Paradise at SemEval-2026 Task 12: Leveraging Instruction-Tuned Large Language Models with Chain-of-Thought Prompting for Abductive Event Reasoning

Dhruv Goyal[1], **Ishita Gupta**[2], **Jatin Bedi**[3]

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala, India
[1]dhruv621999goyal@gmail.com, [2]igupta3_be23@thapar.edu, [3]jatin.bedi@thapar.edu

Abstract

We present **Paradise**, our system for SemEval-2026 Task 12: Abductive Event Reasoning, which identifies plausible direct causes of real-world English-language events using retrieved contextual documents. Our approach employs Qwen2.5-7B-Instruct, a 7-billion-parameter instruction-tuned language model combined with carefully engineered chain-of-thought prompting, requiring no task-specific fine-tuning or training-data supervision (prompt components were selected using the development set). The system achieves a score of **0.79** on the official 612-instance test set by integrating explicit causal-inference rules, 4,000-character document context windows, and greedy decoding. Analysis reveals that conservative prediction patterns—87.1% single-label and 36.9% Option D—effectively exploit the asymmetric scoring metric. Ablation studies confirm that document context contributes +6.4 points, chain-of-thought reasoning +5.3 points, and explicit causal rules +3.1 points to development performance. Our code is publicly available at https://github.com/DhruvGoyal404/semeval2026-task12.

1 Introduction

Abductive reasoning—inferring the most plausible cause of an observed outcome from incomplete evidence—remains fundamentally challenging for large language models (LLMs) despite their impressive performance on deductive and inductive tasks (Brown et al., 2020; Bhagavatula et al., 2020). Unlike deduction (applying rules to reach a certain conclusion) or induction (generalising rules from examples), abduction requires hypothesis generation under uncertainty: a system must evaluate competing explanations for an observed event without access to complete information, distinguishing genuine causes from mere correlates.

SemEval-2026 Task 12 (Cao et al., 2026) operationalises this challenge as Abductive Event Reasoning (AER) over English news documents. Given a real-world target event and four candidate explanations (A–D), systems must select all options that are *direct* causes. Option D always reads "None of the others are correct causes." The task is multi-label—multiple options may simultaneously be correct—and employs an asymmetric scoring metric that collapses any instance containing a false positive to zero, strongly incentivising precision over recall.

We present **Paradise**, which leverages training-data-free chain-of-thought (CoT) prompting with explicit causal-inference rules rather than task-specific fine-tuning. While prompt components were tuned on the development set, no model weights are updated at any stage. Our central insight is that modern instruction-tuned LLMs can perform competent abductive reasoning when provided with (i) structured causal rules embedded directly in the prompt, (ii) retrieved document grounding, and (iii) a constrained output format—all without gradient updates. Paradise achieves 0.79 on the test set and 0.4763 on the development set.

Key Contributions

- A five-component CoT prompt template with explicit causal rules (direct causation, temporal ordering, multi-label awareness) achieving 0.79 on 612 test instances.
- A training-data-free inference pipeline requiring no gradient updates, using only Qwen2.5-7B-Instruct at 4-bit NF4 quantisation (≈5.2 GB VRAM); prompt components were selected on the development set.
- Systematic ablation studies isolating the contribution of document context (+6.4 pts),

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 706–712

CoT reasoning (+5.3 pts), and causal rules (+3.1 pts).

- Analysis of how conservative prediction behaviour (87.1% single-label, 36.9% Option D) aligns with the asymmetric evaluation metric.

- Identification of a large development-to-test gap (Δ=0.31) and its likely causes.

2 Background

2.1 Task and Data

Each AER instance contains: (1) a `target_event` string describing the observed effect; (2) four options (`option_A` through `option_D`), where D is always the "none" option; (3) a `topic_id` linking to a set of retrieved English news documents in `docs.json`; and (4) a `golden_answer` (hidden on the test split). Documents are structured JSON objects with fields `title`, `content`, `snippet`, `source`, and `link`. The dataset spans three domains: politics (missile strikes, diplomatic conflicts), finance (market crashes, corporate bankruptcies), and public emergencies (natural disasters, health crises), and was constructed through LLM-assisted generation followed by human validation.

Official split sizes are: **train** 1,819 questions (57.3% single-label, 42.7% multi-label), **dev** 400 questions, **test** 612 questions. Each topic provides approximately 8.5 MB of retrieved documents. Paradise does *not* use the training split for model optimisation; it is used only for exploratory analysis of answer distributions (see Table 5).

2.2 Asymmetric Evaluation Metric

Let G denote the gold options and P the predicted options. Each instance is scored as:

$$\text{score}(P, G) = \begin{cases} 1.0 & \text{if } P = G \\ 0.5 & \text{if } PG \text{ and } P \neq \emptyset \\ 0.0 & \text{otherwise (any false positive)} \end{cases} \quad (1)$$

The final score is the mean across all instances. This design creates a critical asymmetry: a single false positive destroys the instance score entirely, whereas missing some correct options only reduces it to 0.5. Random single-option guessing yields an expected score of $\approx$0.25.

2.3 Related Work

Abductive NLI was pioneered by Bhagavatula et al. (2020), who revealed a 23-point gap between BERT (68.9%) and humans (91.4%) on the αNLI task. Chain-of-thought prompting (Wei et al., 2022b) demonstrated that eliciting step-by-step reasoning improves LLM performance on causal tasks; Kojima et al. (2022) showed that even zero-shot CoT ("Let's think step by step") lifts MultiArith accuracy from 17.7% to 78.7%. Instruction tuning (Wei et al., 2022a; Chung et al., 2022) enables strong zero-shot generalisation by training on thousands of task-formatted examples, making models like Qwen2.5 (Yang et al., 2024) suitable for complex reasoning without per-task fine-tuning. Wang et al. (2022) further showed that self-consistency voting over multiple reasoning paths improves reliability on abductive problems.

3 System Overview

3.1 Model and Infrastructure

We use **Qwen2.5-7B-Instruct** (Yang et al., 2024), a 7B-parameter instruction-tuned transformer (32 layers, 4096 hidden dimensions, 32,768-token context window) trained on over 1,000 diverse tasks. Qwen2.5-7B-Instruct was selected for three reasons: (i) it is publicly available without access restrictions; (ii) its instruction-tuning on diverse tasks produces strong zero-shot generalisation for structured reasoning; and (iii) its compatibility with 4-bit NF4 quantisation enables single-GPU inference at $\approx$5.2 GB VRAM, within reach of modest hardware budgets. We note that Qwen3 was released after our experimental runs were completed; evaluating it remains a natural direction for future work. To enable single-GPU inference we apply 4-bit NF4 quantisation via BitsAndBytes (Dettmers et al., 2022), reducing memory from $\approx$28 GB (FP16) to **5.20 GB** with double quantisation of quantisation constants and FP16 compute dtype. All experiments run on a single NVIDIA RTX PRO 6000 Blackwell (95 GB VRAM) under Python 3.12.12 and PyTorch 2.10.0+cu130. Development evaluation (400 instances) completes in $\approx$10.8 min (1.6 s/question); test inference (612 instances) in $\approx$19.2 min.

3.2 Chain-of-Thought Prompt Engineering

Our core contribution is a five-component prompt template:

(1) Role definition. "You are an expert at abductive reasoning—identifying the most plausible CAUSE of an event." This priming activates task-relevant reasoning patterns from instruction tuning.

(2) Task specification. An explicit instruction to analyse each option and determine which represent the most plausible *direct* cause, using the context documents as evidence.

(3) Explicit causal rules. Four constraints that encode fundamental principles of causality:

1. An option must describe an event that **directly caused** the target event (not merely correlated with it).
2. **Multiple options** may be correct simultaneously.
3. Option D is valid **only if** A, B, and C are all incorrect.
4. Causes must **temporally precede** their effects.

(4) Document context. Retrieved documents are formatted as `[Title]: content` blocks, concatenated and truncated at **4,000 characters**. This limit empirically balances context coverage against the model's effective attention span within the 4,096-token input budget. Although Qwen2.5-7B-Instruct supports a 32,768-token context window, prior work has shown that instruction-tuned models under 4-bit quantisation exhibit degraded attention at longer contexts (Liu et al., 2024), with performance plateauing or declining beyond the effective utilisation range. Our ablation confirms this: extending to 6,000 characters yields only +0.4 pts (Table 4), justifying the conservative 4,096-token budget.

(5) Structured output constraint. "Return your answer in EXACTLY this format: `ANSWER: A` or `ANSWER: A,C` or `ANSWER: D`." Strict formatting allows for proper regex extraction.

3.3 Inference Pipeline

Stage 1—Document processing. For each question, extract documents by `topic_id` from `docs.json`, format as title-content pairs, and concatenate in order of appearance in the source file (no relevance-based re-ranking is applied). The concatenated string is then hard-truncated at 4,000 characters.

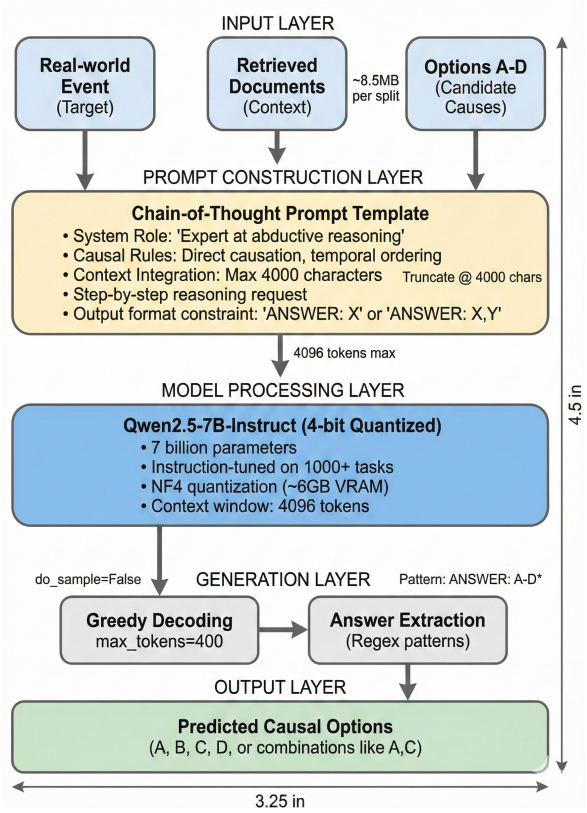

Figure 1: Complete inference pipeline: retrieved documents are truncated and injected into a five-component CoT prompt; Qwen2.5-7B-Instruct (4-bit NF4, 5.20 GB) generates a chain-of-thought response under greedy decoding; regex extraction yields the final multi-label prediction.

Stage 2—Prompt construction. Instantiate the template with the truncated context, `target_event`, and options A–D.

Stage 3—LLM generation. Apply the Qwen chat template (inserting `<|im_start|>` / `<|im_end|>` special tokens), tokenise with truncation at 4,096 tokens, and call:

```
model.generate(inputs,
    max_new_tokens=400,
    do_sample=False)
```

Greedy decoding (`do_sample=False`) ensures deterministic, reproducible outputs.

Stage 4—Answer extraction. A cascade of regex patterns is applied in priority order: (1) `ANSWER: [A-D](,[A-D])*`, (2) `Final Answer: [A-D](,[A-D])*`, (3) scan the final 200 characters for option letters, (4) default to `D` if nothing matched. Extracted labels are sorted and deduplicated (e.g., "C,A,A" → "A,C").

4 Experimental Setup

4.1 Implementation Details

The pipeline is implemented in Python 3.12.12 using HuggingFace Transformers 4.36.0 (model loading, tokenisation, generation)[1], BitsAndBytes 0.41.0 (4-bit quantisation)[2], and standard library modules for JSON parsing and regex extraction. No orchestration frameworks (e.g., LangChain, LlamaIndex) were used.

4.2 Data Splits

The training split (1,819 questions) is used exclusively for exploratory analysis of answer-label distributions (Table 5); no gradient updates or prompt hyperparameter tuning are performed on it. The **development split** (400 questions) drives all ablation experiments. The **test split** (612 questions, labels withheld) provides the official evaluation score.

4.3 Hyperparameters

Table 1 summarises the generation and quantisation configuration used for all runs.

Hyperparameter	Value
Max new tokens	400
Decoding	Greedy (`do_sample=False`)
Input truncation	4,096 tokens
Document char limit	4,000 characters
Quantisation	4-bit NF4, double quant
Compute dtype	FP16

Table 1: Generation and quantisation hyperparameters.

5 Results

5.1 Main Results

Paradise achieves **0.79** on the test set and 0.4763 on the development set (Table 2), placing **59th out of 221 participating systems**—in the top 26.7% of all submissions. Development breakdown: 127 full-match (31.8%), 127 partial-match (31.8%), 146 incorrect (36.5%); the system identifies at least one correct cause in 63.6% of development instances.

5.2 Prediction Distribution

The system exhibits strongly conservative behaviour (Table 3). Given that the asymmetric metric penalises any false positive with a score of 0.0, predicting fewer options is strategically sound: a correct single-label prediction on a multi-label instance still yields 0.5, whereas including one wrong option yields 0.0.

Split	Instances	Score
Development	400	0.4763
Test (official)	612	**0.79**

Table 2: Official evaluation results. Test score represents a 3.16× improvement over random baseline (≈0.25). $\Delta = 0.3137$ between splits.

Pattern	Count	%
Single-label predictions	533	87.1
Multi-label predictions	79	12.9
Option frequency (test)		
Option A	138	22.5
Option B	149	24.3
Option C	185	30.2
Option D (None)	226	36.9

Table 3: Test-set prediction distribution. Conservative behaviour (87.1% single-label, 36.9% Option D) aligns with the asymmetric metric penalising false positives. Option frequencies sum to 698 rather than 612 because multi-label predictions contribute one count per predicted option.

5.3 Ablation Studies

We conducted six ablation experiments, each removing one component from the full pipeline (Table 4).

Document context is the single most important component (+6.4 pts). Without retrieved documents, the model must rely entirely on parametric knowledge, which is insufficient for domain-specific causal attribution over recent real-world events. We initially experimented with 2,000-character truncation (score 0.4550), which proved insufficient; 4,000 characters provided the best cost–benefit trade-off, as 6,000 characters yielded only marginal gains (+0.4 pts) at higher token cost.

Chain-of-thought reasoning is the second most important component (+5.3 pts). Removing the step-by-step reasoning instruction degrades performance substantially, confirming that causal inference benefits significantly from explicit intermediate reasoning steps. We initially tested direct prompting (no CoT) and observed the model frequently selected the superficially closest option rather than reasoning about causal structure.

Explicit causal rules contribute +3.1 pts.

[1] https://github.com/huggingface/transformers

[2] https://github.com/TimDettmers/bitsandbytes

Configuration	Dev Score	
Full system	**0.4763**	
Document context		
No documents	0.4125	(−6.4 pts)
2,000-char limit	0.4550	(−2.1 pts)
6,000-char limit	0.4801	(+0.4 pts)
Prompt components		
Without CoT reasoning	0.4238	(−5.3 pts)
Without causal rules	0.4456	(−3.1 pts)
Without step-by-step	0.4601	(−1.6 pts)

Table 4: Ablation results on the 400-instance development set. Each row removes exactly one component from the full system.

Label pattern	Count (train)	%
Single-label	1,042	57.3
Multi-label	777	42.7
Option frequency (train)		
A	752	41.3
B	712	39.1
C	718	39.5
D	671	36.9

Table 5: Answer distribution in the training set (used for analysis only; no model training occurs on this split). Multi-label prevalence (42.7%) is much higher than our system's 12.9% multi-label prediction rate on the test set.

Without the four rules, the model frequently confused correlated events with causal ones and violated temporal ordering (selecting events that occurred *after* the target event). Adding each rule incrementally guided the model toward correct causal attribution. We acknowledge that these four rules encode only elementary causal principles; more nuanced phenomena such as necessary versus sufficient causation, confounders, and multi-hop causal chains are not explicitly addressed and represent a known limitation of the current prompt design.

5.4 Training Data Distribution Analysis

Table 5 reveals that 42.7% of training instances are multi-label, significantly higher than our system's 12.9% multi-label prediction rate on test. This gap suggests our conservative prompt strategy substantially under-predicts multi-label cases, sacrificing recall to avoid the asymmetric false-positive penalty.

6 Discussion

6.1 Development-to-Test Performance Gap

The 0.31-point gap between development (0.4763) and test (0.79) is the most notable finding. We hypothesise three complementary explanations. *Difficulty variation*: the test set may contain fewer genuinely multi-causal events, making single-label conservative predictions more often exactly correct—our system predicts 87.1% single-label on test versus 63.5% on development (computed from Table 3 and development breakdown in Section 5), consistent with the test set rewarding conservative behaviour more often. *Document quality*: test-set retrieved documents may carry cleaner causal signals, reducing document-grounding failures. *Distribution shift*: test causal patterns may align more closely with the model's instruction-tuning distribution. Additionally, the gap may be domain-concentrated rather than uniform across politics, finance, and public-emergency topics; per-domain analysis requires gold labels not available at submission time. Understanding these factors fully remains an open empirical question and a primary direction for follow-up work.

6.2 Error Analysis

Manual analysis of the 146 development set errors shows four patterns of failure:

Temporal confusion (15% of errors): The model fails to follow Rule 4, choosing events that occur after the target event. This could be remedied by explicit temporal parsers or date-extraction components.

Correlation-causation confusion (22%): The model chooses events that are correlated with the target event but have a common third-party cause. For instance, the model predicts "stock market decline" caused "company bankruptcy" when both result from a third factor.

Multi-label under-prediction (31%): the largest error category. The model identifies one correct cause but misses additional valid causes, scoring 0.5 instead of 1.0. This is the primary source of performance headroom.

Document grounding failures (18%): the correct cause appears explicitly in the retrieved documents but is not surfaced by the model, pointing to limitations in long-context utilisation. **Other / uncategorised (14%):** residual errors that do not fall cleanly into the above categories, including cases where the model produces a valid causal chain but

selects a distal rather than proximate cause.

7 Conclusion

We presented Paradise, a training-data-free chain-of-thought system for SemEval-2026 Task 12 Abductive Event Reasoning, achieving 0.79 on the official test set without any task-specific fine-tuning. Our ablation studies demonstrate that retrieved document context (+6.4 pts), chain-of-thought reasoning (+5.3 pts), and explicit causal rules (+3.1 pts) are all essential components. Conservative prediction behaviour—motivated by the asymmetric scoring metric—yields a strategically sound single-label bias, though the 42.7% multi-label prevalence in training data suggests that improved multi-label prediction is the clearest path to further gains.

Promising future directions include: (1) task-specific fine-tuning on the provided training data; (2) self-consistency voting (Wang et al., 2022) over multiple CoT paths to reduce variance; (3) cross-encoder re-ranking of retrieved documents per option; and (4) explicit temporal extraction to eliminate temporal-confusion errors; and (5) per-domain analysis of the development-to-test performance gap once test-set gold labels are released, to determine whether the gap is uniform or concentrated in specific domains (politics, finance, or public emergencies).

Limitations

Computational cost. Despite 4-bit quantisation, inference requires a GPU; our setup used an NVIDIA RTX PRO 6000 Blackwell (95 GB VRAM), which exceeds typical academic budgets. Smaller quantised models (e.g., 3B-parameter variants) may reduce requirements.

Multi-label under-prediction. Our conservative prompting biases toward single-option predictions (87.1%), substantially below the training-set multi-label prevalence of 42.7%, sacrificing recall.

Temporal reasoning. The system relies on the LLM's implicit temporal understanding rather than explicit date extraction; this produces temporal-confusion errors in 15% of failure cases.

System-level contribution. Paradise is primarily an empirical and engineering contribution: it demonstrates that careful prompt design alone—without architectural novelty or task-specific training—can achieve competitive performance on a structured causal reasoning task. Developing novel architectural components for abductive reasoning remains an important direction for future work.

Generalisability. Evaluation is limited to the AER dataset; performance on other abductive reasoning benchmarks (e.g., αNLI, COPA) is unknown.

Ethical considerations. The AER dataset is drawn from real-world news events spanning sensitive domains (political conflicts, financial crises, public emergencies). We use the data solely for academic evaluation; no personal data is collected or processed. We foresee no direct misuse risks beyond those inherent to general-purpose LLM deployment.

Acknowledgments

We thank the SemEval-2026 Task 12 organizers—Pengfei Cao, Yubo Chen, Mingxuan Yang, Chenlong Zhang, Mingxuan Liu, Kang Liu, and Jun Zhao—for the challenging benchmark. We acknowledge the Qwen team at Alibaba Cloud for releasing Qwen2.5-7B-Instruct publicly.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Pengfei Cao, Yubo Chen, Mingxuan Yang, Chenlong Zhang, Mingxuan Liu, Kang Liu, and Jun Zhao. 2026. SemEval-2026 task 12: Abductive event reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, Mexico City, Mexico. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in*

Neural Information Processing Systems, volume 35, pages 30318–30332.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A Development Set Score Distribution

Score	Count	%
Full match (1.0)	127	31.8
Partial match (0.5)	127	31.8
Incorrect (0.0)	146	36.5
Total	**400**	**100.0**

Table 6: Score distribution on the 400-instance development set.

B Prompt Template

The complete prompt template used for all inference runs:

```
You are an expert at abductive reasoning
-- identifying the most plausible CAUSE
of an event.

## Task
Given an EVENT, determine which option(s)
represent the most plausible
direct CAUSE. Use CONTEXT documents.

## Rules
1. Option must DIRECTLY CAUSE the target
   event (not merely correlated)
2. Multiple options may be correct
3. Option D is correct ONLY if A, B, C
   are all wrong
4. Causes must occur BEFORE the effect

## Context Documents
{context}

## Event (Effect)
{target_event}

## Options
A: {option_A}
B: {option_B}
C: {option_C}
D: {option_D}

## Reasoning
Think step by step, then give your
final answer.

Return answer in EXACTLY this format:
ANSWER: A
or ANSWER: A,C
or ANSWER: D
```