# LOGICAL-COMMONSENSEQA: A Benchmark for Logical Commonsense Reasoning

**Anonymous ACL submission**

## Abstract

Commonsense reasoning often involves evaluating multiple plausible interpretations rather than selecting a single atomic answer, yet most benchmarks rely on single-label evaluation, obscuring whether statements are jointly plausible, mutually exclusive, or jointly implausible. We introduce LOGICAL-COMMONSENSEQA, a benchmark that reframes commonsense reasoning as *logical composition* over pairs of atomic statements using plausibility-level operators (AND, OR, NEITHER/NOR). Evaluating instruction-tuned, reasoning-specialized, and fine-tuned models under zero-shot, few-shot, and chain-of-thought prompting, we find that while models perform reasonably on conjunctive and moderately on disjunctive reasoning, performance degrades sharply on negation-based questions. LOGICAL-COMMONSENSEQA exposes fundamental reasoning limitations and provides a controlled framework for advancing compositional commonsense reasoning.

## 1 Introduction

Commonsense reasoning is central to human cognition and a long-standing challenge in artificial intelligence and natural language understanding. Developmental studies suggest it emerges from intuitive, experience-based frameworks in early childhood and becomes more abstract and socially grounded through embodied experience, social interaction, cultural transmission, and language (Spelke et al., 1992; Spelke and Kinzler, 2007; Gopnik, 2004; Meltzoff et al., 2009). In contrast, large language models (LLMs) acquire "commonsense-like" knowledge indirectly from large-scale text corpora via objectives such as next-token prediction, rather than through grounded interaction or social learning (Bender and Koller, 2020). Despite this, LLMs exhibit emergent behaviors resembling human intuitive reasoning, including inferring intentions and anticipating socially plausible actions (Sap et al., 2019a; Bisk et al., 2020). However, this competence is brittle, as models fail under minimal semantic perturbations, struggle with logical consistency, and rely on statistical shortcuts rather than genuine reasoning (Ettinger, 2020; Geiger et al., 2020; McCoy et al., 2019; Niven and Kao, 2019). To improve the evaluation of commonsense reasoning, we introduce LOGICAL-COMMONSENSEQA, a benchmark that reframes the task as a *logical composition over multiple plausible answers*, requiring models to reason about relationships between plausible statements rather than select a single response.

Early approaches to commonsense reasoning relied on symbolic knowledge bases (Lenat et al., 1990; Liu and Singh, 2004), which were expressive but limited in scalability and coverage. More recent benchmarks shift the evaluation towards the ability of neural models to infer plausible answers from text (Talmor et al., 2019; Sap et al., 2019b,a; Zellers et al., 2019; Bisk et al., 2020; Sakaguchi et al., 2021). However, these benchmarks reduce commonsense reasoning to single-answer prediction, despite the fact that many real-world situations admit multiple plausible interpretations (Min et al., 2020). For example, in the question *"The fox walked from the city into the forest—what was it looking for?"* both *"natural habitat"* and *"shelter from disturbances"* are plausible answers. Restricting the evaluation to a single atomic choice removes this natural ambiguity and obscures whether models recognize jointly plausible, mutually exclusive, or jointly implausible interpretations.

Beyond linguistic ambiguity, commonsense is both socially grounded and compositional. People judge commonsense statements not only by personal belief but also by estimating whether others would share that belief (Whiting and Watts, 2024), implying that commonsense reflects a distribution over plausible interpretations rather than a single canonical truth. At the same time, human reasoning integrates multiple alternatives through composi-

tional inference, recognizing when statements are jointly plausible (AND), partially plausible (OR), or jointly implausible (NEITHER/NOR). These relational judgments are central to everyday reasoning about intentions, plans, and social norms, yet current benchmarks fail to evaluate whether models can reason over such socially grounded and compositional plausibility relationships.

In LOGICAL-COMMONSENSEQA, each instance pairs two atomic answers under one of three symbolic relations—AND, OR, or NEITHER/NOR. These relations are not strict propositional operators, but *plausibility-level composition constructs* indicating joint, partial, or joint implausibility. This formulation preserves the multiple-choice question (MCQ) format of modern benchmarks while explicitly modeling ambiguity and compositional reasoning. We evaluate instruction-tuned, reasoning-specialized, and fine-tuned LLMs under zero-shot, few-shot, and chain-of-thought prompting. Results show that while models perform reasonably well on conjunctive and moderately on disjunctive reasoning, performance degrades sharply on negation-based compositions, revealing persistent gaps in the commonsense reasoning capabilities of state-of-the-art systems.

## 2 Related Work

Commonsense reasoning has been widely studied through symbolic knowledge bases, neural models, and benchmark-driven evaluation. Early symbolic approaches such as CYC (Lenat et al., 1990) and CONCEPTNET (Liu and Singh, 2004) enabled structured inference but suffered from limited coverage and scalability. More recent benchmarks, including COMMONSENSEQA (Talmor et al., 2019), SOCIALIQA (Sap et al., 2019b), ATOMIC (Sap et al., 2019a), PIQA (Bisk et al., 2020), HELLASWAG (Zellers et al., 2019), and WINOGRANDE (Sakaguchi et al., 2021), shift evaluation towards the ability of neural models to infer plausible responses from text. However, these benchmarks adopt a single atomic-answer format that fails to capture the inherent ambiguity of many real-world commonsense questions.

Prior work has emphasized that ambiguity and human disagreement are intrinsic to commonsense reasoning. AMBIGQA (Min et al., 2020) and PROTOQA (Boratko et al., 2020) show that many questions naturally admit multiple valid answers, motivating the need to evaluate systems beyond single gold answers. Separately, logical reasoning benchmarks such as LOGIQA (Liu et al., 2020), RECLOR (Yu et al., 2020), and COM2 (Fang et al., 2024) evaluate deductive inference using formal logical operators, but target logical validity rather than commonsense plausibility. Complementary work in the social sciences argues that commonsense is socially grounded, reflecting shared expectations rather than objective truth (Whiting and Watts, 2024).

Our work differs from these lines of research by introducing LOGICAL-COMMONSENSEQA, a benchmark that evaluates commonsense reasoning as *compositional plausibility*. Instead of open-ended multi-answer annotation or formal deductive logic, LOGICAL-COMMONSENSEQA preserves the MCQ format of modern commonsense benchmarks while explicitly encoding joint plausibility, partial plausibility, and joint implausibility through symbolic composition over human-validated atomic answers.

## 3 The Logical Commonsense Dataset

We construct LOGICAL-COMMONSENSEQA to evaluate compositional, multi-answer commonsense reasoning by extending COMMONSENSEQA (Talmor et al., 2019), a benchmark of 12,247 MCQs with one correct answer and four distractors that probes commonsense relations such as cause–effect, purpose, and spatial association. LOGICAL-COMMONSENSEQA reformulates these single-answer questions into logically composed items that explicitly model relationships among plausible alternatives while preserving full compatibility with the MCQ format.

**Construction Pipeline.** The dataset is built using a three-stage pipeline that integrates neural generation with deterministic symbolic composition, converting atomic answers into logically structured multi-answer instances.

*Stage 1: Generation of Candidate Options.* Starting from COMMONSENSEQA questions and their gold answers, we sample 5,000 instances and use GPT-4o-mini to over-generate a diverse set of atomic answer candidates. The model is prompted to produce both plausible and implausible alternatives, with an emphasis on multi-step causal or situational reasoning rather than shallow lexical cues. This process yields 4-6 plausible and implausible candidate answers per question, spanning physical, social, and situational commonsense domains.

| Operator | Interpretation | Example |
|---|---|---|
| $a$ AND $b$ | Both statements independently plausible | *Q: Sammy wanted to go to where the people were. Where might he go?*<br>A: local events AND social venues |
| $a$ OR $b$ | At least one statement plausible | *Q: Sammy wanted to go to where the people were. Where might he go?*<br>A: local events OR empty parks |
| NEITHER $a$ NOR $b$ | Neither statement plausible | *Q: Sammy wanted to go to where the people were. Where might he go?*<br>A: NEITHER quiet retreats NOR empty parks |

Table 1: Composition operators used in LOGICAL-COMMONSENSEQA to express plausibility relations.

*Stage 2: Refinement and Pruning.* We then use GPT-4o-mini to refine and filter the generated candidates according to three criteria: (1) removing logically inconsistent or factually incorrect answers; (2) eliminating trivial options that can be resolved through keyword matching; and (3) identifying highly plausible options that satisfy most semantic and contextual constraints of the question but fail due to a non-obvious commonsense violation. This stage yields a curated set of three correct and four incorrect atomic options per question, each requiring multi-step inference.

*Stage 3: Deterministic Logical Composition.* In the final stage, we use a symbolic program to deterministically combine pairs of refined atomic options into logical compositions labeled with one of the three relations shown in Table 1. These relations are not strict propositional operators; instead, they encode commonsense judgments about joint plausibility. This procedure yields 14,997 instances in total, with 4,999 instances per base relation type.

The prompts for all stages of the construction pipeline are provided in Appendix A.3.

**Human Validation** To ensure the commonsense validity of the refined atomic options, we conducted a human validation study using the awareness–consensus framework proposed by Whiting and Watts (2024). Two independent annotators evaluated a random sample of 250 questions (5% of Stage 2). For each atomic option, annotators assessed (1) whether they personally believed the option was correct (awareness) and (2) whether they believed most others would agree (consensus). This dual assessment captures both individual plausibility and perceived social agreement, aligning with recent views of commonsense as socially grounded. Disagreements between annotators were resolved through adjudication.

We report label accuracy on the adjudicated subset covering 50% of the question-level test split defined after Stage 2. Performance is high for AND (89.2%), OR (96.4%), and MIXED (88.4%), and reasonable for NEITHER/NOR (73.6%). These numbers reflect a reasonably accurate labeling of composed instances.

Inter-annotator agreement, measured using Cohen's $\kappa$, indicated moderate agreement on human judgments ($\kappa = 0.49$) before adjudication, suggesting that the refined options are both semantically coherent and socially interpretable.

**Resulting Dataset** After Stage 2, we obtain 4,999 curated questions with refined atomic answer sets. In Stage 3, each question is expanded into logically structured instances evenly distributed across the three base relation types, with an additional MIXED setting in which operators are randomly assigned across options. After filtering, the final dataset contains 19,996 instances, with AND, OR, NEITHER/NOR, and MIXED each contributing 4,999 instances.

## 4 Task Formulation

Instances in LOGICAL-COMMONSENSEQA consist of a question in natural language and four options. Unlike standard benchmarks where each option is an atomic answer, each option in our dataset is a *logical composition* of two independent atomic statements combined using one of three plausibility-level operators: AND, OR, or NEITHER/NOR (see Table 1).

Given a question, the model must select the option whose composite plausibility best matches the implied commonsense constraints. While the task remains a simple MCQ problem, it requires models to assess the plausibility of individual statements and reason about their interaction under the specified operator. We include both operator-specific conditions, in which all answer options for a question share the same logical operator, and a MIXED condition, in which different operators appear across options. While the operator-specific conditions allow for controlled evaluation of each relation type, the MIXED condition prevents models from exploiting operator-specific patterns and

requires direct inference over the composed statements. Overall, this formulation preserves the MCQ format while substantially increasing the demands on compositional commonsense reasoning.

## 5 Experiments

We evaluate a diverse set of encoder-only classifiers, encoder–decoder models, and decoder-only LLMs on LOGICAL-COMMONSENSEQA under zero-shot, few-shot, and supervised fine-tuning regimes. Full details on model selection, prompting strategies, and training and inference procedures are provided in Appendices A.1, A.2, and A.5.

Following the dataset construction described in Section 3, we use all 19,996 instances of LOGICAL-COMMONSENSEQA split into 11,996 training, 6,000 development, and 2,000 test examples. Splits are stratified by logical relation, ensuring an even 25% distribution across AND, OR, NEITHER/NOR, and MIXED conditions. We report accuracy and macro $F_1$ as evaluation metrics.

For analysis, we further divide the test set into a human-validated (HV) subset and a non-validated (NV) subset, each comprising 50% of the test data. Table 2 summarizes the main results. Across all model families, two consistent patterns emerge: (1) all models perform reasonably well on AND and moderately on OR, and (2) performance collapses on NEITHER/NOR in zero-and few-shot settings, even for the strongest LLMs. Observed weaknesses are amplified in the MIXED condition, where different operators appear across options and models must implicitly infer the governing relation. F1 drops to 43–56% for all few-shot models, suggesting that these models do not reliably maintain operator-level representations and instead revert to surface-level heuristics. In contrast, fine-tuned models achieve 83–93% F1 across all operators, suggesting that the task is learnable with supervision and that zero- and few-shot failures, particularly on negation, reflect inference-time limitations rather than dataset artifacts.

Additional results—including varying numbers of shots, prompting strategies, and performance on the NV split—are reported in Appendix A.6. A comprehensive error analysis is also provided in Appendix A.7.

*Comparison with CommonsenseQA.* Finally, Table 3 underscores the central motivation for our benchmark. While LLaMA-3.1-8B achieves 72.2% accuracy on COMMONSENSEQA, its performance drops sharply on LOGICAL-COMMONSENSEQA: 72% on AND, 62.2% on OR, 42.7% on MIXED, and only 13.9% on NEITHER/NOR. This discrepancy shows that benchmarks like COMMONSENSEQA substantially overestimate models' commonsense reasoning ability by failing to test relational and compositional plausibility judgments.

| Paradigm | Model | AND | OR | NN | MIX |
|---|---|---|---|---|---|
| 0-Shot | LLaMA-3.3-70B | 80.9 | 70.9 | 13.4 | 53.0 |
| | LLaMA-3.1-8B | 71.9 | 62.2 | 13.1 | 41.8 |
| | Qwen2.5-7B | 79.6 | 68.9 | 12.9 | 53.2 |
| 3-Shot | LLaMA-3.3-70B | 85.7 | 77.5 | 10.7 | 50.4 |
| | LLaMA-3.1-8B | 71.3 | 61.3 | 6.1 | 42.8 |
| | Qwen2.5-7B | 77.9 | 71.9 | 6.8 | 51.4 |
| Fine-tuned | Flan-T5-base | 92.8 | 92.4 | 89.2 | 89.6 |
| | DeBERTa-v3-base | 87.6 | 87.2 | 84.8 | 82.4 |

Table 2: **Macro-F1 on the human-validated test set** across plausibility relations. NN = NEITHER/NOR, MIX = MIXED

| Dataset | Acc | F1 |
|---|---|---|
| **CommonsenseQA** | | |
| LLaMA-3.1-8B | 72.2 | 72.2 |
| **LOGICAL-COMMONSENSEQA** | | |
| AND | 72.0 | 71.9 |
| OR | 62.2 | 62.2 |
| NEITHER/NOR | 13.9 | 13.1 |
| MIXED | 42.7 | 41.8 |

Table 3: 0-Shot performance of LLaMA-3.1-8B on the original COMMONSENSEQA versus each logical operator in LOGICAL-COMMONSENSEQA.

## 6 Conclusions and Future Work

We introduce LOGICAL-COMMONSENSEQA, a benchmark that reframes commonsense reasoning as selecting among logically composed answer options. By combining human-validated atomic statements with symbolic operators that encode conjunction, disjunction, and negation of plausibility, our work exposes reasoning failures obscured by traditional single-answer benchmarks. Experiments reveal that while LLMs handle conjunctive and, to a lesser extent, disjunctive reasoning, they fail sharply on negation-based composition, highlighting persistent gaps in plausibility evaluation and operator grounding. Future work includes extending our resource to generative settings and structured reasoning frameworks, expanding the operator set, scaling awareness-consensus validation, and studying transfer to open-ended QA, dialog, and planning tasks.

## Limitations

LOGICAL-COMMONSENSEQA offers a controlled framework for probing compositional commonsense reasoning, but several limitations remain. The logical operators (AND, OR and NEITHER/NOR) capture plausibility-level relations rather than full propositional semantics, and do not cover richer structures such as implication, exclusivity or temporal and causal reasoning. Extending the operator space would enable a more complete evaluation of symbolic and relational inference.

Model coverage is another limitation. Decoder-only LLMs are evaluated only via prompting, while smaller encoder-based models receive supervised training. A broader comparison across instruction-tuned and specialized architectures may clarify which components most improve logical composition.

## Ethical Considerations

LOGICAL-COMMONSENSEQA is derived from the publicly available COMMONSENSEQA dataset and consists of generic, hypothetical scenarios that do not reference identifiable individuals or personal contexts. To the best of our knowledge, the dataset does not contain private or sensitive information.

We used an LLM to initially generate atomic answer options and increase coverage and diversity. These candidates are subsequently refined and partially validated by human annotators using an awareness–consensus framework. As a result, the dataset reflects socially grounded plausibility judgments rather than objective ground truth, and may thus inherit cultural or contextual biases.

## References


Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. Protoqa: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Tianqing Fang, Zeming Chen, Yangqiu Song, and Antoine Bosselut. 2024. Complex reasoning over logical queries on commonsense knowledge graphs. *arXiv preprint arXiv:2403.07398*.

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

Alison Gopnik. 2004. Finding our inner scientist. *Daedalus*, 133(1):21–28.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Douglas B Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. 1990. Cyc: toward programs with common sense. *Communications of the ACM*, 33(8):30–49.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

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.

Andrew N Meltzoff, Patricia K Kuhl, Javier Movellan, and Terrence J Sejnowski. 2009. Foundations for a new science of learning. *science*, 325(5938):284–288.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. In *Conference on Empirical Methods in Natural Language Processing*.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. 1992. Origins of knowledge. *Psychological review*, 99(4):605.

Elizabeth S Spelke and Katherine D Kinzler. 2007. Core knowledge. *Developmental science*, 10(1):89–96.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158.

Mark E Whiting and Duncan J Watts. 2024. A framework for quantifying individual and collective common sense. *Proceedings of the National Academy of Sciences*, 121(4):e2309535121.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

# A Appendix

## A.1 Model Selection

This section describes the architectures and evaluation settings of all models considered in our experiments, covering encoder-only, encoder-decoder, and decoder-only LLMs.

**Encoder Models** We fine-tune DeBERTa-v3-base (He et al., 2021), a pretrained encoder which encodes the concatenation of the question and one composite option and applies a classification head to score the four candidate answers.

**Encoder–Decoder and Multi-Step Reasoning Models** We also fine-tune sequence-to-sequence models to generate the output label. Flan-T5-base (Chung et al., 2024) is fine-tuned to output a single token corresponding to the chosen option. Additionally, we evaluate Entailer-11B (Tafjord et al., 2022), a model designed to produce faithful chains-of-reasoning by recursively generating premises through backward-chaining and verifying them via self-querying. Although originally developed for faithful explanation generation, we use Entailer as a strong encoder-decoder baseline to examine whether reasoning-oriented models better capture the compositional structures in our dataset.

**Decoder-Only LLMs** We evaluate several open-weight instruction-tuned LLMs in zero- and few-shot settings: Llama-3.3-70B-Instruct, Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-7B-Instruct (Qwen et al., 2025). These models are not fine-tuned; instead, they are prompted to output a single capital letter corresponding to the predicted option. All models are

evaluated on each logical subset defined in Table 1 as well as on the Mixed setting.

### A.2 Prompting Setup

In this section, we describe the prompting strategy used for decoder-only models. For decoder-only LLMs, we adopt a unified multiple-choice prompting scheme. Each prompt presents the question, the four composite options and an instruction to answer with a single letter (A-D). We evaluate 0-, 1-, 2- and 3-shot settings using a fixed set of labeled examples.

Decoding uses `temperature` $= 0.0$, `top-p` $= 0.9$, and `max_new_tokens`$= 3$. Outputs not corresponding to a valid answer option are treated as incorrect.

**Chain-of-Thought (CoT)** We evaluate CoT prompting for LLAMA-3.1-8B-INSTRUCT only, due to computational cost. CoT runs use a larger generation budget (`max_new_tokens` $=$ 200-300).

### A.3 Prompt Templates

This section provides the exact prompt templates used in our experiments. Prompts differ only in the number of in-context examples and the logical operator expressed in the answer options.

**Zero-shot Prompt** In the zero-shot setting, we don't provide any labeled examples.

> Answer the following commonsense question by selecting the correct option. Respond with **only** the single capital letter (A, B, C, or D).
>
> Question: <QUESTION>
>
> A. <OPTION A>
> B. <OPTION B>
> C. <OPTION C>
> D. <OPTION D>
>
> Answer:

**Few-shot Prompting** For few-shot (1-, 2-, and 3-shot) evaluation, we prepend labeled examples before the target question. **The 1- and 2-shot prompts are strict prefixes of the 3-shot prompt**, using the first one or two examples respectively.

For operator-specific subsets (AND, OR, NEITHER/NOR), all in-context examples match the operator being evaluated. For the MIXED setting, examples are sampled across operators.

**Conjunctive Reasoning (AND) Prompt**

> Question: Sammy wanted to go to where the people were. Where might he go?
>
> A. social venues AND quiet retreats
> B. local events AND social venues
> C. sports arenas AND quiet retreats
> D. sports arenas AND train platforms
>
> Answer: B
>
> Question: The fox walked from the city into the forest, what was it looking for?
>
> A. suitable prey AND shelter from disturbances
> B. urban garden AND farm fields
> C. natural habitat AND farm fields
> D. suitable prey AND neighborhood pets
>
> Answer: A
>
> Question: What home entertainment equipment requires cable?
>
> A. wireless speaker AND portable projector
> B. television AND home theater system
> C. wireless speaker AND gaming console
> D. digital frame AND portable projector
>
> Answer: B
>
> Now answer the following commonsense question by selecting the correct option. Respond with **only** the single capital letter (A, B, C, or D).
>
> Question: <TARGET AND QUESTION>
>
> A. <OPTION A>
> B. <OPTION B>
> C. <OPTION C>
> D. <OPTION D>
>
> Answer:

**Disjunctive Reasoning (OR) Prompt**

> Question: Sammy wanted to go to where the people were. Where might he go?
>
> A. quiet retreats OR empty parks
> B. local events OR empty parks
> C. sports arenas OR empty parks
> D. train platforms OR empty parks
>
> Answer: B
>
> Question: The forgotten leftovers had gotten quite old; he found it covered in mold in the back of his what?
>
> A. living room OR dining area
> B. utility drawer OR kitchen counter
> C. living room OR utility drawer
> D. cold storage OR utility drawer
>
> Answer: D
>
> Question: What do people use to absorb extra ink from a fountain pen?
>
> A. absorbent paper OR notebook cover
> B. ink storage OR writing surface
> C. notebook cover OR writing surface
> D. ink storage OR fabric swatch

Answer: A

Now answer the following commonsense question by selecting the correct option. Respond with **only** the single capital letter (A, B, C, or D).

Question: <TARGET OR QUESTION>

A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>

Answer:

### Negation-Based Reasoning (NEITHER/NOR) Prompt

Question: Sammy wanted to go to where the people were. Where might he go?

A. NEITHER local events NOR train platforms
B. NEITHER sports arenas NOR train platforms
C. NEITHER social venues NOR sports arenas
D. NEITHER social venues NOR quiet retreats

Answer: B

Question: The only baggage the woman checked was a drawstring bag. Where was she heading with it?

A. NEITHER family gathering NOR local gathering
B. NEITHER airport travel NOR local gathering
C. NEITHER local gathering NOR short trip
D. NEITHER business engagement NOR local gathering

Answer: C

Question: The forgotten leftovers had gotten quite old, he found it covered in mold in the back of his what?

A. NEITHER cold storage NOR dining area
B. NEITHER sealed container NOR living room
C. NEITHER food cabinet NOR utility drawer
D. NEITHER living room NOR utility drawer

Answer: D

Now answer the following commonsense question by selecting the correct option. Respond with **only** the single capital letter (A, B, C, or D).

Question: <TARGET NEITHER/NOR QUESTION>

A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>

Answer:

### Mixed Prompt

Question: The fox walked from the city into the forest, what was it looking for?

A. natural habitat AND shelter from disturbances
B. farm fields OR neighborhood pets
C. city park AND neighborhood pets
D. farm fields AND neighborhood pets

Answer: A

Question: Google Maps and other highway and street GPS services have replaced what?

A. historical navigation logs AND tourist information centers
B. NEITHER printed road maps NOR route planning software
C. manual navigation techniques AND tourist information centers
D. manual navigation techniques AND route planning software

Answer: D

Question: Where is a business restaurant likely to be located?

A. corporate office clusters AND at university campuses
B. business park entrances AND near conference venues
C. NEITHER near conference venues NOR at university campuses
D. NEITHER near conference venues NOR near tourist sites

Answer: B

Now answer the following commonsense question by selecting the correct option. Respond with **only** the single capital letter (A, B, C, or D).

Question: <TARGET MIXED QUESTION>

A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>

Answer:

**Chain-of-Thought Prompting** For CoT evaluation, we add a brief instruction asking the model to reason about the plausibility of each atomic statement and how the logical operator applies before providing the final answer.

> Carefully reason about each atomic statement and the logical operator. Then provide the final answer as a single capital letter (A–D).

**Decoding Constraints** All decoder-only models are evaluated with near-deterministic decoding (`temperature`=0.0, `top-p`=0.9). Outputs are constrained to a single capital letter (A–D); any other output is treated as incorrect.

### A.4 Annotation Guidelines

This section describes the guidelines provided to human annotators for validating refined atomic answer options in Stage 2 of dataset construction.

**Annotation Task** Each annotation item consists of a commonsense question and a set of automatically labeled answer options. Options are pre-labeled as *correct* or *incorrect* by the refinement pipeline. Annotators are instructed to evaluate each option *independently*, without reconsidering or changing the question itself.

The goal of annotation is only *verification*: annotators judge whether the assigned labels are reasonable under commonsense reasoning.

**Evaluation Criteria** For each option, annotators provide judgments along two dimensions, following the awareness–consensus framework of Whiting and Watts (2024):

1. **Correctness (Awareness).** Annotators indicate whether they personally believe the option is a valid answer to the question.
    - For options labeled as *correct*: "Do you believe this option is actually a valid answer?"
    - For options labeled as *incorrect*: "Do you believe this option is actually wrong or implausible?"
2. **Commonsense (Consensus).** Annotators indicate whether they believe most adults would agree with the judgment using everyday knowledge.

Each criterion is rated using a three-point scale: `Yes`, `No`, or `Maybe`. Annotators are required to provide a brief comment for any `No` or `Maybe` response (e.g., "could be true in some contexts" or "requires specialized knowledge").

**Definition of Commonsense** Annotators are instructed to interpret commonsense as knowledge that an average adult could reasonably be expected to possess, without relying on formal education, technical expertise, or domain-specific facts. Options requiring specialized scientific, historical, or technical knowledge are marked accordingly.

**Handling Ambiguity** Annotators are encouraged to consider realistic context and pragmatic reasoning. When an option is plausible only under rare or contrived interpretations, it should be marked as `Maybe` rather than `Yes`. The goal is to assess typical human judgments, not theoretical possibility.

**Adjudication** Two annotators independently evaluate each item. Disagreements, as well as any option receiving a `Maybe` are resolved through adjudication by a third reviewer. Only disputed cases are adjudicated; non-disputed options retain their original labels.

**Illustrative Examples** Annotators are provided with worked examples illustrating borderline cases, such as partially plausible distractors, context-dependent interpretations, and options that are topically related but logically invalid. These examples are used to clarify expectations but are not treated as templates.

### A.5 Training and Inference Details

This section reports the training configurations and inference settings used across all evaluated models.

Encoder-only and encoder-decoder models are fine-tuned with AdamW ($5 \times 10^{-5}$ learning rate, weight decay $0.01$) for up to 10 epochs with early stopping (patience 3), using a maximum sequence length of 512 and an effective batch size of 8.

Decoder-only LLMs are evaluated without gradient updates under 0–3 shot prompting (with/without CoT) and outputs are constrained to a single label (A–D). Most experiments run on NVIDIA P100 GPUs. Only the largest models and CoT evaluations are run on A100 GPUs.

### A.6 Additional Results

This section reports supplementary experimental results that complement the main findings discussed in Section 5. Table 4 reports accuracy and macro-F1 across all logical relations for human-validated (HV) and non-human-validated (NV) test subsets, including instruction-tuned and fine-tuned models. Table 5 presents macro-F1 for decoder-only LLMs under 0–3 shot prompting and CoT settings. Encoder-only and encoder–decoder models without fine-tuning exhibit uniformly low performance. Overall trends mirror the main paper: reasonable performance on AND and OR, and collapse on NEITHER/NOR across settings.

### A.7 Error Analysis

We analyze errors made by LLAMA-3.1-8B in the zero-shot setting. Table 6 summarizes representative failure patterns revealed by LOGICAL-COMMONSENSEQA. Because each option is a compositional pair of atomic statements, errors expose not only mis-classifications but also

| Model | AND | | | | OR | | | | NN | | | | MIX | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NV A | NV F1 | HV A | HV F1 | NV A | NV F1 | HV A | HV F1 | NV A | NV F1 | HV A | HV F1 | NV A | NV F1 | HV A | HV F1 |
| LLaMA-70B | 76.4 | 76.2 | 80.8 | 80.9 | 72.4 | 72.3 | 71.2 | 70.9 | 13.6 | 13.4 | 14.0 | 13.4 | 53.7 | 53.4 | 53.2 | 53.0 |
| LLaMA-8B | 67.2 | 66.7 | 72.0 | 71.9 | 63.2 | 62.6 | 62.2 | 62.2 | 8.6 | 7.9 | 13.9 | 13.1 | 44.5 | 43.9 | 42.7 | 41.8 |
| Qwen2.5-7B | 76.2 | 76.2 | 79.6 | 79.6 | 70.4 | 70.5 | 68.8 | 68.9 | 13.6 | 11.8 | 13.3 | 12.9 | 54.0 | 53.7 | 53.2 | 53.2 |
| Entailer-11B | 5.6 | 5.5 | 3.2 | 3.2 | 20.8 | 20.8 | 18.4 | 18.4 | 22.0 | 22.0 | 24.0 | 24.1 | 16.8 | 16.8 | 22.4 | 22.3 |
| Flan-T5-base | 38.8 | 38.0 | 37.2 | 36.2 | 63.2 | 62.2 | 65.2 | 64.3 | 12.0 | 11.4 | 16.8 | 16.3 | 32.8 | 32.5 | 33.6 | 33.0 |
| Flan-T5-base (FT) | 94.8 | 94.8 | 92.8 | 92.8 | 97.6 | 97.6 | 92.4 | 92.4 | 93.2 | 93.2 | 89.2 | 89.2 | 87.2 | 87.2 | 89.6 | 89.6 |
| DeBERTa-v3-base | 28.8 | 28.6 | 27.2 | 27.2 | 23.6 | 23.8 | 23.2 | 23.1 | 30.8 | 30.9 | 27.2 | 27.4 | 28.8 | 28.8 | 28.4 | 28.4 |
| DeBERTa-v3-base (FT) | 90.8 | 90.8 | 87.6 | 87.6 | 88.0 | 88.0 | 87.2 | 87.2 | 89.2 | 89.2 | 84.8 | 84.8 | 84.8 | 84.8 | 82.4 | 82.4 |

Table 4: **0-shot vs. fine-tuned:** Accuracy (A) and macro-F1 across four logical relations. NV = non-human-validated labels; HV = human-validated subset. NN = NEITHER/NOR, MIX = MIXED, FT = fine-tuned.

| Model | AND | | OR | | NN | | MIX | |
|---|---|---|---|---|---|---|---|---|
| | NV F1 | HV F1 | NV F1 | HV F1 | NV F1 | HV F1 | NV F1 | HV F1 |
| **0-shot** | | | | | | | | |
| LLaMA-70B | 76.2 | 80.9 | 72.3 | 70.9 | 13.4 | 13.4 | 53.4 | 53.0 |
| LLaMA-8B | 66.7 | 71.9 | 62.6 | 62.2 | 7.9 | 13.1 | 43.9 | 41.8 |
| Qwen-7B | 76.2 | 79.6 | 70.5 | 68.9 | 11.8 | 12.9 | 53.7 | 53.2 |
| **1-shot** | | | | | | | | |
| LLaMA-70B | 80.3 | 84.8 | 76.9 | 79.5 | 7.9 | 11.3 | 54.7 | 53.1 |
| LLaMA-8B | 72.1 | 67.0 | 61.7 | 62.0 | 5.9 | 5.3 | 45.0 | 41.7 |
| Qwen-7B | 73.9 | 79.6 | 75.0 | 71.6 | 5.7 | 7.4 | 57.4 | 55.4 |
| **2-shot** | | | | | | | | |
| LLaMA-70B | 78.7 | 83.6 | 74.0 | 76.8 | 9.1 | 12.2 | 53.1 | 51.7 |
| LLaMA-8B | 71.4 | 75.4 | 61.6 | 63.9 | 6.0 | 6.6 | 46.1 | 41.2 |
| Qwen-7B | 72.3 | 77.5 | 70.6 | 69.6 | 5.7 | 6.9 | 55.2 | 52.0 |
| **3-shot** | | | | | | | | |
| LLaMA-70B | 80.0 | 85.7 | 75.1 | 77.5 | 7.8 | 10.7 | 52.4 | 50.4 |
| LLaMA-8B | 75.8 | 71.3 | 61.0 | 61.3 | 5.5 | 6.1 | 47.0 | 42.8 |
| Qwen-7B | 73.4 | 77.9 | 73.2 | 71.9 | 4.6 | 6.8 | 55.5 | 51.4 |
| **CoT (0-shot)** | | | | | | | | |
| LLaMA-8B | 65.3 | 67.9 | 63.0 | 62.5 | 10.7 | 8.2 | 45.8 | 42.7 |

Table 5: Macro-F1 across logical relations under 0–3 shot prompting and chain-of-thought (CoT). NV = non-human-validated labels; HV = human-validated subset. NN = NEITHER/NOR, MIX = MIXED.

| Op | Failure Pattern | Representative Error (Gold-Pred) |
|---|---|---|
| **AND** | Single-statement dominance | **Q:** Heat applied to combustible materials?<br>Gold: emits toxic fumes AND decomposes into gases<br>Pred: ignites rapidly AND melts into liquid |
| **OR** | Plausibility collapse | **Q:** Heat applied to combustible materials?<br>Gold: ignites rapidly OR melts into liquid<br>Pred: melts into liquid OR burns without oxygen |
| **NEITHER** | Negation inversion | **Q:** Where do you buy a glass of wine?<br>Gold: NEITHER public park NOR home kitchen<br>Pred: NEITHER licensed restaurant NOR home kitchen |
| | Plausibility dominance | **Q:** Where might a child with no home go?<br>Gold: NEITHER school event NOR family gathering<br>Pred: NEITHER youth shelter NOR foster care |

Table 6: Representative error patterns for LLaMA-3.1-8B (0-shot) across AND, OR, and NEITHER/NOR.

deeper limitations highlighting how models fail to compose plausibility across logical operators.

**Conjunctive and Disjunctive Reasoning** Across both AND and OR settings, the dominant failure mode is the model's tendency to recognize only a *single* plausible statement and ignore how plausibility must be composed across clauses. In AND questions, this appears as *single-statement dominance*: the model anchors on a highly plausible clause and fails to reject an implausible partner. For example, in the combustibility question, it predicts *ignites rapidly AND melts into liquid*, correctly identifying ignition while overlooking that combustible materials do not melt.

The same underlying behavior manifests in OR questions. Rather than exploiting the requirement that at least one clause be plausible, the model treats OR as indicating thematic similarity. In the combustibility example, it predicts *melts into liquid OR burns without oxygen*, selecting two implausible but fire-related statements. These errors indicate that models do not enforce the compositional constraint imposed by the operator; instead, they rely on surface-level plausibility cues and semantic similarity.

**Negation-Based Reasoning** NEITHER/NOR proves most challenging. We observe *negation inversion*, where the model selects the *most plausible* pair of statements despite the operator requiring both to be implausible. For example, in the wine-purchase question, the model predicts *NEITHER licensed restaurant NOR home kitchen*, incorrectly negating the true location. A second pattern, *plausibility dominance*, reflects the model's tendency to retain highly plausible statements even under negation. For example, in the homeless-child example, the model selects *NEITHER youth shelter NOR foster care*, two locations that are plausible options. These behaviors indicate a broader difficulty with composing negation and plausibility simultaneously.