

FactNLI: Dynamic and Automated Fact-Based Augmentation of NLI Benchmarks

Yihan Lu, Junzhe Zhang, Huixuan Zhang, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
luyihan@stu.pku.edu.cn, junzhezhang@stu.pku.edu.cn
zhanghuixuan@stu.pku.edu.cn, wanxiaojun@pku.edu.cn

Abstract

Natural Language Inference (NLI) is a core task for language understanding, yet existing NLI datasets are static and increasingly saturated, making it difficult to diagnose model robustness in semantically enriched contexts. We propose **FactNLI**¹, a dynamic and automated framework that constructs more challenging NLI instances by iteratively fusing retrieved facts into the premise and hypothesis. A truth-set and graph-based filter removes contradictory and redundant candidates before fusion. Applied to SNLI and MNLI, FactNLI substantially increases input length and factual density, and accuracy drops consistently across a range of models (up to 30%). Length-controlled ablations show that this degradation is not explained by length or generic extra text alone, while human audits confirm high, though not perfect, label preservation and overall data quality.

1 Introduction

Natural Language Inference (NLI) is a foundational task in language understanding: given a premise, decide whether a hypothesis is entailed, contradicted, or neutral (Dagan et al., 2006). Despite its centrality, we argue that current benchmarks face two core weaknesses: **simplicity** and **staticity**. Modern systems often report very high scores on standard NLI benchmarks, yet these scores may not reveal how robustly they preserve NLI decisions when the inputs contain longer, semantically aligned factual context. Regarding **simplicity**, widely used datasets such as the Stanford Natural Language Inference corpus (SNLI; Bowman et al., 2015) and the Multi-Genre Natural Language Inference corpus (MNLI; Williams et al., 2018) concentrate on short, lexically cued items that hinge on single, local relations. Such items underrepresent longer contexts, multi-fact dependencies, and

diverse semantic phenomena, making it easy for models to succeed without robust content coverage or careful semantic comparison (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019).

Regarding **staticity**, these datasets are typically collected once (via crowdsourcing/annotation) and then frozen. As a result, they do not keep pace with evolving knowledge, domains, or usage, and they offer no mechanism to systematically adjust difficulty or enrich instances over time—an issue evident across widely used static resources such as SICK (Marelli et al., 2014), and XNLI (Conneau et al., 2018).

To address the issue of *simplicity*, earlier studies turned to human-in-the-loop adversarial collection: by iteratively eliciting model failures, they constructed more challenging and diverse examples. ANLI (Nie et al., 2020a) proceeds in rounds of human–model interaction to expose systematic weaknesses; Dynabench (Kielia et al., 2021) extends this paradigm into a platform for continuously updating datasets against deployed models; and WANLI (Liu et al., 2022) leverages worker–model collaboration to generate candidates that are subsequently filtered by humans. While such approaches improve hard-example coverage, they remain static once collected, incur substantial annotation costs, and do not provide a mechanism for continuously injecting verifiable evidence or systematically scaling difficulty.

In parallel, model-based synthetic augmentation has rapidly developed: large language models are prompted to generate new premise–hypothesis pairs, sometimes with templates or instructions, followed by lightweight filtering. Hosseini et al. (2024) showed that large-scale synthetic NLI can enhance domain generalization; more broadly, recent surveys confirm the potential of LLM-driven synthetic data to reduce annotation cost and expand coverage across domains and longer contexts (Nadăș et al., 2025). However, such methods rely

¹Code and data are available at <https://github.com/LuYanjie12/FactNLI>.

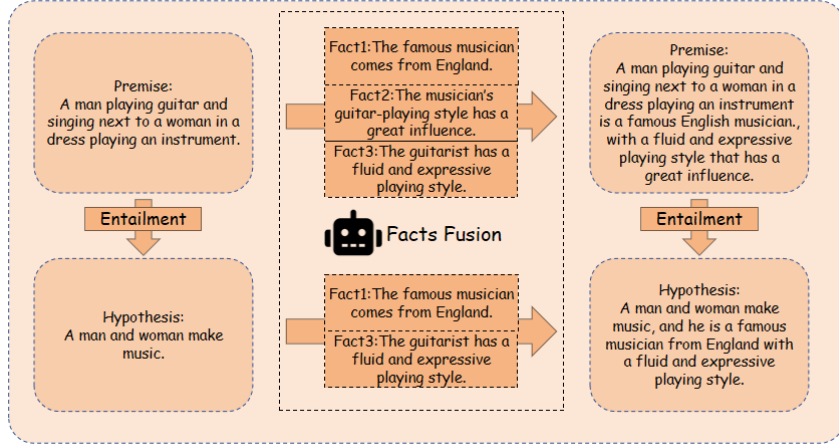


Figure 1: A one-level augmentation example with selected facts.

heavily on the generative model itself, often introducing unverifiable content, label drift, and stylistic artifacts. Even with partial human review, they lack explicit alignment with external evidence, making controllable difficulty and traceability difficult to guarantee.

To address these limitations, we propose a **dynamic, automated, fact-enhanced augmentation framework** that operates on existing NLI samples. For each (premise, hypothesis, label), we (i) retrieve premise-conditioned evidence from Wikipedia to ground the instance in verifiable content; (ii) filter candidate facts for non-contradiction and non-redundancy; and (iii) fuse accepted facts into the premise and a conservative subset into the hypothesis. A multi-level framework enables tunable difficulty by progressively composing more evidence while maintaining high audited label fidelity. Here *dynamic* describes the construction framework: it can be rerun with different root examples, retrieval corpora, generators, or augmentation depths without redesigning the task. The SNLI/MNLI-based benchmark released in this work is a fixed instantiation of that framework. Figure 1 shows an example with one-level augmentation.

Applied to SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), our augmentation (i) reveals consistent, depth-controlled performance drops across diverse model families (up to 30% at higher depths); (ii) expands example length by multiple folds (up to 10 $\times$) and increases semantic density via multi-fact composition; and (iii) preserves labels and instance coherence, as verified by ablations and human evaluation—thereby exposing

substantial headroom that static benchmarks conceal. To summarize, we list our main contributions as follows:

- We introduce a dynamic and automated fact-enhanced augmentation framework that enriches existing NLI items via a retrieval–filtering–fusion process, preserves original labels, and enables depth-controlled difficulty.
- We apply our augmentation framework to SNLI and MNLI and release new challenging benchmarks for the research community.
- We conduct a comprehensive evaluation across a range of LLM-based NLI models, showing consistent, depth-controlled performance degradation that highlights a substantive challenge understated by current static benchmarks.

2 Related Work

2.1 Classic NLI benchmarks and human-in-the-loop curation

The Stanford Natural Language Inference corpus (SNLI; Bowman et al., 2015) first established a large-scale, three-way classification benchmark, but its short, lexically cued pairs made it easy for models to exploit shallow cues rather than deep inference. The Multi-Genre NLI corpus (MNLI; Williams et al., 2018) broadened coverage across genres, yet still preserved relatively short contexts, leaving similar vulnerabilities. The SICK dataset (Marelli et al., 2014) focused on compositional semantics, but its small size limited robustness. XNLI (Conneau et al., 2018) extended MNLI to

15 languages, enabling cross-lingual evaluation but again freezing into a static test set. HANS (McCoy et al., 2019) directly targeted heuristic shortcuts, showing that high accuracy on SNLI/MNLI does not imply robust inference. DocNLI (Yin et al., 2021) scaled inference to full documents, though without mechanisms for dynamic difficulty control.

To increase difficulty, ANLI (Nie et al., 2020a) introduced iterative adversarial collection where annotators probe system weaknesses, raising difficulty but at significant annotation cost. Dynabench (Kielbaso et al., 2021) generalized this into a platform for continuous interactive collection, though instances ultimately re-freeze into fixed test sets. ChaosNLI (Nie et al., 2020b) proposed retaining distributions over human judgments, highlighting interpretive variability, and Jiang and de Marneffe (2022) further analyzed sources of label disagreement, but such work still lacks attached, verifiable evidence for each instance. In sum, while human-in-the-loop methods raise challenge and capture subjectivity, they remain expensive, hard to scale, and insufficient for long-context inference.

2.2 Synthetic augmentation for NLI

Synthetic generation has emerged as a practical response to data scarcity. STaTA (Vu et al., 2021) integrates self-training with LM-generated premise-hypothesis pairs, increasing volume but offering no guarantees of evidence alignment. The GAL framework (He et al., 2022) likewise generates unlabeled text and relies on teacher models for annotation, lowering labeling cost while risking label drift. Li et al. (2023) provide a systematic analysis showing that synthetic data can improve classification in some settings, yet task subjectivity moderates gains and raises concerns about unverifiable artifacts. WANLI (Liu et al., 2022) combines automatic filtering with selective human review after GPT-3 expansion of “challenging pockets” in MNLI, achieving strong out-of-domain improvements but at renewed human cost. More recent work emphasizes domain-diverse augmentation (Hosseini et al., 2024), underscoring the promise of scaling with LLMs while exposing persistent issues of model priors and style biases.

3 Method

Natural Language Inference (NLI) determines the relationship between a premise and a hypothesis (Dagan et al., 2006). The outcome is one of three

labels. *Entailment* means the premise provides sufficient information to conclude the hypothesis is true. *Contradiction* means the premise provides sufficient information to conclude the hypothesis is false. *Neutral* means the premise neither supports nor rules out the hypothesis.

3.1 Overview

Our goal is to automatically transform standard NLI samples into more challenging ones by adding verified facts to the premise and hypothesis in a multi-level manner. Given an input (p, h, l) with premise p , hypothesis h , and label $l \in \{\textit{entailment}, \textit{neutral}, \textit{contradiction}\}$, we iteratively expand p and h with external evidence while preserving logical consistency and the original label. After L levels, we obtain an enriched sample $(p^{(L)}, h^{(L)}, l)$ in which $p^{(L)}$ and $h^{(L)}$ incorporate verified new facts.

As shown in Fig. 2, our method consists of three steps at every level: (i) Facts acquisition — Retrieve clean, verifiable facts for augmentation. (ii) Truth-set and graph filtering — Remove conflicting and duplicate fact pairs to ensure non-redundancy and consistency. (iii) Premise and hypothesis fusion — Merge selected facts into the premise and hypothesis while preserving the original label. We present details of each step as follows.

3.2 Facts Acquisition

To construct the snippet pool used in later fusion, at level t we use the previous premise $p^{(t-1)}$ as the query to retrieve candidate Wikipedia pages. From each page, we take the introductory summary and split it into sentences. We compare sentence embeddings to the query embedding via cosine similarity. We keep sentences with score $\geq \tau$, where τ is a fixed, moderate threshold. The retained sentences form the snippet pool $P^{(t)}$. This pool is the only evidence used in subsequent steps (Fig. 2a).

Given only $P^{(t)}$, we apply a strong LLM (GPT-4o) to extract atomic facts. Specifically, the extraction prompt prohibits information beyond the retrieved snippet and requires each fact to be supported by exactly one snippet from $P^{(t)}$. The output is the candidate fact set $C^{(t)}$. We use Wikipedia because its public pages and inspectable source passages support evidence-level auditing; the same framework can be instantiated with other retrieval corpora. We preserve the retrieved page and source passage associated with each extracted and injected fact as provenance. Full prompt templates with

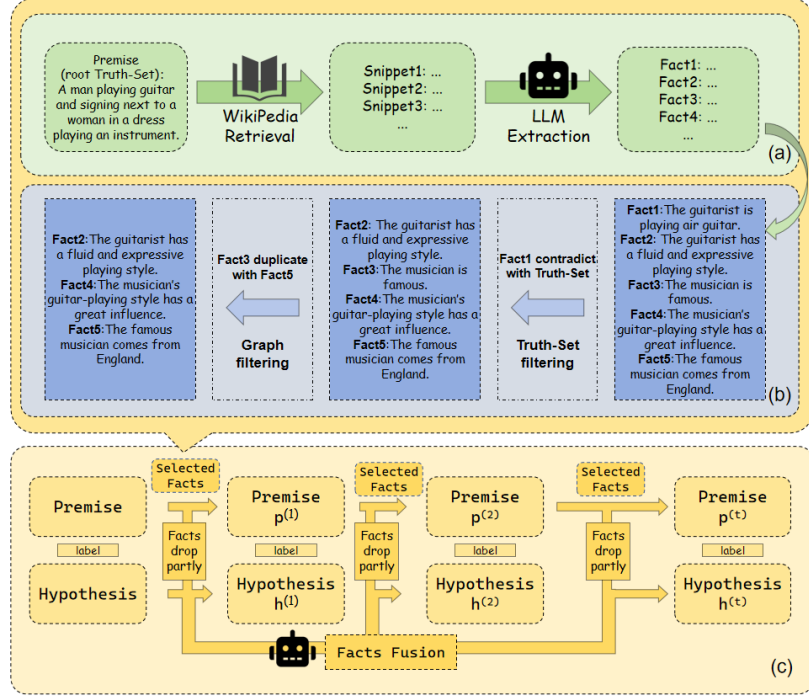


Figure 2: Overview of the full framework. (a) Premise-conditioned retrieval and fact extraction. (b) Truth-set and graph-based filtering of facts, where the initial truth set is derived from the original premise. (c) The complete multi-level augmentation framework; partial fact dropping denotes the subset of premise-side facts retained in the hypothesis.

additional rules appear in Appendix F.1. Further implementation details are provided in Section 4.1.

3.3 Truth-Set and Graph Filtering

To ensure the facts to be fused do not conflict with or duplicate the current premise, we maintain a truth set $T^{(t)}$ including facts fused at earlier levels and extracted from the original premise. We then construct an entailment graph G to check pairwise relations among candidate facts and remove conflicting or redundant items.

At level t , the candidates set $C^{(t)}$ is first checked against truth-set: for each $f \in C^{(t)}$, a lightweight NLI model compares f with every element of $T^{(t-1)}$ and assigns a relation in $\{\text{entailment, contradiction, neutral}\}$. Any f that contradicts an element of $T^{(t-1)}$ is discarded. Any f that is entailed by $T^{(t-1)}$ is marked as redundant and discarded. Facts that are neutral with respect to all elements of $T^{(t-1)}$ are retained. Denote the retained facts by $V^{(t)}$.

We discard conflicts and redundancies among the retained facts by constructing an undirected graph $G^{(t)}$ on $V^{(t)}$. An edge connects $u, v \in V^{(t)}$ if and only if their relation is *neutral*; pairs in *entailment* or *contradiction* are excluded to prevent

conflict and redundancy. We take the **maximum clique** of the neutral-edge graph $G^{(t)}$ (constructed on $V^{(t)}$) as the level- t fact set, denoted $F^{(t)}$. A clique guarantees pairwise neutrality among all selected facts; choosing the largest such clique yields the broadest subset that is simultaneously non-conflicting and non-redundant under our criterion.

We then update the truth set by

$$T^{(t)} = T^{(t-1)} \cup F^{(t)}.$$

This two-stage procedure removes conflicts with prior facts, suppresses items already implied by history, and at each level admits a maximal cluster of pairwise neutral (non-entailed, non-contradictory) facts, thereby preserving global consistency as T grows (Fig. 2b). Detailed algorithms are provided in Appendix G.

3.4 Premise and Hypothesis Fusion

At each level $t = 1, \dots, d$, we have selected a set of novel facts $F^{(t)}$ from the candidates after filtering. We then update the premise by fusing $F^{(t)}$ with the previous anchor $p^{(t-1)}$ via a constrained LLM prompt to produce a short paragraph $p^{(t)}$. The prompt requires $p^{(t)}$ to preserve and entail $p^{(t-1)}$,

cover the content in $F^{(t)}$, and introduce no information beyond $p^{(t-1)} \cup F^{(t)}$. Because fusion is performed by an LLM, these are generation constraints rather than formal guarantees; their empirical reliability is assessed in Section 4.6.

In contrast, for the hypothesis, we always enhance from the original hypothesis h . At level t , for each $i \in \{1, \dots, t\}$ we select a subset $\hat{F}^{(i)} \subseteq F^{(i)}$ from level i , and then form the aggregate $F^* = \bigcup_{i=1}^t \hat{F}^{(i)}$. We fuse F^* with h to obtain $h^{(t)}$. The prompt likewise requires $h^{(t)}$ to preserve and entail h , reflect F^* , and add nothing beyond $h \cup F^*$; we verify the resulting label preservation empirically rather than treating it as guaranteed solely by construction. The detailed prompt templates are provided in Appendix F.2. The final enhanced sample at level t is $(p^{(t)}, h^{(t)}, l)$.

Note that $h^{(t)}$ is always generated directly from the original hypothesis h (rather than from $h^{(t-1)}$) to avoid semantic drift and label reinterpretation across levels. This also enables controlled difficulty scaling: by aggregating F^* at higher t , we inject more filtered, semantically aligned content into the hypothesis without compounding generation artifacts.

Finally, across the L levels we obtain L enhanced premise–hypothesis pairs $\{(p^{(t)}, h^{(t)}, l)\}_{t=1}^L$ (Fig. 2c). We provide a complete case in Appendix K.

4 Experiments

4.1 Experimental Setup

Augmentation Data Source. We apply our framework to *all* instances from the following official splits (no sub-sampling): **SNLI** (Bowman et al., 2015) test; **MNLI** (Williams et al., 2018) `validation_matched` and `validation_mismatched`. Dataset sizes are shown in Table 1.

We choose these splits for two practical reasons: (i) many widely used pretrained or instruction-tuned models have been (directly or indirectly) fine-tuned on the training portions of SNLI/MNLI, so evaluating on the official held-out splits miti-

gates train–test contamination; and (ii) the public test split of MNLI does not release gold labels, hence evaluation customarily uses the labeled `validation_matched` and `validation_mismatched` splits instead.

Both corpora adopt a unified instance schema (p, h, l) comprising a *premise*, a *hypothesis*, and a *label*. In their original construction protocols, human annotators wrote hypotheses conditioned on sampled premises and assigned labels to the resulting pairs.

Artifact Use. We use only public artifacts: SNLI and MNLI, Wikipedia evidence, and published model APIs or released checkpoints, under their original release terms. We do not collect private user data, and we audit generated examples for label preservation, contradiction, factuality, redundancy, and readability.

Implementation Details. We deliberately separate *generation* from *filtering*: GPT-4o (OpenAI et al., 2024) performs snippet-faithful fact extraction and constrained fusion, while smaller models handle screening.

All screening steps rely on lightweight models for scalability: sentence-level *relevance* is computed with ALL-MINILM-L6-V2 embeddings (Reimers and Gurevych, 2019; Wang et al., 2020). Before comparison, we normalize whitespace and case, remove empty strings, and delete exact normalized duplicates; these operations do not perform semantic filtering. *Entailment/contradiction/neutral* edges in the truth-set graph are assigned by a compact cross-encoder based on DEBERTA-V3-BASE (He et al., 2021, 2023), following standard cross-encoder reranking practice (Nogueira and Cho, 2020). This design emphasizes GPT-4o for content generation while delegating relevance and entailment filtering to efficient small models.

We intentionally isolate generation from inference to avoid leakage and bias. Premise updates are generated only from the previous anchor and selected facts, and hypotheses are generated only from the root hypothesis and the aggregated facts—never from any premise text. Gold labels remain fixed, and E/C/N decisions (and relevance scoring) are handled by lightweight discriminative models rather than GPT-4o. This separation reduces label and cross-side leakage, keeps generation orthogonal to the core reasoning task, and yields more impartial examples.

DATASET	SPLIT	SIZE
SNLI	TEST	9,824
MNLI	VALIDATION_MATCHED	9,815
MNLI	VALIDATION_MISMATCHED	9,832

Table 1: Dataset sizes used in augmentation.

Tested Models. We evaluate seven models—ROBERTA-LARGE (Liu et al., 2019), DeBERTa-V3-LARGE (He et al., 2023), GPT-4o (OpenAI et al., 2024), LLAMA-3-8B-INSTRUCT (Grattafiori et al., 2024), QWEN2.5-14B-INSTRUCT (Qwen et al., 2025), DEEPSEEK-V3 (DeepSeek-AI et al., 2025), and DEEPSEEK-R1 (Guo et al., 2025)—in a zero-shot transfer setting on the original datasets ($L=0$, *orig*) and on FactNLI at three *levels* ($L=1, 2, 3$). Here zero-shot transfer means that no model is adapted on FactNLI. RoBERTa and DeBERTa use the public three-way NLI checkpoints FacebookAI/roberta-large-mnli and cross-encoder/nli-deberta-v3-large, respectively. Each receives (p, h) as a sequence pair, and its prediction is the argmax of the three classifier logits rather than a label-token probability.

4.2 Main results

Across all datasets and models, augmentation produces a *consistent accuracy drop* that grows with the *level*, as shown in Table 2. The largest stepwise decline typically occurs at the first enhancement ($0 \rightarrow 1$), with smaller but still nontrivial declines at higher levels, indicating that the enhanced sets increase task hardness while preserving labels.

Beyond the seven general-purpose models in Table 2, we also evaluate specialized NLI models trained on standard and adversarial NLI-style data. These models achieve strong scores on the original benchmarks (e.g., 92.9 on SNLI and 90.9 on MNLI for RoBERTa-large), but still drop substantially on FactNLI, reaching 62.9 and 69.8 at $L=3$ for the two specialized baselines. We also compare FactNLI with related adversarial and counterfactual resources and conduct a controlled transfer study between the original and augmented distributions. The transfer results position FactNLI as an evidence-dense but learnable NLI subdistribution: training on FactNLI improves performance on augmented levels while retaining substantial performance on original pairs, and balanced joint training performs best overall. These supplementary results are reported in Appendix C, Appendix D, and Appendix E.

To summarize, dynamic fact-enhanced augmentation consistently reduces accuracy across models and datasets in a level-controlled manner, turning standard NLI benchmarks into stronger stress tests for text inference abilities.

4.3 Analysis of Different Levels

We treat the original data as level $L=0$ (*baseline*) and track length as the enhancement level rises to $L=1, 2, 3$. We report *word* counts for both premise and hypothesis.

Across datasets, baselines are short: premises average 14–20 words and hypotheses 7–10. By $L=3$, premises reach 101–118 and hypotheses 70–76. In relative terms, premise length grows by $5\text{--}8\times$ and hypothesis length by $7\text{--}10\times$, with steady gains at each level (Fig. 3a), indicating increased semantic load on both sides with high audited preservation of the inherited labels.

In the fact-tracked subset, both premise-side and hypothesis-side fact counts rise monotonically with depth. Crucially, deeper levels do not merely pad the premise; they consolidate filtered facts into the hypothesis, increasing the amount of semantically aligned content that must be considered in the NLI decision (Fig. 3b).

Together, the length and fact-count trends show that higher levels add semantically aligned factual content rather than merely padding the input. This structural growth complements the accuracy trend in Section 4.2, supporting L as a predictable control over difficulty.

4.4 Difficulty Analysis

As shown in Section 4.2, accuracy drops steadily as we increase the enhancement level L . We investigate the underlying causes of this increased difficulty through controlled ablations.

We run two length-controlled ablations that keep surface form comparable while altering the informational content: (i) a rewrite-only variant that expands (p, h) to match the $L=1$ length distribution without adding new facts, and (ii) an unrelated-facts control that adds extra sentences of similar style and length but deliberately unrelated to the premise–hypothesis pair. Averaged across four models, accuracy is 85.2 on the original data, 67.2 on FactNLI- $L=1$, 78.5 on the length-matched rewrite-only control, and 77.5 on the length-matched unrelated-facts control. Thus, length or added text alone does not explain the substantially larger FactNLI drop; the principal difference is the presence of semantically aligned, entity-linked facts that models must selectively integrate or ignore. Detailed experimental results and further discussion are provided in Appendix H.

In addition, our confusion-matrix analysis (Ap-

MODEL	SNLI				MNLI			
	$ACC_{L=0}$	$ACC_{L=1}$	$ACC_{L=2}$	$ACC_{L=3}$	$ACC_{L=0}$	$ACC_{L=1}$	$ACC_{L=2}$	$ACC_{L=3}$
ROBERTA-LARGE	89.3	70.8	64.7	60.8	90.4	73.7	70.4	68.1
DEBERTA-V3-LARGE	92.4	77.1	69.0	67.2	90.1	74.2	71.9	69.7
LLAMA-3-8B-INSTRUCT	55.6	43.2	41.6	41.1	64.3	55.3	53.4	52.3
QWEN2.5-14B-INSTRUCT	82.4	64.7	59.4	56.7	82.3	62.3	59.5	58.8
GPT-4o	84.8	62.1	58.7	55.3	83.5	67.5	67.0	66.7
DEEPSEEK-V3	81.3	64.8	62.7	61.5	82.3	69.7	68.5	67.0
DEEPSEEK-R1	78.9	60.2	56.6	53.1	82.2	71.7	70.9	63.3
Average	80.7	63.3	59.0	56.7	82.2	67.8	65.9	63.7

Table 2: Zero-shot transfer accuracy on FactNLI at varying levels (L).

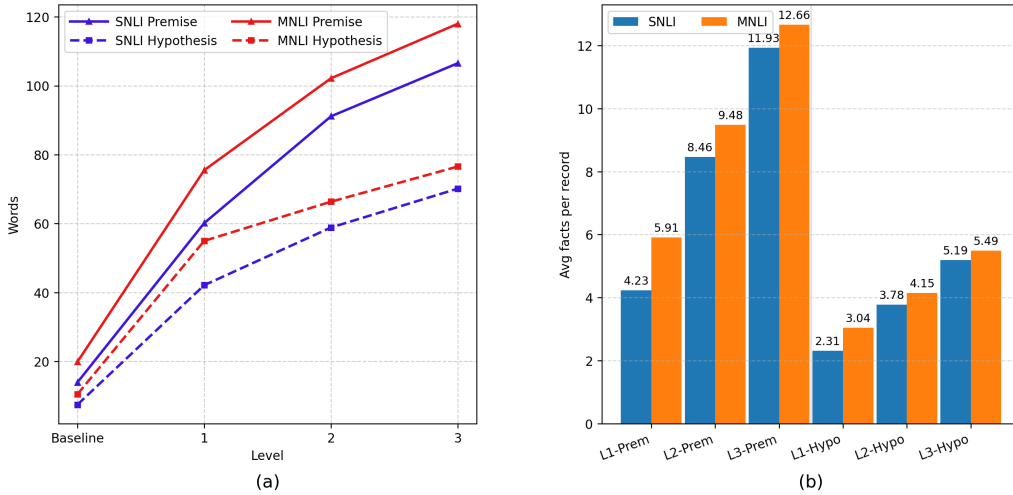


Figure 3: **Length growth and fact fusion by level.** (a) Average word counts vs. level ($L=0$ baseline, then $L=1, 2, 3$) for SNLI and MNLI, shown separately for *premise* (solid) and *hypothesis* (dashed). (b) Average fused facts per record at levels $L=1, 2, 3$ on both sides (Prem/Hypo); numbers above bars indicate the exact averages. Baseline ($L=0$) is omitted for clarity.

pendix I) shows that the performance drop is not explained by a collapse toward predicting entailment: instead of shifting mass predominantly to the entailment column, errors are distributed across labels (e.g., frequent confusion between contradiction and neutral), consistent with models struggling to identify which of many aligned facts is decisive.

These results indicate that the increased difficulty is driven primarily by the need to selectively reason over *semantically aligned*, entity-linked injected facts, rather than by length or generic extra context.

4.5 Ablations

Setup. To isolate the effect of the *truth set and entailment-graph filter*, we re-run the framework with the filter disabled (NO-FILTER)—which means we no longer maintain a truth set or an entailment graph—while holding everything else fixed. We report results at $L=3$ (primary endpoint) and include level sensitivity at $L=1, 2$. Our base NLI

pool is constructed by sampling from SNLI and MNLI in fixed proportions.

Results. The truth set and entailment-graph filter enforce global consistency by vetoing contradictions and suppressing redundancies. When we disable them, three failure modes emerge: (i) *label drift*—often driven by two mechanisms: items originally labeled *Contradiction* flip to *Neutral* or *Entailment* when we inject premise-entailed facts into both the premise and hypothesis (neutralizing the contradiction), and items originally labeled *Entailment* flip to *Contradiction* when conflicting facts are introduced; (ii) *corpus conflicts*—predominantly mismatches in time, location, or participants/roles (e.g., inconsistent dates, places, or entities across fused sources), as well as cross-source negations; and (iii) *redundant fusion*—already-entailed content is restated via near-duplicates or mechanical repetition across segments, inflating length (and sometimes superficial

MODEL	ACC _{L=0} (%)	ACC _{L=3} (%) FULL	ACC _{L=3} (%) NO-FILTER
ROBERTA-LARGE	90.3	59.3	26.0
GPT-4O	83.3	56.0	31.7

Table 3: Model accuracies at $L=0$ and $L=3$ under the Full framework and the NO-FILTER ablation.

coverage) without adding genuine new information.

Consequently, NO-FILTER shows a larger accuracy drop in Table 3 and may appear superficially “harder,” but this comes with degraded coherence: reduced label preservation, increased internal contradiction, and elevated redundancy.

In short, removing the filter makes instances look harder primarily by introducing inconsistency and noise rather than principled difficulty. To assess quality more accurately, we define additional evaluation metrics in Section 4.6.

4.6 Ex-post human audit

Design. Following common practice in NLI, fact verification, and factuality evaluation, we run a *blinded, stratified, double-annotation* audit (Bowman et al., 2015; Williams et al., 2018; Thorne et al., 2018; Nie et al., 2020a; Kiela et al., 2021; Maynez et al., 2020; Pagnoni et al., 2021). The submitted audit contains 180 Full-framework items (60 per level) and a separate 60-item $L=3$ NO-FILTER control. In the fact-visible protocol, annotators see the enhanced pair and a compact fact list, but not the original texts, labels, model predictions, or system variant. Two annotators label each item; a senior third annotator adjudicates disagreements. We include attention checks and a brief calibration round (Nangia et al., 2021), and report Cohen’s κ or weighted κ as appropriate. Unless noted otherwise, percentages are macro-averaged across strata. To match the model-input condition directly, we additionally reannotate the same 180 Full-framework items under a pair-only protocol that shows only the enhanced premise and hypothesis.

Annotation Task and Metrics. Beyond accuracy, we report five diagnostics: *label preservation* (agreement between the annotator-assigned label and the original gold label, computed only after annotation), *internal contradiction*, *redundancy*, *factuality* against construction evidence, and *readability*. The audit elicits one five-part judgment per item: absolute NLI label (E/N/C),

METRIC	$L=1$	$L=2$	$L=3$
LABEL PRES. (%)	98.3	96.7	95.0
CONFLICT (%)	1.7	5.0	6.7
REDUNDANCY (%)	11.7	10.0	13.3
FACTUALITY (%)	100.0	100.0	98.3
READABILITY	4.1	3.9	3.9

Table 4: Ex-post human audit by level ($L \in \{1, 2, 3\}$), macro-averaged over datasets.

internal contradiction (yes/no), redundant fusion (none/some/many), factuality w.r.t. construction evidence (supported/partially/unsupported), and readability (Likert 1–5) (Thorne et al., 2018; Maynez et al., 2020; Pagnoni et al., 2021). This design tests whether humans can infer the NLI relation from the enhanced case itself, and whether accuracy drops reflect principled difficulty rather than incoherent editing. Guidelines are in Appendix L.1.

Sampling. We pre-specify 240 audit items, stratified by dataset, level, and system variant; details are in Appendix L.2. The pair-only and fact-visible protocols produce identical E/N/C judgments item by item (180/180 agreement), so fact visibility does not affect the level-wise preservation estimates in this audited sample. We report the two protocols separately in Appendix L.3 because pair-only annotation validates labels under the model-input condition, whereas fact-visible annotation supports provenance and factuality checks.

Results and Analysis. Because *many* redundancy cases and *fully unsupported* factuality cases are rare, we report redundancy and factuality as binary outcomes. Table 4 summarizes level-wise results for the full framework. Pair-only label preservation is 98.3% at $L=1$ (Wilson 95% CI [91.1, 99.7]), 96.7% at $L=2$ ([88.6, 99.1]), and 95.0% at $L=3$ ([86.3, 98.3]). These intervals characterize sample-level reliability rather than guaranteeing every generated item. Conflicts are rare, redundancy is controlled, and readability decreases slightly but stays in the “clear” range. Complementary component audits report 90% retrieval relevance, 94% extraction supportedness, 92% atomicity, 92.5% truth-set decision accuracy, and 96% graph-relation accuracy (Appendix B).

The residual 5% error estimated at $L=3$ warrants caution for close model comparisons. At error rate ϵ , a model’s measured accuracy can change by at most ϵ , and a pairwise gap by at most 2ϵ . Thus, the estimated error cannot explain the principal $L=0$ -

METRIC	FULL	NO-FILTER
LABEL PRES. (%)	95.0	65.0
CONFLICT (%)	6.7	31.7
REDUNDANCY (%)	13.3	28.3
FACTUALITY (%)	98.3	98.3
READABILITY	3.9	3.3

Table 5: Full framework vs. NO-FILTER at $L=3$, macro-averaged over datasets.

DIMENSION	κ
ABSOLUTE NLI LABEL (E/N/C)	0.94
INTERNAL CONTRADICTION (YES/NO)	0.83
REDUNDANCY (NONE/SOME/MANY)	0.68
FACTUALITY VS. FACT LIST (3-WAY)	0.98
READABILITY (LIKERT 1-5; WEIGHTED)	0.57

Table 6: Inter-annotator agreement (IAA). Cohen’s κ for categorical tasks; linearly weighted κ for ordinal scales.

to- $L=3$ drops of 24.1 points on SNLI and 18.5 points on MNLI, although gaps below 10 points are not worst-case robust. A sensitivity analysis that rescores the seven models on the 60 audited $L=3$ items with adjudicated labels raises aggregate correct decisions from 235/420 to 239/420 while leaving all seven descriptive ranks unchanged; full results are in Appendix L.3.

Table 5 shows that, at $L=3$, disabling the filter (NO-FILTER) clearly degrades corpus quality, causing label drift, contradictions, repetition, and poorer readability. Factuality remains comparable, supporting the fusion mechanism. Because this drop is unacceptable for evaluation or downstream use, we treat the filtered framework as the reliable setting. Final IAAs are in Table 6; agreement is consistently high across dimensions.

In summary, human evaluation shows that our augmentation raises difficulty while preserving quality. The filter is crucial: it maintains labels, limits contradictions, and controls redundancy, whereas NO-FILTER causes a clear quality drop.

Cross-model Robustness. Replacing GPT-4o extraction and fusion with Qwen2.5-14B-Instruct yields the same qualitative monotonic degradation across all three tested solvers, but with different drop magnitudes. We find no consistent self-generator advantage; generator choice therefore affects the exact distribution without accounting for the overall difficulty trend. Three repeated generations per backbone further show modest aggregate variation. Details are in Appendix J.

5 Conclusion

We presented a *dynamic, automated fact-based* augmentation framework that turns existing NLI examples into more challenging, evidence-grounded ones. By iteratively retrieving, distilling, filtering, and fusing atomic facts, the method increases semantic richness while preserving labels; the level parameter L provides a simple control over difficulty. Across SNLI and MNLI, FactNLI produces consistent accuracy reductions across models, while ablations and human audits show that the truth-set and graph filter is essential for maintaining label preservation, low contradiction, controlled redundancy, and strong evidence grounding.

Limitations

This work has several limitations. First, the framework is built around English Wikipedia and English NLI benchmarks, so its coverage is constrained by the breadth and bias of those sources. Second, fact extraction and fusion still rely on LLMs, which introduces residual generation bias and a small amount of label noise despite filtering. Third, we evaluate the method on SNLI and MNLI and on a finite set of models; broader transfer to other tasks, domains, and languages remains future work. Finally, the pipeline is API- and generation-heavy, so the practical cost is dominated by hosted-model latency rather than the graph filter itself.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (L253001), Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology) and National Engineering Research Center of New Electronic Publishing Technologies. We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the contact author.

References

- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and

- Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *Preprint*, arXiv:2111.09543.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). *Preprint*, arXiv:2006.03654.
- Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. [Generate, annotate, and learn: NLP with synthetic text](#). *Transactions of the Association for Computational Linguistics*, 10:826–842.
- Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. [A synthetic data approach for domain generalization of NLI models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2212–2226, Bangkok, Thailand. Association for Computational Linguistics.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. [Investigating reasons for disagreement in natural language inference](#). *Transactions of the Association for Computational Linguistics*, 10:1357–1374.
- Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. [Learning the difference that makes a difference with counterfactually-augmented data](#). *Preprint*, arXiv:1909.12434.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ring-shia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. [Synthetic data generation with large language models for text classification: Potential and limitations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [WANLI: Worker and AI collaboration for natural language inference dataset creation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. [A sick cure for the evaluation of compositional distributional semantic models](#). In *International Conference on Language Resources and Evaluation*.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. 2025. [Synthetic data generation using large language models: Advances in text and code](#). *IEEE Access*, 13:134615–134633.
- Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. [What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. [What can we learn from collective human opinions on natural language inference data?](#) *Preprint*, arXiv:2010.03532.
- Rodrigo Nogueira and Kyunghyun Cho. 2020. [Passage re-ranking with bert](#). *Preprint*, arXiv:1901.04085.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. [Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- 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, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. [STraTA: Self-training with task augmentation for better few-shot learning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. [DocNLI: A large-scale dataset for document-level natural language inference](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

A Runtime and Scalability

The maximum neutral clique step is NP-hard in the worst case, but each filtering step in our setting operates on a small local candidate pool, typically

10–20 facts per instance. We implement clique selection with Bron–Kerbosch search with pivoting. In practice, this graph step is not the dominant runtime cost; retrieval and LLM-based extraction/fusion account for most of the wall-clock time. Table 7 reports the observed cost for 100 instances with 20 concurrent API calls.

Stage	$L=1$	$L=2$	$L=3$
Retrieval	42s	48s	55s
LLM extraction	238s	276s	322s
Truth-set filtering	14s	20s	22s
Graph filtering	3s	5s	5s
LLM fusion	121s	146s	179s
Tokens	5.2k	6.1k	7.3k
Total	418s	495s	583s

Table 7: Runtime for 100 generated instances.

These measurements show that graph filtering contributes only a small fraction of total runtime. For substantially larger candidate pools, the exact clique step can be replaced by a greedy or approximate compatible-set selection strategy without changing the overall retrieve–filter–fuse framework.

We also benchmark the complete pipeline at 100, 500, and 1,000 roots under concurrency 20. API-call counts include retries, and token usage separates input from output. Table 8 reports end-to-end measurements under the service configuration used for generation.

Across the measured range, wall time and billed usage scale approximately proportionally while throughput remains stable. At 100 roots, retrieval, extraction, filtering, and fusion take 55, 322, 27, and 179 seconds, respectively, confirming that hosted extraction and fusion dominate runtime. We restrict this near-linear scaling claim to the measured range of 100–1,000 roots and do not extrapolate beyond it.

B Intermediate Pipeline Diagnostics

To quantify intermediate component quality, we use the same ALL-MINILM-L6-v2 embedding space used for sentence-level relevance screening. We report two similarity diagnostics: *premise–evidence similarity*, measuring how closely retrieved Wikipedia evidence matches the original premise, and *fact–source similarity*, measuring how closely extracted atomic facts align with their source Wikipedia sentences.

Diagnostic	Average similarity
Premise–evidence	0.43
Fact–source	0.85

Table 9: Embedding-based intermediate diagnostics.

The premise–evidence score is moderate because retrieved Wikipedia snippets are often longer than the original premise and include additional background context. By contrast, fact–source similarity is high because atomic facts are direct abstractions of individual source sentences. These diagnostics suggest that retrieved evidence remains meaningfully related to the premise while extracted facts stay tightly grounded in their source sentences.

At $t=0$, the truth set is initialized by applying the same atomic-fact extraction procedure to the original premise. This initial truth set is not intended to exhaustively encode all background knowledge implied by a short premise; rather, it provides an initial factual compatibility set used to block contradiction and redundancy during later augmentation steps.

Human audit of intermediate components.

The embedding scores above are useful automatic checks, but they do not directly verify whether each intermediate decision is correct. We therefore add a component-level human audit that evaluates retrieval, fact extraction, and filtering before final fusion. The audit samples intermediate artifacts from generated examples across SNLI and MNLI, balanced across enhancement levels where applicable. Each artifact is independently judged by two annotators; disagreements are adjudicated by a third annotator. Annotators see only the local evidence needed for the decision, such as a premise–evidence pair, a source-sentence–fact pair, a current-truth-set–candidate pair, or a candidate-fact pair with the system-assigned relation. They do not see final model predictions or downstream evaluation labels.

Table 10 defines the annotation units and labels. Retrieval is marked correct when the retrieved sentence or passage is topically and entity-wise relevant to the original premise. Fact extraction is evaluated along two axes: supportedness, requiring the extracted fact to be entailed by the source sentence, and atomicity/self-containedness, requiring the fact to express one standalone proposition without unresolved pronouns or bundled claims. Truth-set filtering is marked correct when the accept/reject

Roots	Wall time	Roots/hour	API calls	Input/output tokens	Retry	Cost
100	583s	617	920	2.10M/0.45M	2.0%	\$9.80
500	2,940s	612	4,610	10.60M/2.25M	2.2%	\$49.00
1,000	5,980s	602	9,230	21.30M/4.51M	2.4%	\$98.40

Table 8: End-to-end scalability of the augmentation pipeline.

decision matches the annotator’s judgment about whether the candidate is compatible with the current truth set and non-redundant. Graph filtering is marked correct when the pairwise relation used for neutral-clique construction correctly identifies compatible, redundant/entailed, or conflicting fact pairs.

Component	Input unit	Judgment
Retrieval	Premise + evidence	Relevant / not relevant
Fact extraction	Source sentence + fact	Supported / not supported
Fact extraction	Extracted fact	Atomic+self-contained / not
Truth set	Truth set + candidate	Correct accept/reject
Graph filter	Fact pair + relation	Correct relation

Table 10: Intermediate component audit design.

We report the main audit results in Table 11. For retrieval, fact supportedness, and atomicity, the rate is the percentage of audited units receiving a positive judgment after adjudication. For truth-set and graph-filter decisions, the rate is decision accuracy against the adjudicated human label. The final column records the number of audited units, so that the table makes the scale of each component audit explicit.

Component	Metric	Rate	<i>N</i>
Retrieval	Evidence relevance	90.0%	40
Fact extraction	Supportedness	94.0%	50
Fact extraction	Atomic+self-contained	92.0%	50
Truth set	Decision correctness	92.5%	40
Graph filter	Relation correctness	96.0%	50

Table 11: Human audit of intermediate components.

We also report inter-annotator agreement for the same component-level tasks in Table 12. Since all intermediate judgments are categorical, we use Cohen’s κ before adjudication. This agreement table separates component reliability from final adjudicated accuracy, making it possible to distinguish system errors from ambiguous annotation cases.

Task	κ	Disagreement rate
Evidence relevance	0.76	12.5%
Fact supportedness	0.82	8.0%
Fact atomicity	0.74	12.0%
Truth-set decision	0.70	15.0%
Graph relation	0.68	16.0%

Table 12: IAA for intermediate component audit.

Finally, Table 13 provides an optional compact error breakdown for the two filtering components. This table is useful when a decision is incorrect but the failure mode matters: a rejected fact may actually be compatible, an accepted fact may be redundant, or a graph edge may confuse entailment with contradiction or neutrality.

Component	Error type	Count	Rate
Truth set	False reject	2	5.0%
Truth set	False accept: redundant	2	5.0%
Truth set	False accept: conflicting	1	2.5%
Graph filter	Compatible marked non-compatible	3	6.0%
Graph filter	Redundant/entailed marked compatible	2	4.0%
Graph filter	Conflict missed	2	4.0%

Table 13: Filtering error breakdown.

C Specialized NLI Baselines

To test whether FactNLI remains challenging for models explicitly trained on robust NLI-style data, we evaluate specialized NLI baselines in addition to the general-purpose models in the main results.

Model	Train	SNLI	MNLI	L1	L2	L3
RoBERTa-large	SNLI/MNLI/FEVER/ANLI	92.9	90.9	76.8	70.4	62.9
DeBERTa-v3-large	MNLI/FEVER/ANLI/LingNLI/WANLI	92.8	91.0	79.4	73.1	69.8

Table 14: Specialized NLI baseline accuracy.

Compared with their vanilla counterparts, these models achieve modest gains but still show clear degradation as the FactNLI level increases. This suggests that FactNLI is not merely an artifact of zero-shot transfer or weak task adaptation; it remains challenging for models optimized for NLI robustness.

D Comparison with Related Augmentation Resources

FactNLI is related to adversarial and counterfactual NLI resources, but it targets a different construction regime: automatic, fact-grounded, rerunnable augmentation of existing NLI pairs. Table 15 summarizes the comparison. For quality metrics such as readability and label preservation, we apply the same human-audit protocol used in Section 4.6: annotators judge the enhanced premise–hypothesis pair with the provided fact list, without access to the original/root texts, original labels, model predictions, or system identity.

Aspect	FactNLI	ANLI	CF-SNLI
Paradigm	Fact-grounded aug.	Adversarial	Counterfactual
Human effort	Audit only	High	High
Rerunnable construction	Yes	No	No
Difficulty control	$L=1/2/3$	$R=1/2/3$	No
Readability	4.1 / 3.9 / 3.9	4.0 / 3.8 / 3.7	4.3
Label preservation	98.3 / 96.7 / 95.0	Human labels	Human edits
Scale cost	Low	High	High

Table 15: Comparison with related NLI resources.

ANLI is a human-and-model-in-the-loop adversarial benchmark, and CF-SNLI (Kaushik et al., 2020) focuses on human-written counterfactual edits. In contrast, FactNLI augments existing NLI examples through retrieval, filtering, and controlled fusion. Its main advantage is not that it replaces adversarial or counterfactual datasets, but that it provides a low-cost, evidence-grounded way to rerun construction and scale difficulty while preserving strong human-audited quality.

E Bidirectional Transfer between Original NLI and FactNLI

To characterize the distributional relation between original NLI and FactNLI, we conduct a controlled transfer experiment on 5,668 released SNLI-based roots with complete $L=1$ – $L=3$ variants. We construct a label-stratified, root-disjoint split of 3,967/566/1,135 train/dev/test roots and train the same non-NLI-tuned RoBERTa-base under Original, FactNLI-Mix, and balanced joint conditions. All conditions use three seeds and matched training budgets. This is a root-level re-split of the complete paired released corpus, not the official SNLI train/test split.

Training data	$L=0$	$L=1$	$L=2$	$L=3$
Original	83.11±0.40	57.06±2.56	46.99±3.81	43.26±3.01
FactNLI-Mix	81.74±3.86	74.68±1.15	63.09±1.09	61.41±1.55
Original+FactNLI	83.44±0.57	76.64±0.73	65.50±1.12	63.68±0.99

Table 16: Bidirectional transfer accuracy (mean ± standard deviation over three seeds).

Original-only transfer falls from 83.11 on $L=0$ to 43.26 on $L=3$. FactNLI-Mix retains 81.74 on original pairs while improving the augmented levels to 74.68/63.09/61.41. Joint training performs best overall, reaching 83.44 on $L=0$ and 76.64/65.50/63.68 on $L=1/2/3$. These results position FactNLI as an evidence-dense English NLI subdistribution that provides complementary, learnable training signal.

F Prompt

F.1 Fact Extraction

Fact Extraction Prompt

```
[Task Description]
You are a meticulous information extractor. Your sole job
is to read the given SNIPPETS and return only
literal, atomic facts present in those snippets. You
must not infer, generalize, or add outside knowledge.

[Inputs]
{SNIPPETS}

[Rules]
1) Source-only: use only snippet content; preserve tense,
   modality, quantifiers, and negation exactly as
   written.
2) Minimality: each fact must be irreducible (cannot be
   further split).
3) No cross extraction: each fact must be fully supported
   by a single snippet. Do not combine evidence across
   multiple snippets.
4) No ambiguous references: avoid vague or deictic
   pronouns (e.g., it, they, he, she, this, that,
   former, latter, here, there) ; replace with the
   explicit noun phrase from the snippet.
5) If nothing is extractable, return {"facts": []} exactly.

[Output JSON]
{
  "facts": [
    {
      "text": "<atomic fact1>",
    },
    {
      "text": "<atomic fact2>",
    },
    ...
  ]
}
```

F.2 Premise and Hypothesis Fusion

The fusion procedure for the premise and the hypothesis is identical. In both cases, the text to be fused is provided as the ANCHOR to the fusion step.

Premise and Hypothesis Fusion Prompt

[Task Description]

You are a precise composer. Begin with the given ANCHOR text verbatim, then add ONLY the supplied atomic facts from FACT_LIST_JSON. Do not introduce any new information beyond those facts.

[Inputs]

{ANCHOR_TEXT}
{FACT_LIST_JSON}

[Rules]

- 1) Anchor-centric: write a short paragraph that revolves around ANCHOR_TEXT.
- 2) Facts-only: integrate all and only the provided facts; each fact appears once.
- 3) Fidelity: preserve polarity, modality, time expressions, and named entities as given.
- 4) Minimal glue: use neutral connectors ("Additionally,", "At <time>,", "In <location>,", "Meanwhile,") without adding new information.

[Output JSON]

```
{
  "fusion_result": "<fused_paragraph>"
}
```

G Algorithm

Algorithm 1 Truth-set filtering and maximum neutral clique selection.

```

1: Input: truth set  $T$ , candidate facts  $F$ , NLI model  $\mathcal{M}$ , thresholds  $\tau_e, \tau_c$ 
2: Output: maximum neutral clique  $S^* \subseteq F$  and induced entailment subgraph  $G^{\Rightarrow}[S^*]$ 
3: Predicates:
4:  $\text{ENTAILS}(a, b) \triangleq \text{Pr}_{\mathcal{M}}(a \Rightarrow b) \geq \tau_e$ 
5:  $\text{CONTRADICTS}(a, b) \triangleq \text{Pr}_{\mathcal{M}}(a \perp b) \geq \tau_c$ 
6:  $\text{NEUTRAL}(a, b) \triangleq \neg \text{ENTAILS}(a, b) \wedge \neg \text{ENTAILS}(b, a) \wedge \neg \text{CONTRADICTS}(a, b) \wedge \neg \text{CONTRADICTS}(b, a)$ 
7: Stage 1: Truth-set filtering
8:  $R \leftarrow \emptyset$ 
9: for  $f \in F$  do
10:   if  $\forall t \in T : \neg \text{CONTRADICTS}(t, f) \wedge \neg \text{CONTRADICTS}(f, t) \wedge \neg \text{ENTAILS}(t, f)$  then
11:      $R \leftarrow R \cup \{f\}$ 
12:   end if
13: end for
14: Stage 2: Entailment edges and neutral graph on  $R$ 
15:  $E^{\Rightarrow} \leftarrow \emptyset$ ;  $\text{Adj}(u) \leftarrow \emptyset, \forall u \in R$ 
16: for  $i = 1$  to  $|R|$  do
17:   for  $j = i+1$  to  $|R|$  do
18:      $u \leftarrow R[i]; v \leftarrow R[j]$ 
19:     if  $\text{ENTAILS}(u, v)$  then
20:        $E^{\Rightarrow} \leftarrow E^{\Rightarrow} \cup \{(u \rightarrow v)\}$ 
21:     end if
22:     if  $\text{ENTAILS}(v, u)$  then
23:        $E^{\Rightarrow} \leftarrow E^{\Rightarrow} \cup \{(v \rightarrow u)\}$ 
24:     end if
25:     if  $\text{NEUTRAL}(u, v)$  then
26:        $\text{Adj}(u) \leftarrow \text{Adj}(u) \cup \{v\}; \text{Adj}(v) \leftarrow \text{Adj}(v) \cup \{u\}$ 
27:     end if
28:   end for
29: end for
30: Stage 3: Maximum neutral clique (Bron-Kerbosch + pivot)
31:  $S^* \leftarrow \emptyset$ 
32: Function  $\text{BK\_PIVOT}(C, P, X)$ :
33:   if  $P = \emptyset$  and  $X = \emptyset$  then
34:     if  $|C| > |S^*|$  then
35:        $S^* \leftarrow C$ 
36:     end if
37:     return
38:   end if
39:   Choose  $u \in P \cup X$  maximizing  $|P \cap \text{Adj}(u)|$ 
40:   for each  $v \in P \setminus \text{Adj}(u)$  do
41:      $\text{BK\_PIVOT}(C \cup \{v\}, P \cap \text{Adj}(v), X \cap \text{Adj}(v))$ 
42:    $P \leftarrow P \setminus \{v\}; X \leftarrow X \cup \{v\}$ 
43: end for
44: End Function
45:  $\text{BK\_PIVOT}(\emptyset, R, \emptyset)$ 
46: Stage 4: Induced entailment subgraph
47:  $G^{\Rightarrow} \leftarrow$  subgraph of  $(R, E^{\Rightarrow})$  induced by  $S^*$ 
48: return  $S^*, G^{\Rightarrow}$ 

```

H Analysis of Difficulty

The previous analysis shows that our multi-level augmentation substantially increases both sequence length and the number of fused facts, and that model accuracy consistently decreases with higher levels. A natural question is whether this difficulty is merely a side effect of longer, noisier inputs, or whether it stems from models having to infer with additional, semantically aligned facts. We address this with two controlled ablations that keep length and surface form comparable while manipulating the informational content of the inputs.

Model	Orig.	FactNLI $L=1$	Rewrite	Unrelated
GPT-4o	84.8	62.1	78.6	80.7
DeepSeek-V3	81.3	64.8	76.9	71.9
DeBERTa-v3-large	92.4	77.1	80.4	85.2
Qwen2.5-14B-Instruct	82.4	64.7	78.0	72.1

Table 17: Difficulty ablations with length-controlled inputs.

Length-controlled rewrite-only ablation. First, we isolate the effect of sequence length. For each original (p, h) pair, we construct a rewrite-only variant $(\tilde{p}, \tilde{h})$ by prompting a LLM to expand p and h into longer paraphrases that preserve their meaning, while targeting the same length distribution as our fact-augmented data (we use the $L=1$ setting as the target length, since matching $L=3$ would require extreme expansion and makes it difficult for the rewrite-only variant to preserve the original meaning without introducing new information). Crucially, no external facts are added in this condition: all content in $(\tilde{p}, \tilde{h})$ is a rephrasing or elaboration of the original sentence pair. We then evaluate the same models on the original data, the FactNLI $L=1$ data, and the rewrite-only (length-matched) data.

Across all evaluated setups (Table 17), rewriting to match FactNLI length leads to only modest drops relative to the original benchmark, whereas FactNLI- $L=1$ produces substantially larger declines for the same models and datasets. This pattern indicates that longer sequences and more tokens, by themselves, do not fully account for the observed difficulty; the specific way in which we inject external factual content matters. Notably, DEBERTA-V3-LARGE shows a more pronounced drop, which we conjecture reflects the reduced robustness of smaller models to longer-context inputs and out-of-distribution generalization, while stronger models remain comparatively stable.

Unrelated-facts ablation. To test whether the injected facts themselves truly participate in inference and are responsible for the additional difficulty, rather than generic noise from extra sentences, we design an *unrelated-facts* control. Starting from the same original (p, h) pairs, we follow the FactNLI fusion protocol but replace filtered Wikipedia facts with commonsense or encyclopedic statements that are deliberately unrelated to the premise and hypothesis (different entities, topics, or events), while keeping length and discourse style comparable to FactNLI. In other words, this variant adds irrelevant information around (p, h) .

As shown in Table 17, augmenting with unrelated factual sentences again yields only small changes in accuracy compared to the original data, while FactNLI causes a much larger drop on the same models. Since sequence length and the amount of added text are comparable across FactNLI and the unrelated-facts control, this result suggests that the extra difficulty is not driven simply by more context or more information, but by the presence of semantically aligned facts that interact with the entities and events in (p, h) and must be selectively integrated or ignored.

Taken together, these two ablations show that FactNLI increases difficulty in a way that goes beyond one-dimensional length scaling. Models do not fail merely because the inputs are longer or noisier; they fail when they must jointly reason over the original premise–hypothesis pair and a large set of newly injected, entity-aligned facts, treating *all* atomic statements in the combined context as candidates for inference. This is precisely the kind of evidence that our augmentation is designed to introduce, and it explains why performance on FactNLI is substantially lower than on both the original benchmarks and length-matched controls.

I Confusion Matrices

J Open-Source Generation

Our main experiments use GPT-4o as the backbone generator in the retrieve–filter–fuse framework. To verify that the framework is not tied to a single proprietary model, we additionally instantiate the generative components with the open-source **Qwen2.5-14B-Instruct** model on the SNLI dataset.

Generation setup. We keep the retrieval procedures, fact-filtering steps, and label-preservation checks identical to the GPT-4o setup, and only replace the backbone generator with Qwen2.5-14B-Instruct. For each base SNLI pair and enhancement level $L=0, 1, 2, 3$, Qwen expands the original premise and hypothesis into longer, more discursive versions and integrates retrieved Wikipedia facts at levels $L=1, 2, 3$, ensuring entity and topic alignment. Prompts follow the same structure as in Appendix F, adapted to Qwen’s chat template.

Generator–solver interaction. We evaluate three solvers on matched SNLI roots generated by GPT-4o and Qwen2.5-14B-Instruct while holding retrieval, filtering, and label checks fixed. Table 18

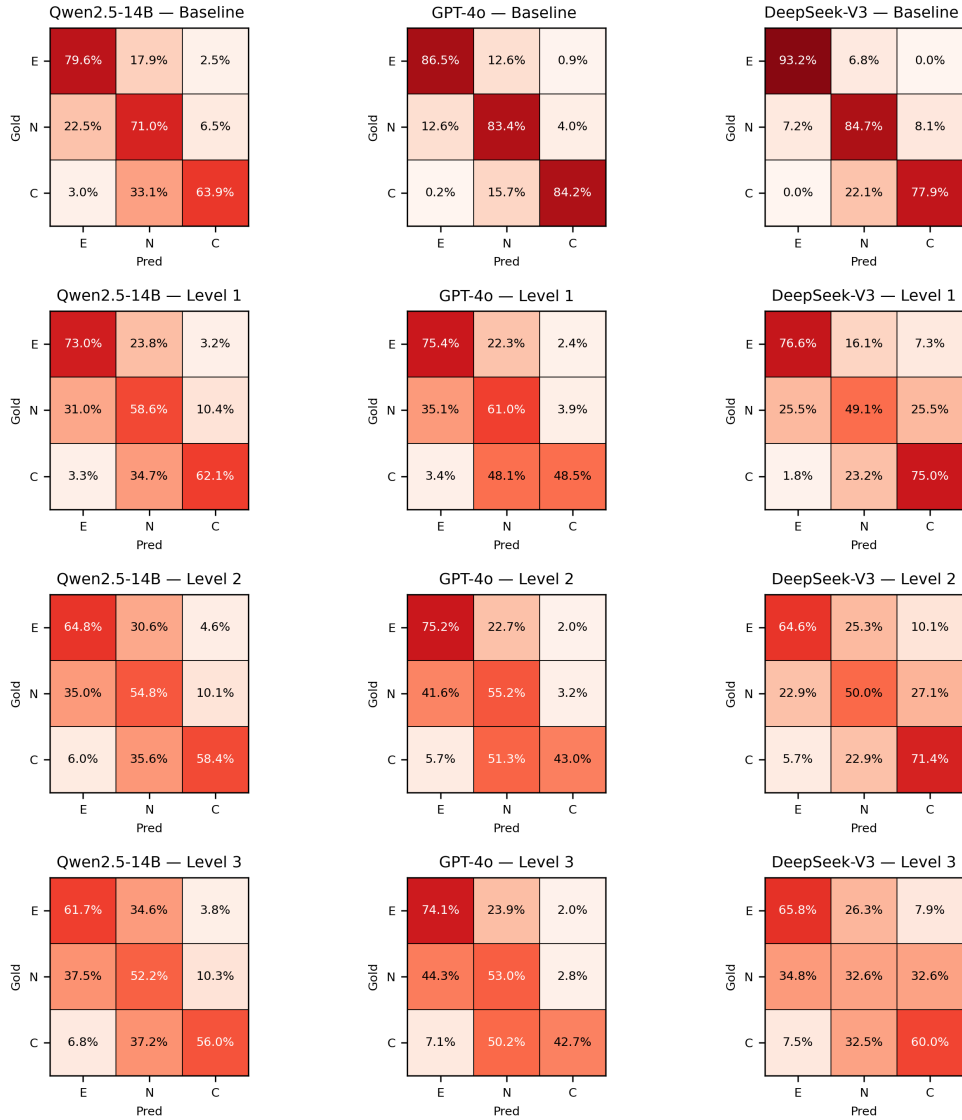


Figure 4: Confusion matrices across enhancement levels and models.

reports $L=3$ accuracy and the corresponding $L=0$ -to- $L=3$ drop.

Generation source	GPT-4o	Qwen2.5-14B	DeepSeek-V3
GPT-4o	55.3 (−29.5)	56.7 (−25.7)	61.5 (−19.8)
Qwen2.5-14B	50.0 (−34.8)	51.6 (−30.8)	45.8 (−35.5)

Table 18: Generator–solver interaction at $L=3$; parentheses show the accuracy drop from $L=0$.

There is no consistent self-generator advantage: GPT-4o drops by 29.5 points on GPT-generated data versus 34.8 on Qwen-generated data, while Qwen drops by 25.7 points on GPT-generated data versus 30.8 on its own. Both generators yield mono-

tonic degradation for all three solvers, so the qualitative trend is not GPT-4o-specific. The different magnitudes nevertheless show that generator choice affects the resulting distribution.

Repeated-generation stability. For API-based regeneration, aggregate stability is more meaningful than item-level textual identity. We therefore repeat generation three times on matched roots for each backbone. Table 19 shows modest within-generator variation.

Generator	Runs	Label preservation	Fact-count SD	Accuracy SD
GPT-4o	3	95.1±0.8%	0.6	1.1
Qwen2.5-72B	3	93.2±1.1%	0.8	1.5

Table 19: Aggregate stability over three generation runs.

In addition, we conduct a small-scale manual audit following Section 4.6. As Table 20 shows, Qwen-generated data have somewhat lower label preservation, factual coverage, and readability than GPT-generated data but retain strong aggregate quality. Together, the cross-generator and repeated-run results mitigate dependence on a single generator while showing that the generator remains a distributional factor.

Model	Label (%)	Conflict (%)	Redun. (%)	Fact. (%)	Read.
GPT-4o	96.7	4.5	11.7	99.4	4.0
Qwen	93.3	5.0	5.3	91.0	3.5

Table 20: Human audit on GPT- and Qwen-generated data.

K Case Study

The following example shows the 3-level augmentation process of an example from SNLI with label *entailment*.

Original. Premise: A woman in a blue coat is looking through the front window of a small bookstore on a city street.

Hypothesis: A woman is looking through a bookstore window.

Level 1. Facts:

- Bookstores sell books to the public.
- Independent bookstores are often located in shopping districts or neighborhood streets.
- Storefront windows are commonly used to display merchandise to pedestrians.
- Window displays can include books, signs, and posters for customers outside the shop.

Premise: A woman in a blue coat is looking through the front window of a small bookstore on a city street. The bookstore sells books to the public, and its storefront window displays books, signs, and posters for pedestrians outside the shop.

Hypothesis: A woman is looking through a bookstore window that is used to display books and signs to people outside the shop.

Level 2. Facts:

- Many bookstores organize shelves by genre or subject.
- Bookstores may advertise author readings, book signings, or local literary events.
- Posters in shop windows are used to announce upcoming events.
- A display table can feature new releases or recommended books.
- Customers often browse bookstore displays before entering the store.

Premise: A woman in a blue coat is looking through the front window of a small bookstore on a city street. The window displays books, signs, and posters for pedestrians, including a poster announcing an upcoming author reading. Inside the shop, a display table features new releases, and the shelves are organized by genre for customers who browse before entering.

Hypothesis: A woman is looking through a bookstore window that displays books and a poster for an upcoming author reading, while new releases are arranged on a table inside the shop.

Level 3. Facts:

- Book signings allow readers to meet an author and have books signed.
- Reading groups and book clubs often meet in bookstores or community spaces.
- A chalkboard sign outside a store can list daily recommendations.
- Some bookstores provide chairs or small reading areas for visitors.

Premise: A woman in a blue coat is looking through the front window of a small bookstore on a city street. The window displays books, signs, and posters for pedestrians, including a poster announcing an upcoming author reading and book signing. A chalkboard sign outside lists daily recommendations, and inside the shop, a display table features new releases near shelves organized by genre, with a few chairs set aside as a small reading area for visitors.

Hypothesis: A woman is looking through a bookstore window with books, signs, and a poster for

an author reading and book signing, while the shop also has new releases on display and a small reading area inside.

L Human Audit

L.1 Fact-visible Annotation Guideline

Fact-visible Evaluation Criteria

Purpose. Evaluate the *enhanced* NLI items for *absolute* quality. Judge: (i) the NLI relation, (ii) factual correctness *with respect to the provided fact list*, (iii) internal consistency, (iv) redundancy, and (v) readability. **Do not** use outside sources or personal knowledge.

Materials per item.

- **Sample texts:** a *premise* and a *hypothesis*.
- **Fact list:** 3–15 *atomic facts* (short, self-contained statements) used to construct or justify the sample.

Core labeling principles (from standard NLI practice).

1. **Judge the hypothesis relative to the premise**, not real-world truth.
2. **Use only minimal, text-licensed inference:** paraphrase, synonymy, simple hypernymy/hyponymy, obvious arithmetic, and straightforward temporal/order reasoning.
3. **Use the fact list as the only external support.** If a central claim is not supported or clearly entailed by the facts, mark it *unsupported*.
4. **Prefer certainty over plausibility.** If the premise does not guarantee truth or falsity, choose **Neutral**.

How to read.

1. Read the premise and hypothesis carefully.
2. Read the fact list: skim once for coverage, then re-read to verify specific claims in the texts against the facts.
3. **Factuality rule:** judge factuality *only* against the fact list. New content not supported (or clearly entailed) by the listed facts is *unsupported*.

What to answer (absolute, five-part judgment).

1. **Absolute NLI label (E/N/C):** choose **Entailment** (premise makes the hypothesis certainly true), **Contradiction** (certainly false), or **Neutral** (neither guaranteed nor refuted).
2. **Internal contradiction (yes/no):** mark “yes” if the premise and hypothesis (or content inside them) contradict themselves or each other, independent of the fact list.
3. **Redundancy (none/some/many):** rate repetition or trivial paraphrase without a new reasoning step: **None** (no noticeable repetition), **Some** (occasional), **Many** (frequent, length-inflating).
4. **Factuality w.r.t. fact list (supported/partially/unsupported):** check each *central claim* against the facts only.
5. **Readability (Likert 1–5):** 1 (unreadable), 2 (poor), 3 (adequate), 4 (good), 5 (clear and natural).

Allowed vs. not allowed.

- **Allowed:** paraphrase, synonymy, simple hypernymy/hyponymy, basic arithmetic, direct temporal reasoning, transitivity—when licensed by the text/facts.
- **Not allowed:** web search, outside knowledge, multi-hop world knowledge not in the facts, speculative assumptions.

Submission checklist (per item).

- Absolute NLI label (E/N/C).
- Internal contradiction: yes/no (+ short note if “yes”).
- Redundancy: none/some/many (+ brief reason if “many”).
- Factuality vs. fact list: supported/partially/unsupported (cite fact IDs if helpful).
- Readability: 1–5.

Reference Examples

A. NLI label (E/N/C) examples

- **Entailment (E).**

Premise: “The match was postponed due to heavy rain.”

Hypothesis: “Weather caused the match to be delayed.”

- **Contradiction (C).**

Premise: “The museum is closed on Mondays.”

Hypothesis: “The museum is open every Monday.”

- **Neutral (N).**

Premise: “A chef entered the kitchen.”

Hypothesis: “The chef prepared pasta.”

B. Internal contradiction example

- **Yes (contradictory).**

Premise: “The event starts at 8 pm and starts at 9 pm.”

Hypothesis: (any)

C. Redundancy examples

- **None.** “The match took place in Paris.” (no repetition)

- **Some.** “The match took place in Paris, France.” (light paraphrase once)

- **Many.** “The match took place in Paris, which is in France. The game occurred in France, specifically Paris.” (repeated content)

D. Factuality vs. fact list examples

- **Facts:** F1 “The Eiffel Tower is in Paris.” F2 “Roland Garros is a tennis venue in Paris.”

- **Supported.**

Premise: “The final was held at Roland Garros in Paris.”

Hypothesis: “The final took place in Paris.”

Why: Premise aligns with F2; hypothesis follows from the premise.

- **Partially.**

Premise: “The final was held at a major stadium.”

Hypothesis: “The final took place in Paris.”

Why: “Major stadium” is underspecified by F1/F2; location remains unclear.

- **Unsupported.**

Premise: “The final was held in Berlin.”

Hypothesis: “The final took place in Paris.”

Why: Conflicts with the premise and not backed by F1/F2.

E. Readability anchors (Likert)

- **1 (unreadable).** “Win match rain delay heavy because.” (severe errors)

- **3 (adequate).** “The match was delayed due to rain; wording is a bit awkward but clear.”

- **5 (clear).** “Heavy rain delayed the match.” (natural and fluent)

L.2 Human Evaluation Setup

Scope and blinding. Annotators see only the *enhanced* premise and hypothesis plus a compact *fact list* (3–15 atomic facts) for each item. They do *not* see original/root texts, original labels, model predictions, or the system variant (Full vs. ablation). Item order is individually randomized to avoid order and fatigue effects.

Sampling. We draw a stratified sample from the enhanced pool with strata defined by *dataset* (SNLI, MNLI) and *level* $L \in \{1, 2, 3\}$. The sampling frame contains 9,824 SNLI items per level and 19,647 MNLI items per level. Within each (dataset, L) stratum, we uniformly sample without replacement a target of n_s items—SNLI: $n_s=20$ per level; MNLI: $n_s=40$ per level. This yields

$$N_{\text{enhanced}} = (3 \times 20)_{\text{SNLI}} + (3 \times 40)_{\text{MNLI}} = 180.$$

Where feasible, we balance the root label distribution (E/N/C) *at sampling time only* to ensure coverage; annotators are not shown root labels.

Separately, from the NO-FILTER ablation sets we sample $L=3$ only: 20 SNLI items and 40 MNLI items, i.e.,

$$N_{\text{ablation}} = 20_{\text{SNLI}} + 40_{\text{MNLI}} = 60.$$

Thus the combined total is

$$N_{\text{all}} = N_{\text{enhanced}} + N_{\text{ablation}} = 240.$$

L.3 Pair-only Label Validation and Sensitivity Analysis

To match the information available to evaluated models, we reannotate the same 180 Full-framework items under a pair-only protocol. Annotators see only the enhanced premise and hypothesis and assign an absolute E/N/C label; the fact list,

inherited label, model predictions, and system identity remain hidden. Table 21 compares the resulting preservation rates with the original fact-visible protocol. The E/N/C judgments are identical item by item across the two protocols (180/180 agreement).

Protocol	$L=1$	$L=2$	$L=3$
Fact-visible	98.3 (59/60)	96.7 (58/60)	95.0 (57/60)
Pair-only	98.3 (59/60)	96.7 (58/60)	95.0 (57/60)

Table 21: Label preservation under fact-visible and pair-only protocols. Counts are shown in parentheses.

The Wilson 95% confidence intervals for pair-only label preservation are [91.1, 99.7] at $L=1$, [88.6, 99.1] at $L=2$, and [86.3, 98.3] at $L=3$. These estimates provide sample-level validation rather than a guarantee for every generated item.

We further rescore all seven models on the 60 audited $L=3$ items using both the inherited and adjudicated labels. As shown in Table 22, adjudication adds four correct model–item decisions in aggregate, from 235/420 to 239/420, while the descriptive rank order remains unchanged.

Model	Inherited correct/60 (Acc.)	Adjudicated correct/60 (Acc.)	Rank
DeBERTa-v3-large	40/60 (66.7)	41/60 (68.3)	1→1
DeepSeek-V3	37/60 (61.7)	38/60 (63.3)	2→2
RoBERTa-large	36/60 (60.0)	37/60 (61.7)	3→3
Qwen2.5-14B	34/60 (56.7)	35/60 (58.3)	4→4
GPT-4o	33/60 (55.0)	33/60 (55.0)	5→5
DeepSeek-R1	30/60 (50.0)	30/60 (50.0)	6→6
Llama-3-8B	25/60 (41.7)	25/60 (41.7)	7→7
Mean	56.0	56.9	unchanged

Table 22: Sensitivity of $L=3$ results to adjudicated labels on the 60-item audited subset.

This small-scale analysis supports ordering stability on the audited subset, while close model gaps should not be treated as statistically conclusive.