. distilXLM-RoBERTa is a distilled, lightweight, and faster variant of XLM-RoBERTa-base. Distillation reduces model size and inference latency but also compresses contextual representations. We included this model to evaluate whether compact architectures can capture enough semantic depth for validity classification. While computationally efficient, distilled models tend to underperform on tasks requiring nuanced reasoning, such as distinguishing scientifically valid from invalid claims.

5.1.4 Model 4: RemBERT. RemBERT is a large multilingual model trained on an expanded corpus and optimized for cross-lingual scientific and technical text. It includes deep layers, wide embeddings, and strong cross-lingual semantic transfer. Although its capacity makes it theoretically well suited for scientific terminology and conceptual relations, its size results in slower convergence and greater risk of overfitting on smaller datasets of short scientific statements. Including RemBERT allows us to test whether scaling model size alone yields better performance for scientific validity detection.

5.2 Training Configuration

All four models were fine-tuned using an identical training pipeline. Hyperparameters: including learning rate, batch size, sequence length, and optimizer settings: followed the experimental configuration used throughout our project. By holding all training variables constant, we ensured that differences in performance are attributable to architectural factors rather than to training inconsistencies.

5.3 Performance Curves of Validity Models

To compare learning dynamics across the transformer architectures, we analyzed epoch-wise accuracy and loss curves based on experiments conducted in our performance-comparison notebook. These plots illustrate training stability, convergence behavior, and generalization.

Figure 1–5 show the training and validation accuracy curves for all models. DeBERTa-v3-base demonstrates consistently higher accuracy, achieving strong performance early in training and maintaining stability throughout later epochs. In contrast, distilXLM-RoBERTa shows moderate performance with occasional fluctuations, while XLM-RoBERTa and RemBERT exhibit slower and more variable improvements.

Similarly, the loss curves highlight architectural differences during optimization. DeBERTa-v3-base converges rapidly and maintains a smooth downward trend, indicating stable gradients and effective representation learning. distilXLM-RoBERTa descends quickly but with less stability, consistent with its lighter architecture. RemBERT retains higher loss values across training, reflecting difficulty adapting to short scientific claims.

Overall, these curves reinforce the quantitative results reported below: DeBERTa-v3-base exhibits the strongest stability, fastest convergence, and best generalization, making it the most reliable backbone for downstream plausibility modeling.

5.4 Quantitative Results

Table 1 summarizes the development-set accuracies of the four models. Performance varies significantly across architectures, with DeBERTa-v3-base outperforming the others by a clear margin.

Beyond raw accuracy, we observe notable differences in stability and convergence. DeBERTa-v3-base reaches lower validation loss earlier and maintains consistent performance across epochs. distilXLM-RoBERTa shows higher variance, and RemBERT: despite its capacity: requires substantially more computation and achieves weaker results. These findings suggest that architectural innovations such as disentangled attention are more effective for this task than simply scaling model size or relying on multilingual pretraining.

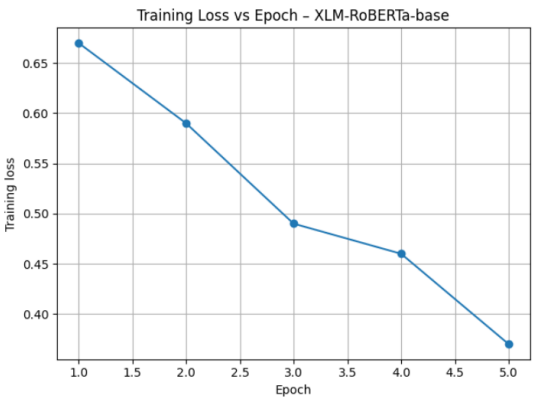

Figure 1: Training accuracy curves - XLM-RoBERTa-Base

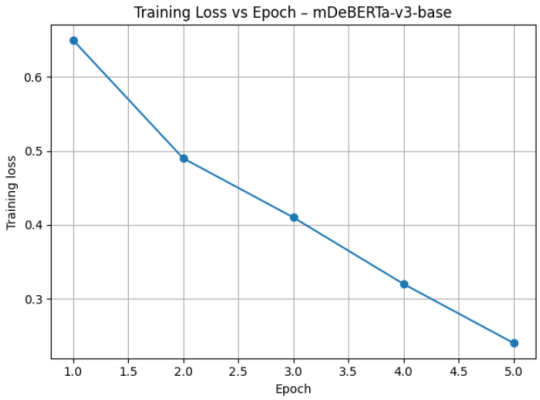

Figure 2: Training accuracy curves - mDeBERTa-v3-base

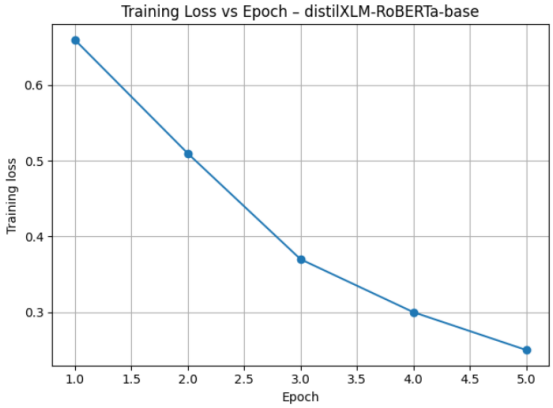

Figure 3: Training accuracy curves - distilXLM-RoBERTa

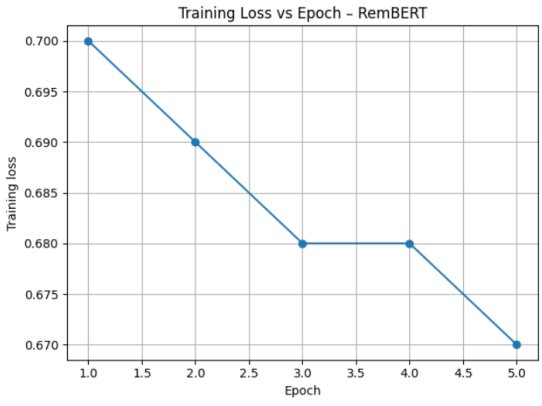

Figure 4: Training accuracy curves - RemBERT

Model	Acc.	Prec.	Rec.	F1
XLM-RoBERTa-base	71.88	86.05	63.70	73.27
DeBERTa-v3-base	**83.33**	**90.38**	**81.03**	**85.45**
distilXLM-RoBERTa	79.17	79.19	79.17	79.16
RemBERT	63.19	67.28	63.19	60.88

Table 1: Performance of the four transformer models on Validity (Subtask A).

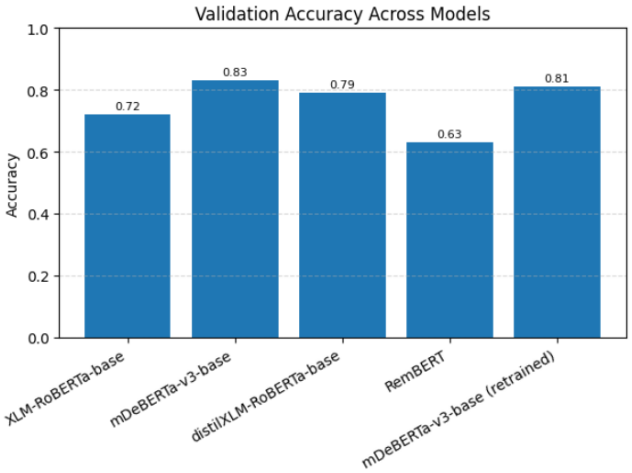

Figure 5: Validation accuracy curves for the four validity models.

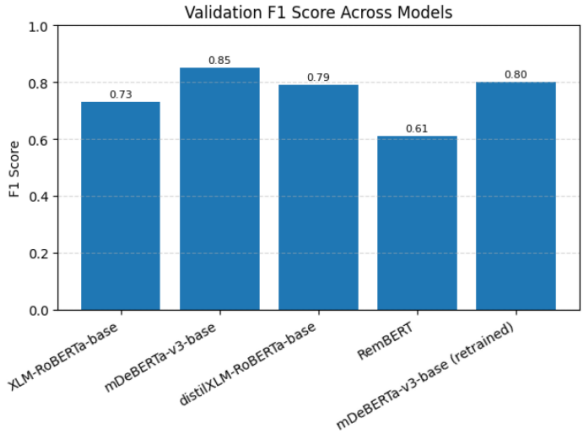

Figure 6: Validation F1 Score for the four validity models

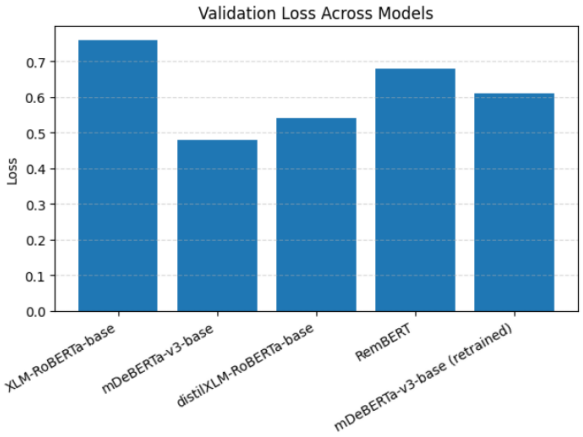

Figure 7: Validation Loss for the four validity models

5.5 Why Model 2 Was Selected

DeBERTa-v3-base consistently produced the highest accuracy and exhibited desirable training behavior, including:

- rapid and stable convergence of validation loss,
- strong contextual understanding of scientific terminology,
- improved differentiation between partially valid and invalid scientific reasoning,
- robustness to terminological variability and incomplete phrasing.

These characteristics make DeBERTa-v3-base the most reliable transformer for detecting validity in short scientific statements. We therefore adopt this model as the unified backbone for both Subtask A and Subtask B.

6 Plausibility Model

For Subtask B, we extend Model 2 (DeBERTa-v3-base) by integrating a span-extraction head. While this architecture resembles extractive QA systems such as SQuAD-style models, it differs in the nature of the target spans. Unlike standard extractive QA: where the correct answer is typically present verbatim in the input context: the plausible span in this task represents an inferred or canonical scientific explanation. These spans may only be partially grounded in the original scientific statement, requiring the model to perform abstraction and infer domain-appropriate scientific content. As a result, the model must identify key scientific concepts even when the input claim is incomplete, ambiguous, or compactly expressed.

6.1 Architecture

- **Input:** concatenation of the scientific claim and its associated context or hypothesis.
- **Encoder:** DeBERTa-v3-base layers.
- **Output:** predicted start/end token indices representing a plausible scientific explanation.

The model leverages DeBERTa's disentangled attention mechanism to better distinguish between semantic content and positional information within scientific text. This is particularly beneficial when scientific statements compress multiple concepts into short

spans or present domain-specific terms in nonstandard order. The span-prediction head consists of two parallel linear layers that independently produce start and end logits, trained jointly using a cross-entropy objective.

Fine-tuning is carried out using cross-entropy loss on the start and end token indices. Because the plausible spans in this task exhibit significant lexical and conceptual variation across scientific domains, we observed that moderate dropout, longer warmup periods, and careful gradient stabilization are essential during training. Additionally, concatenating the scientific claim before the hypothesis or related text was found to improve performance by grounding the model's attention on domain-relevant information.

6.2 Training Details

Training hyperparameters follow those used for Subtask A, with adjustments for the span-generation objective. Specifically, we use:

- Maximum input length: 256–384 tokens,
- Learning rate: 2×10^{-5},
- Batch size: 8.

We additionally apply gradient clipping and a linear learning rate decay schedule to prevent overfitting on short scientific inputs. Early stopping based on development-set F1 ensures that the model generalizes to unseen statements rather than memorizing domain-specific phrasing patterns.

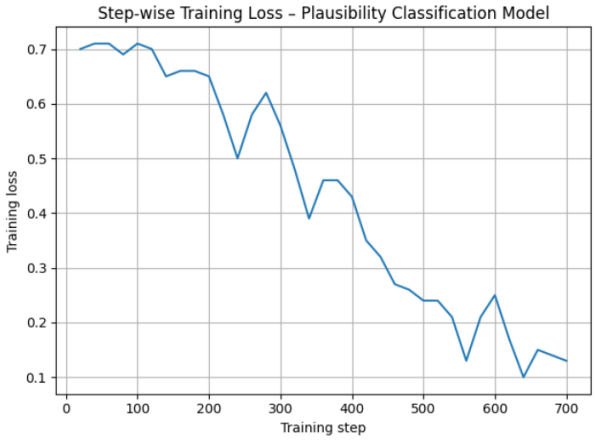

Figure 8: Step-wise Training Loss for the Plausibility Model

7 Results on Plausibility

Table 2 shows the F1 score and accuracy for plausibility prediction.

Model	F1	Accuracy
DeBERTa-v3-base (span-based)	77.76	77.78

Table 2: Plausibility results using the best model.

Qualitative analysis suggests that the model generates explanations that align with common scientific assumptions or domain-typical reasoning patterns, effectively bridging gaps in incomplete or underspecified scientific statements.

7.1 Prediction Analysis

To better understand the behavior of the plausibility model, we conducted a qualitative review of predicted spans on the development set. The model performs well when the input scientific statement contains at least one domain-relevant cue, even if expressed indirectly or with minimal context. In such cases, the span head is able to infer a canonical scientific explanation grounded in the underlying concept rather than relying solely on surface wording. For example, when a claim includes partial phrases such as "gives energy" or "because of sunlight," the model often produces scientifically coherent spans such as "plants use sunlight to make food" or "cellular respiration releases energy."

However, the model struggles when the input statement lacks any meaningful scientific anchor. For hypotheses requiring multi-step reasoning or implicit causal links, the model frequently defaults to high-frequency or broadly applicable scientific explanations (e.g., "to get energy" or "it moves because of force"). Although these predictions are scientifically plausible in a general sense, they may not align closely with the specific concept intended in the gold span.

Further inspection reveals that the model occasionally selects spans that are semantically valid but either too narrow or too broad relative to the target reference. These boundary-related inconsistencies highlight the inherent ambiguity in mapping short scientific statements to concise explanatory spans. They also suggest that integrating generative modeling components or question-conditioned prompting could further improve robustness in future work.

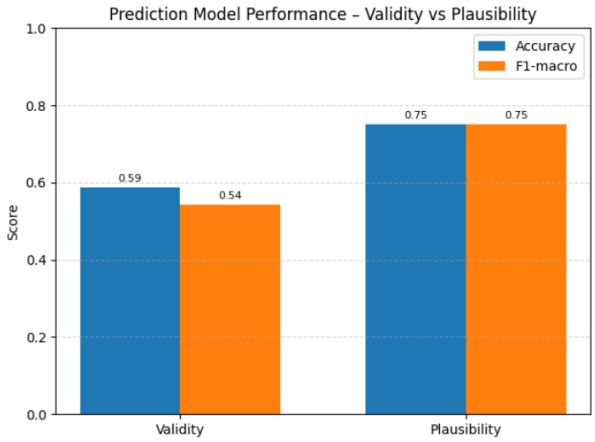

Figure 9: Prediction Model Performance

8 Error Analysis

To better understand the limitations of our system, we conducted a qualitative error analysis on both the validity and plausibility predictions. The majority of errors fall into three recurring categories, each reflecting a distinct linguistic or conceptual challenge present in short scientific statements.

8.1 Ambiguous Scientific Language

A substantial portion of misclassifications originates from statements that contain vague or underspecified expressions, such as "because of energy" or "it just happens." These claims lack explicit scientific detail, making it difficult for the model to determine whether any valid reasoning is present. In several cases, a statement may be conceptually sound in intent but expressed too implicitly or incompletely for the model to reliably interpret.

8.2 Overly Generic Predictions

For the plausibility task, the model occasionally defaults to high-frequency or broadly applicable scientific explanations (e.g., "to get energy" or "it needs sunlight"). Although such predictions are scientifically reasonable in general, they may not correspond to the specific concept implied by the input statement. This suggests that the span extraction head sometimes over-relies on distributional patterns learned during training rather than grounding predictions precisely in the input claim.

8.3 Boundary Errors in Span Extraction

The span-based architecture also exhibits classic extraction-related errors. These include selecting only part of the intended explanation (e.g., predicting "sunlight" instead of "plants use sunlight to make food"), shifting the predicted span slightly earlier or later than the intended region, or in rare cases producing terms not present in the input. These errors highlight sensitivity to token alignment and the inherent difficulty of mapping compact scientific statements to concise explanatory spans that are not strictly extractive.

9 Discussion

Our findings indicate that DeBERTa-v3-base is well suited for tasks requiring fine-grained semantic understanding of short scientific statements. Its disentangled attention mechanism appears to enhance the model's ability to identify key scientific relationships and distinguish valid reasoning from incorrect or incomplete assertions. The architecture's stability across domains further supports its use as a unified backbone for both validity classification and plausibility inference.

10 Limitations

- The training data is moderately sized; larger domain-specific corpora may improve generalization.
- The span-extraction formulation assumes that plausible explanations correspond to contiguous spans. Generative models (e.g., T5, Phi-3) may produce more natural explanations in cases requiring abstraction beyond the input.
- The model may learn stylistic or distributional biases present in scientific statements, potentially limiting robustness across specialized subdomains.

11 Conclusion

This work presented a comparative evaluation of four transformer-based architectures for SemEval-2026 Task 11, addressing the challenges of assessing scientific validity and reconstructing plausible scientific explanations from short, domain-specific statements. Our experiments demonstrated that DeBERTa-v3-base consistently outperforms the other models in both accuracy and training stability for the validity classification subtask. Building on this backbone, we introduced a span-prediction head for plausibility generation and observed strong performance despite the inherent ambiguity and variability of the task.

The results underscore the importance of fine-grained semantic modeling and robust contextual representations when working with compact scientific claims. Our analysis highlights the effectiveness of architectural innovations such as disentangled attention for capturing subtle scientific cues. Looking forward, incorporating generative reasoning, domain adaptation, or hybrid extractive–generative frameworks may further enhance performance in scenarios where plausible explanations require inference beyond the literal text. Overall, this study establishes a strong baseline and contributes to a deeper understanding of transformer behavior in scientific claim evaluation.

References

[1] He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention.

[2] Conneau, A. et al. (2019). Unsupervised cross-lingual representation learning at scale.

[3] He, P., Liu, X., Gao, J., & Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations (ICLR)*.

[4] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of EMNLP*.

[5] Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A., & Riedel, S. 2019. Language Models as Knowledge Bases? In *Proceedings of EMNLP*.

[6] Beltagy, I., Lo, K., & Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of EMNLP*.

[7] SemEval-2026 Task 11. 2026. Disentangling Content and Formal Reasoning in Scientific Statements. https://sites.google.com/view/semeval-2026-task-11