BEYOND ATOMIC CLAIMS: PROBABILISTIC LOGICAL COHERENCE FOR LONG-FORM LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form factuality assessment typically follows a *decompose-then-verify* paradigm: a response is decomposed into atomic claims, each claim is verified against evidence, and the resulting judgments are aggregated into a final score. While effective, this approach evaluates claims in isolation and cannot assess whether they form a coherent argument. We formalize *logical coherence* as a property of the joint distribution over claim truth values induced by a Markov random field whose factors encode discourse relations expressed within the response. To bridge discourse and inference, we propose a two-level taxonomy that maps twelve PDTB discourse senses onto five inferential couplings. We also introduce LOCOBENCH, a benchmark that holds claims fixed while varying only their relational structure. Using gold relation graphs, our logical coherence scores satisfy 82.7% of 202 ordering constraints, compared with 34.7–37.1% when relations are extracted automatically, identifying relation extraction as the primary bottleneck. Across various baselines, our method is the only one that both orders responses by coherence and remains invariant under meaning-preserving edits. Furthermore, it localizes incoherence through claim-level posteriors rather than just detecting it.

1 INTRODUCTION

The dominant paradigm for evaluating the factuality of long-form generation is *decompose-then-verify*: a response is decomposed into atomic claims, each claim is verified against retrieved evidence, and the resulting verdicts are aggregated into a final score (Min et al., 2023; Wei et al., 2024; Song et al., 2024; Marinescu et al., 2025). Although highly effective and now standard, this framework implicitly assumes that claims can be evaluated in isolation. In doing so, it discards the relational structure that transforms a collection of statements into a coherent argument.

This limitation manifests in at least two ways. First, a response may be internally inconsistent. For example, a report may assert both that no passengers were harmed and that a news service reported three deaths. Claim-level verification identifies one claim as true and the other as false, but it does not capture the more fundamental defect: the response contradicts itself. Second, the inferential relations between claims may be invalid even when the claims themselves are true. A sequence of factually correct statements connected by an unsupported "therefore" does not constitute a sound argument, yet every constituent claim may verify successfully. These examples highlight that factuality and logical coherence are distinct and complementary dimensions of response quality. Neither can be reduced to the other: a response may consist entirely of true claims while remaining logically incoherent, or it may exhibit coherent reasoning despite being factually incorrect. More recently, (Xie et al., 2026) showed empirically that well established fact-checkers like SAFE (Wei et al., 2024), VeriScore (Song et al., 2024), and FactCheck-GPT (Wang et al., 2024) fail to capture the logical dimension of the response in 60%, 78%, and 48% of cases, respectively.

Contributions. The limitations of claim-level factuality assessment motivate a complementary question: *to what extent is the content of a response supported by the argument it constructs as a whole?* We address this question by formalizing logical coherence as a property of the joint distribution over claim truth values induced by a Markov random field (MRF), where variables represent claims and factors encode discourse relations expressed within the response. Our contributions are fourfold. First, we introduce a probabilistic framework for logical coherence together with a two-level discourse taxonomy. Inspired by PDTB 3.0 (Rashmi et al., 2019) and RST (Mann & Thompson, 1988), we

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

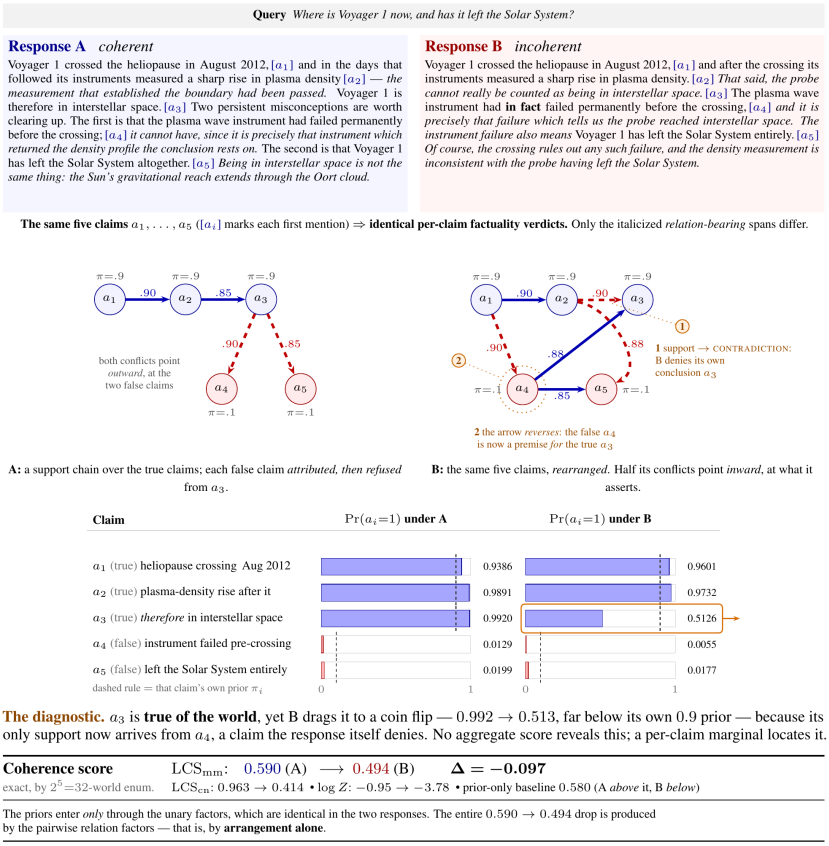


Figure 1: A simple example showing identical claims but different coherence. Both responses assert the same five claims and receive the same claim-level factuality score, yet differ substantially in coherence due to the relations established between those claims.

define twelve discourse senses that compile deterministically into five inferential couplings, yielding an interpretable discourse layer while preserving a compact pairwise MRF. Second, we characterize the factorization required for coherence modeling. We derive coherence-specific factor tables for the inferential couplings and show that factor tables commonly used in claim–evidence factuality assessment are unsuitable for coherence, substantially compressing the metric's dynamic range. Third, we identify desiderata for coherence scoring and develop coherence scores defined on the induced joint distribution, including scores that remain monotonic under conflict resolution. We further show how coherence composes naturally with existing factuality pipelines through claim-level priors. Finally, we introduce LoCoBench, a benchmark with construction-defined coherence orderings and gold relation graphs that disentangles relation-extraction error from coherence-scoring error. Using identical scoring procedures while varying only the relation graph, our method satisfies 82.7% of 202 coherence constraints with gold relations, compared with 34.7–37.1% when relations are extracted automatically, identifying relation extraction as the principal bottleneck. Across five baseline families, our approach is the only one that both ranks responses by coherence ($37/50$ versus $30/50$ for the strongest baseline) and remains invariant under meaning-preserving edits ($20/20$). Additional details and analyses are provided in the Appendix.

2 Probabilistic Logical Coherence

In this section, we formalize logical coherence as a probabilistic property of the dependencies among claims in a response. To this end, we introduce a pairwise Markov random field grounded in a two-level taxonomy of discourse relations, characterize the factor tables required for coherence assessment, and derive a coherence score that integrates naturally with claim-level factuality evaluation.

2.1 A MOTIVATING EXAMPLE

We begin with a motivating example. Figure 1 presents two responses to a query about Voyager 1. Both responses contain the same five claims about the probe: three true and two false. Consequently, any factuality metric based on independent claim verification assigns them identical scores, since every claim-level judgment is the same, yielding an aggregate factuality score of $3/5$ ($=0.6$).

The difference lies entirely in the relations between claims. In Response A, the two false claims are explicitly attributed and then refuted, yielding a coherent argument. In Response B, by contrast, the false claim a_4 ("the plasma wave instrument had failed") is used to support the true claim a_3 ("therefore in interstellar space"), while a_3 is contradicted elsewhere in the response and both a_4 and a_5 are ultimately rejected. Thus, although the two responses contain identical claims, they induce fundamentally different argument structures. Only two changes in the relation graph (callouts **1** and **2** in Figure 1) account for the difference: the support relation $a_2 \to a_3$ becomes a CONTRADICTION, and the edge between a_3 and a_4 reverses direction. What Response A uses to reject a falsehood, Response B uses to justify a conclusion. As a result, Response A attains a score of 0.590, compared with 0.494 for Response B. Since the claim priors are identical in both cases, the entire drop is attributable to the arrangement of relations rather than the claims themselves. Crucially, the score not only detects incoherence but also localizes it. The true claim a_3 receives a posterior of 0.992 in Response A but only 0.513 in Response B, despite having a prior of 0.9. Its sole supporting premise is a claim that the response itself later denies, preventing support from propagating through the graph. In other words, coherence is compromised not by mentioning a falsehood but by reasoning from it.

2.2 THE COHERENCE MARKOV RANDOM FIELD

We now formalize logical coherence using a Markov Random Field (MRF) (Koller & Friedman, 2009) that provides a probabilistic encoding of the relations between the claims of a response.

Definition 1 (Coherence MRF)**.** *Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ be the atomic claims of a response in assertion order, each binary with $a_i = 1$ meaning "the claim holds". Let $\mathcal{R} = \{(s_r, t_r, \tau_r, p_r)\}_{r=1}^{|\mathcal{R}|}$ be a set of* relations*, where $s_r \neq t_r$ index claims, τ_r is an* inferential coupling *and $p_r \in [0,1]$ is the relation's* weight*. Given priors $\pi \in [0,1]^n$, the* coherence MRF *induced by $(\mathcal{A}, \mathcal{R}, \pi)$ is:*

$$\mathbb{P}(\mathbf{a}) \;=\; \frac{1}{Z} \prod_{i=1}^{n} \phi_i(a_i) \prod_{r \in \mathcal{R}} \psi_{\tau_r, p_r}(a_{s_r}, a_{t_r}), \qquad Z \;=\; \sum_{\mathbf{a} \in \{0,1\}^n} \prod_i \phi_i(a_i) \prod_r \psi_r, \tag{1}$$

where Z is a normalization constant known as the partition function*, $\phi_i(a_i) = [1 - \pi_i,\ \pi_i]$ over the states $(0, 1)$ of a_i, while $\psi_{\tau,p}$ is called a* potential *and is given in Section 2.4, respectively.*

We note that coherence depends on claim order only through the induced relation graph: Eq. 1 contains no notion of position, so any edit that preserves the graph leaves the induced distribution and all coherence scores unchanged. Furthermore, coherence can be evaluated independently of external evidence: under uniform priors ($\pi_i \equiv \frac{1}{2}$), the model assesses only the internal support structure of the response. Thus, coherence and factuality remain distinct yet complementary dimensions that can be combined within a unified framework as shown later in Section 2.7.

2.3 TWO LEVELS: INFERENTIAL COUPLINGS AND DISCOURSE SENSES

Definition 2 (Inferential couplings)**.** *A* coupling *τ is specified by the set $V(\tau) \subseteq \{0,1\}^2$ of joint assignments (a_s, a_t) it* penalizes *i.e., the configurations at which the relation is violated:*

Coupling	*Reading*	*Directed*	$V(\tau)$
ENTAILMENT	$a_s \Rightarrow a_t$	*yes*	$\{(1,0)\}$
CONTRADICTION	$a_s \Rightarrow \neg a_t$	*yes*	$\{(1,1)\}$
EQUIVALENCE	$a_s \Leftrightarrow a_t$	*no*	$\{(0,1),(1,0)\}$
EXCLUSIVE	*exactly one of a_s, a_t holds*	*no*	$\{(0,0),(1,1)\}$
CO_NECESSITY	*at least one of a_s, a_t holds*	*no*	$\{(0,0)\}$
NONE	*no coupling*	—	*$\varnothing$ (no factor)*

In addition to the standard NLI relations (Bowman et al., 2015) we consider EQUIVALENCE, EXCLUSIVE, and CO_NECESSITY, respectively. While CONTRADICTION excludes only the both-true

assignment, EXCLUSIVE also excludes the both-false assignment, capturing *exhaustive* alternatives. For instance, "no one was harmed" and "three people died" can neither both hold nor both fail. Conversely, CO_NECESSITY is the dual coupling, excluding only the both-false assignment and thus capturing disjunction. Together, CONTRADICTION, EXCLUSIVE, and CO_NECESSITY provide three distinct ways of constraining compatibility between claims.

Logical coherence is expressed through discourse relations, whereas couplings provide their probabilistic realization (Mitchell et al., 2022; Marinescu et al., 2025). We therefore propose a two-level representation in which discourse senses are deterministically compiled into couplings and strength priors via a mapping $\mathcal{C}$. This preserves an interpretable discourse vocabulary while yielding a compact inferential model. A strength prior of $\perp$ indicates that the coupling strength is estimated directly.

Level-2 discourse sense (PDTB)	$\mathcal{C}$ coupling	prior	note
Cause-Effect, Effect-Cause, Evidence	ENTAILMENT	$\perp$	directed
Condition, Instantiation	ENTAILMENT	$\perp$	directed
Restatement	EQUIVALENCE	0.90	definitional identity
Contrast	CONTRADICTION	$\perp$	opposed, *not* exhaustive
Concession	CONTRADICTION	$\perp$	a tension the text may resolve
Alternative	EXCLUSIVE	$\perp$	exhaustive: exactly one holds
Disjunction	CO_NECESSITY	$\perp$	at least one holds
Precedence, Succession	NONE	—	**ordering-only: no factor**

Importantly, a Concession introduces a conflict only to resolve it. Penalizing it as a standard contradiction would therefore mischaracterize the relation. When a resolving *holding* claim h is present, we accordingly discount the conflict weight: $p_r^{\text{eff}} = p_r(1 - \lambda\,\pi_h)$, with $\lambda = 0.45$.

2.4 The Pairwise Potentials and Why The Factuality Model is Wrong Here

Let τ be a coupling between a source claim a_s and a target claim a_t. The potential $\psi_{\tau,p}(a_s, a_t)$ is a 2×2 table indexed by the joint assignments $(a_s, a_t) \in \{0,1\}^2$, ordered as $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. Here, p is the coupling strength and π_s is a fixed parameter of the potential and it is independent of the claim priors π_i from Definition 1. We assume hereafter that $\pi_s = \frac{1}{2}$.

ENTAILMENT	CONTRADICTION	CO_NECESSITY	EQUIVALENCE	EXCLUSIVE
$[1-\pi_s, \pi_s, \mathbf{1}-\mathbf{p}, p]$	$[1-\pi_s, \pi_s, p, \mathbf{1}-\mathbf{p}]$	$[\mathbf{1}-\mathbf{p}, \pi_s, \pi_s, p]$	$[p, \mathbf{1}-\mathbf{p}, \mathbf{1}-\mathbf{p}, p]$	$[\mathbf{1}-\mathbf{p}, p, p, \mathbf{1}-\mathbf{p}]$

Factuality models instead use *no-prior* potentials, such as $[p, p, \mathbf{1} - \mathbf{p}, p]$ for ENTAILMENT (Marinescu et al., 2025). These coincide with our coherence potentials when the source claim is true, but differ when it is false: coherence yields the uninformative row $[\frac{1}{2}, \frac{1}{2}]$, whereas factuality yields $[p, p]$. This difference reflects a change in semantics. In factuality assessment, the source is external evidence whose reliability itself provides information about the claim. In coherence assessment, the source is another claim in the same response, so the relation should constrain the target conditional on the source rather than contribute independent support for it. Therefore, the consequence is significant. Consider a single ENTAILMENT edge $a_1 \to a_2$ with $p = 0.9$ and priors $\pi_1 = \pi_2 = 0.9$. Under the coherence model, we can show that the source marginal increases to $\mathbb{P}(a_1{=}1) = 0.9365$, reflecting the fact that a claim cited as a premise is reinforced by the argument structure. Under the factuality model, however, the same marginal decreases to 0.8913, falling below its prior. The source claim is effectively penalized for being used as a premise, a behavior that is appropriate for uncertain external evidence but undesirable for coherence assessment (see also Appendix C for more details).

2.5 Logical Coherence Scores

We next define a *Logical Coherence Score* (LCS) as a functional of the joint distribution $\mathbb{P}(\mathbf{a})$ induced by Eq. 1. The score should be bounded in $[0, 1]$ and interpretable, monotonic under conflict resolution, sensitive to logical inconsistencies without ad hoc parameters, and computable using standard probabilistic inference. Most importantly, it should quantify coherence as a property of the inferred distribution itself. In contrast, a score based solely on the relation graph $\mathcal{R}$ would measure characteristics of the graph itself (e.g., number of claims, connectivity of the graph, number and types of relations, etc.) rather than the coherence of the argument encoded by the model.

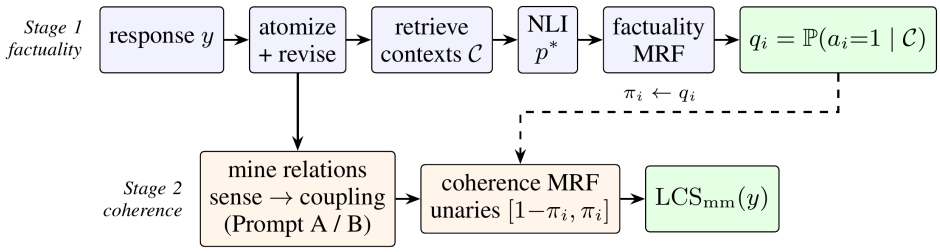


Figure 2: A two-stage pipeline that combines logical coherence with factuality.

Given a response y with atomic claims $\{a_1, \ldots, a_n\}$ and its corresponding coherence MRF, the *mean-marginal Logical Coherence Score* $\mathrm{LCS}_m m(y)$ is defined by:

$$\mathrm{LCS}_{\mathrm{mm}}(y) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(a_i = 1)\,. \tag{2}$$

The score admits a natural interpretation as the expected fraction of claims upheld by the response's argument as a whole. Conflicts reduce the posterior support of the claims they involve, whereas entailment propagates support through the graph; claims disconnected from the relation structure retain their prior beliefs. For completeness, in Appendix D we define three alternative scores: a two-term consistency probability $\mathrm{LCS}_{\mathrm{cn}}(y)$, a reified-node score $\mathrm{LCS}_{\mathrm{re}}(y)$, and a normalized log-partition score $\mathrm{LCS}_{\mathrm{lp}}(y)$. Among the four candidates, only $\mathrm{LCS}_{\mathrm{mm}}(y)$ and $\mathrm{LCS}_{\mathrm{lp}}(y)$ are monotonic under conflict resolution, making them the only scores that satisfy all of our desiderata.

2.6 Estimating Relations from the Response

So far, we have assumed that the relation set $\mathcal{R}$ is given. In practice, relations must be extracted from the response itself. For a relation $r = (a_s, a_t, \tau)$, we decompose its weight as $p_r = \mathbb{P}(\tau \mid a_s, a_t, y) \cdot \mathbb{P}(a_t \mid a_s, \tau, y)$, separating uncertainty about the relation type from uncertainty about its strength.

We estimate these quantities using a two-prompt LLM-based relation miner. Specifically, Prompt A predicts the discourse sense between claims, which is deterministically mapped to a coupling, while Prompt B estimates relation strength from a binary plausibility judgment. Following Kadavath et al. (2022), strength is computed from the re-normalized probability of an affirmative response; when token probabilities are unavailable, we use sampling-based confidence estimates (Bhattacharjya et al., 2025). Additional details on relation extraction are provided in Appendix A and E, respectively.

A naive ALL_PAIRS strategy considers all $n(n-1)$ ordered claim pairs, leading to quadratic complexity and many spurious relations. We therefore propose three alternative candidate-generation policies: WINDOWED, which considers forward pairs within a distance w ($w = 4$ in our experiments); SEM-SIM, which augments WINDOWED with semantically similar long-range pairs; and BIDIRECTIONAL, which considers both directions within the window. The latter addresses a structural limitation of forward-only extraction, since directed relations may point backward in claim order.

2.7 Composing Coherence with Factuality

The claim priors π_i in Definition 1 provide a natural interface to factuality assessment. Given factuality estimates $q_i = P(a_i = 1 \mid \mathcal{C})$ from an external system such as FactReasoner (Marinescu et al., 2025), we set $\pi_i \leftarrow q_i$. Figure 2 shows the resulting two-stage pipeline: Stage 1 estimates claim factuality from retrieved evidence, while Stage 2 extracts discourse relations and constructs a coherence MRF over the same claims. The stages interact only through the priors.

This design separates factuality from coherence while allowing them to compose naturally. Uniform priors ($\pi_i \equiv \frac{1}{2}$) recover a purely coherence-based assessment, whereas factuality-derived priors incorporate external evidence. Since priors are applied only at scoring time, the same relation graph can be reused under different prior settings. Consequently, a claim receives high support only when it is both externally plausible and internally supported by the response's argument.

3 THE LOCOBENCH BENCHMARK

Evaluating coherence requires responses that vary in argument structure while holding claim content fixed, together with gold relation graphs specifying the intended discourse structure. To this end, we introduce LOCOBENCH, a synthetic benchmark of *coherence ladders*. A ladder consists of five responses (also called *rungs*) built from the same claim set and topic but differing along a single structural dimension, such as conflict resolution, entailment-chain integrity, or claim ordering. Each response is paired with its own gold relation graph and a set of expected ordering constraints (see Appendix F), enabling coherence measures to be evaluated independently of relation extraction.

Specifically, LOCOBENCH contains four ladder families. CONFLICT ladders progressively resolve contradictions while holding all claims fixed, yielding a strictly increasing coherence ordering. CHAIN ladders break and repair entailment links, testing sensitivity to missing support rather than explicit contradiction. ORDER ladders apply meaning-preserving reorderings and therefore require invariance, while CONTROL ladders perform edits that leave the relation graph unchanged and should produce identical scores throughout. Together, these families distinguish genuine coherence measures from metrics that merely track contradictions, surface edits, or response length.

The benchmark is generated through a five-stage pipeline: topic selection, claim generation, relation-plan construction, response realization, and controlled perturbation. Starting from a pool of atomic claims, we construct a relation plan containing discourse relations, non-relations, validity labels, and target strengths. Responses are then generated to realize the plan, and perturbation operators modify only the intended relations while preserving the underlying claim content. The resulting gold graph for each rung is obtained by applying the same perturbations directly to the relation plan.

To ensure quality, every ladder is validated by an independent multi-model committee distinct from the generator. Validation requires accurate recovery of planned relations, complete claim coverage, and high-quality prose free of annotation leakage. Ladders are admitted or rejected as a whole, ensuring that all reported comparisons remain meaningful. The released corpus contains 14 ladders (70 responses) spanning four domains, with 1,115 claims, 603 edge-producing gold relations, and 202 defined ordering constraints.

Human study. To validate LOCOBENCH, four annotators rated 14 response pairs drawn from seven of the benchmark's fourteen families, yielding 56 judgments in a fully crossed design ($\alpha = 0.748$). The following findings are particularly relevant for interpreting the main results from Section 4. First, the constructed coherence orderings are meaningful to human readers: annotators reproduced 8 of 10 correct orderings of the response pairs by majority vote and 32 of 40 pairwise comparisons overall, with their explanations predominantly citing relational phenomena such as contradiction, unsupported conclusions, and broken reasoning chains rather than surface fluency. This suggests that the 82.7% agreement achieved by the gold-graph model and reported in Section 4 reflects human notions of coherence rather than artifacts of the generation process. Second, the two invariance families behave differently. Annotators largely agreed that CONTROL ladders were equivalent (6 of 8 judgments), but consistently distinguished ORDER ladders (0 of 8), often citing disrupted anaphora and discourse flow after sentence reordering. Since such effects are not represented in the relation graph, the model's perfect ORDER invariance reported in Section 4 should be interpreted as a limitation of its scope rather than evidence that sentence order never affects coherence.

Appendix F includes additional details about the LOCOBENCH construction and human evaluation.

4 EXPERIMENTS

In this section, we empirically evaluate the proposed logical coherence framework. In addition to LOCOBENCH, we consider five datasets commonly used for long-form factuality assessment: BIO, ASKHIST, BOOKS, ELI5, and LFOBJ (Min et al., 2023; Wei et al., 2024; Marinescu et al., 2025). This allows us to assess coherence both in a controlled setting (LOCOBENCH) and on realistic long-form generations from diverse domains.

We consider five families of baselines, as follows: (**F1**) *Lexical cohesion* – DiscoScore's reference-free variants RC, LC, and ENTITY GRAPH (Zhao et al., 2023), which score noun repetition and entity continuity across sentences but do not model entailment, contradiction, or discourse relations; (**F2**) *Local inconsistency detection* – CONTRADICTION RATE, defined as one minus the fraction of

Table 1: Results on LoCoBench dataset using GOLD and MINED relation graphs.

Measure	**Increase** (of 50)		**Invariance** (of 20)	
	`llama`	`gpt-oss`	`llama`	`gpt-oss`
a. *LCS on* GOLD *relation graphs*				
$\mathrm{LCS}_{\mathrm{mm}}$	37 (**74**%)		20 (**100**%)	
$\mathrm{LCS}_{\mathrm{lp}}$	36 (72%)		20 (100%)	
$\mathrm{LCS}_{\mathrm{cn}}$	34 (68%)		20 (100%)	
b. *baselines: model-free*				
RC (cohesion)	30 (60%)		16 (80%)	
LC (cohesion)	28 (56%)		16 (80%)	
LENGTH (control)	20 (40%)		12 (60%)	
ENTITY GRAPH	0 (0%)		20 (100%)†	
CLAIM COUNT (control)	0 (0%)		20 (100%)†	
c. *LCS on* MINED *relation graphs — rows are the candidate policy, sub-columns the miner*				
WIN / $\mathrm{LCS}_{\mathrm{mm}}$	20 (40%)	25 (50%)	0 (0%)	0 (0%)
WIN / $\mathrm{LCS}_{\mathrm{lp}}$	35 (70%)	29 (58%)	0 (0%)	0 (0%)
WIN / $\mathrm{LCS}_{\mathrm{cn}}$	23 (46%)	25 (50%)	0 (0%)	0 (0%)
BID / $\mathrm{LCS}_{\mathrm{mm}}$	28 (56%)	21 (42%)	0 (0%)	0 (0%)
BID / $\mathrm{LCS}_{\mathrm{lp}}$	25 (50%)	28 (56%)	0 (0%)	0 (0%)
BID / $\mathrm{LCS}_{\mathrm{cn}}$	34 (68%)	24 (48%)	0 (0%)	0 (0%)
d. *baselines: call an LLM — both axes move with the backend*				
ROSCOE-SC (mean)	16 (32%)	17 (34%)	9 (45%)	5 (25%)
CONTRADICTION RATE	12 (24%)	12 (24%)	9 (45%)	8 (40%)
JUDGE, G-Eval	0 (0%)	21 (42%)	13 (65%)	5 (25%)
JUDGE, direct rating	4 (8%)	18 (36%)	14 (70%)	5 (25%)
ROSCOE-SC (max)	0 (0%)	0 (0%)	20 (100%)†	20 (100%)‡

WIN/BID abbreviate the WINDOWED and BIDIRECTIONAL candidate policies. † vacuous: the measure never varies *within* a family, so it cannot fail an invariance pair.

claim pairs classified as contradictory by an NLI model (Manakul et al., 2023), measuring explicit pairwise contradictions only; (**F3**) *Reasoning consistency* – ROSCOE Self-Consistency (ROSCOE-SC) (Golovneva et al., 2023), reported in both its original max-based form and a mean-based variant, which aggregates pairwise contradiction probabilities between claims but does not model relation types or global argument structure; (**F4**) *LLM judges* – G-Eval's coherence dimension (Liu et al., 2023; Fabbri et al., 2021) and a direct coherence-rating prompt, denoted G-EVAL and JUDGE, respectively, which assess coherence through holistic LLM judgments rather than an explicit structural representation; (**F5**) *Control baselines* – CLAIM COUNT and response LENGTH, which should not track coherence and are included to detect corpus-level confounders (Steen & Markert, 2022).

In all experiments, we consider two relation sources: GOLD, which uses the annotated relation graph available in LoCoBench, and MINED, which uses relations extracted from the response by the relation miner of Section 2.6. For the MINED source, we report results using both `llama-3.3-70b` (Touvron et al., 2023) and `gpt-oss-120b` (OpenAI et al., 2025) open-weight models[1].

Results on the LoCoBench corpus. In Table 1 we evaluate two complementary properties, namely whether a metric correctly ranks responses with increasing coherence (*Increase* axis - 50 assertions), or remains unchanged under meaning-preserving edits that preserve the underlying relation graph (*Invariance* axis - 20 assertions). A useful coherence measure should satisfy both. We see again that using GOLD relation graphs, LCS is the only method that performs strongly on both axes. The mean-marginal score $\mathrm{LCS}_{\mathrm{mm}}$ correctly satisfies $37/50$ ordering constraints while remaining perfectly invariant ($20/20$), and the other LCS variants exhibit similar behavior. No baseline matches this combination. The strongest baseline, the lexical cohesion metric RC, reaches $30/50$ on ordering but drops to $16/20$ on invariance, indicating sensitivity to surface form rather than argument structure. Contradiction-based measures capture explicit conflicts but miss missing-support failures, while ROSCOE's max-based aggregation collapses entirely ($0/50$). LLM judges are the least stable: G-Eval ranges from $0/50$ to $21/50$ on ordering depending solely on the judge model, despite identical inputs

[1]The models were deployed on A100 GPUs and served using vLLM with greedy decoding (temperature 0).

Table 2: Results on the factuality datasets using MINED relation graphs.

	`llama-3.3-70b`			`gpt-oss-120b`		
Measure	mean	sd	d	mean	sd	d
LCS: WINDOWED *candidate policy*						
$\mathrm{LCS}_{\mathrm{mm}}$	0.976	0.069	90	0.948	0.090	148
$\mathrm{LCS}_{\mathrm{lp}}$	0.685	0.450	45	0.671	0.433	85
$\mathrm{LCS}_{\mathrm{cn}}$	0.803	0.246	74	0.711	0.253	134
LCS: BIDIRECTIONAL *candidate policy*						
$\mathrm{LCS}_{\mathrm{mm}}$	0.980	0.059	79	0.960	0.078	111
$\mathrm{LCS}_{\mathrm{lp}}$	0.661	0.465	35	0.620	0.460	58
$\mathrm{LCS}_{\mathrm{cn}}$	0.804	0.241	59	0.721	0.251	87
baselines that call an LLM						
CONTRADICTION RATE	0.997	0.008	48	0.996	0.009	54
ROSCOE-SC (mean)	0.996	0.011	73	0.994	0.012	69
ROSCOE-SC (max)	0.625	0.467	51	0.564	0.496	**2**
JUDGE, G-Eval	1.000	0.003	**5**	0.917	0.101	**10**
JUDGE, direct	1.000	0.000	**3**	0.829	0.162	16

model-free baselines	mean	sd	d
RC (cohesion)	0.502	0.180	224
LC (cohesion)	0.184	0.064	209
ENTITY GRAPH	0.972	0.119	17
LENGTH (control)	0.207	0.046	136

and rubric. A similar instability can be observed on the *Invariance* dimension. In addition, Table 19 in Appendix evaluates whether the proposed coherence scores satisfy the ordering constraints defined by the dataset. We can see that with gold relation graphs, the scores satisfy 167 of 202 constraints (82.7%), including perfect performance on the ORDER and CONTROL ladders. Replacing gold graphs with mined ones reduces performance to 34.7–37.1%, indicating that relation extraction rather than probabilistic scoring is the dominant source of error.

The comparison reveals complementary failure modes. Cohesion metrics partially track coherence because coherent responses often reuse entities consistently, but they lack any representation of contradiction or support. Contradiction-based measures detect local inconsistencies but cannot reason about entailment chains. LLM judges can sometimes order responses effectively, but improvements on the *Increase* axis are often accompanied by losses on *Invariance*, making their behavior highly backend-dependent. Several baselines achieve perfect invariance, but only vacuously because they produce nearly constant scores and therefore fail to distinguish coherent from incoherent responses.

Replacing gold graphs with MINED relations changes the picture. On the *Increase* axis, several mined-graph configurations remain competitive, with `llama`/WINDOWED $\mathrm{LCS}_{\mathrm{lp}}$ reaching $35/50$ and `llama`/BIDIRECTIONAL $\mathrm{LCS}_{\mathrm{cn}}$ reaching $34/50$, outperforming most baselines. However, every mined configuration scores $0/20$ on *Invariance*, whereas all gold-graph configurations remain at $20/20$. Since the underlying responses differ only by meaning-preserving edits, this failure cannot be attributed to the coherence model itself; it reflects the lack of performance of current relation extraction pipelines. The overall conclusion is therefore twofold: a probabilistic coherence model is necessary to capture properties beyond cohesion, contradiction detection, and holistic judging, but the main bottleneck of the end-to-end pipeline is accurate and stable extraction of discourse relations.

Results on the factuality corpus. To assess discriminative power on realistic generations, we apply the full two-stage pipeline to a stratified sample of 250 responses drawn from the five long-form factuality benchmarks (50 responses each). Since these datasets do not provide coherence labels, we evaluate a necessary condition for usefulness: the ability of a measure to assign distinct scores to different responses. Table 2 reports, for each measure, the mean, standard deviation, and the number of distinct values (d) produced across the 250 responses. The results reveal a striking contrast. Both LLM judges saturate on natural text, producing only 3–5 distinct scores under `llama-3.3-70b`, while ROSCOE's published max-based aggregation collapses to just two values under `gpt-oss-120b`. In contrast, every LCS variant remains highly discriminative, with at least 35 distinct values across all score, candidate policy, and backend combinations, and substantially

more in most settings. The effect is consistent across miners and candidate policies: BIDIRECTIONAL extraction reduces discrimination relative to WINDOWED, likely due to additional spurious relations, yet still exceeds the best judge configuration. Together with the LOCOBENCH results, these findings expose complementary failure modes of current coherence judges: they are highly unstable on controlled coherence perturbations yet largely saturated on natural prose, whereas the proposed probabilistic scores remain both discriminative and structurally grounded.

5 RELATED WORK

Atomic factuality. FActScore (Min et al., 2023) introduced per-atom precision against a knowledge source; SAFE (Wei et al., 2024) added search-augmented verification and an $F_1@K$ metric, VeriScore (Song et al., 2024) restricted attention to *verifiable* claims, and FLEEK (Bayat et al., 2023) retrieves evidence to detect and correct errors. All aggregate independent per-claim verdicts and are therefore blind to inter-claim structure by construction. FactReasoner (Marinescu et al., 2025) replaced verdict counting with probabilistic inference over a factor graph of claim and evidence variables with pairwise factors derived from NLI relations. Although it is closely related to our work, FactReasoner's graph has no claim–claim edges and its marginals cannot express internal consistency. More recently, Xie et al. (2026) introduced LoReFact to address the gap directly, classifying inter-atom dependencies as a labeling task over claim pairs. In contrast, our proposed approach is to treat them as *factors in a joint distribution*, which lets a local conflict propagate to distant claims and makes the logical coherence score a functional of the whole graph rather than a per-edge average.

Discourse coherence. Entity-grid models (Barzilay & Lapata, 2005) and their neural successors (Li & Jurafsky, 2017) score *local coherence* from entity-mention distributions, evaluated against *synthetic perturbations*. We extend this methodology in our work by making the perturbation an operator on a declared relation graph rather than on raw text, so its intended effect is known per item. DiscoScore (Zhao et al., 2023) uses BERT-based discourse representations. What these share is that coherence is read off surface statistics or learned representations: none exposes an inspectable relation graph, and none yields per-claim attributions. Our proposed inventory of discourse senses is inspired by the PDTB 3.0 (Rashmi et al., 2019) and RST (Mann & Thompson, 1988) frameworks but treats them as an interpretable layer compiling to a smaller inferential vocabulary. Closest to our relation mining stage is Liu & Strube (2025), who label sentence pairs with PDTB 3.0 senses via a discourse parser and develop both neural and prompting architectures that integrate these signals.

Consistency checking and judges. SelfCheckGPT (Manakul et al., 2023) introduces a black-box method for hallucination detection that estimates the factuality of LLM-generated text by measuring the consistency of multiple stochastic samples from the same model using NLI, BERT scores or prompting strategies. ROSCOE (Golovneva et al., 2023) scores reasoning chains with a self-consistency term. G-Eval (Liu et al., 2023) is an LLM judge scoring coherence from a rubric; judges are strong on aggregate correlation and weak on per-claim attribution and on any guarantee that a meaning-preserving edit cannot change the score. Steen & Markert (2022) argue coherence measures must be checked against system-level confounders, which we adopt in our control baselines.

6 CONCLUSION

We introduced a probabilistic framework for logical coherence that models discourse relations as constraints over the joint distribution of claim truth values. To evaluate coherence independently of content, we introduced LOCOBENCH, a benchmark that holds claims fixed while varying only their relational structure. Human evaluation confirms that the benchmark's orderings largely align with human judgments and are driven by relational phenomena such as contradiction and broken support. Using gold relation graphs, we show that the proposed coherence measure satisfies 82.7% of 202 coherence constraints and is the only method that both orders responses by coherence and remains invariant under validated meaning-preserving edits. Performance drops substantially when relations are extracted automatically (34.7–37.1%), identifying relation extraction as the primary bottleneck. Overall, our results suggest that coherence is fundamentally a relational property that cannot be captured reliably by cohesion metrics, contradiction detectors, or holistic LLM judgments alone, and that future progress depends primarily on more accurate discourse-relation extraction.

AI USE STATEMENT

We used generative AI tools to assist with language editing, grammar correction, software development, and benchmark construction. In particular, AI-assisted coding tools were used during the implementation of the proposed algorithms, and large language models were used to generate the synthetic LOCOBENCH benchmark. All AI-generated artifacts, including text, code, and dataset instances, were reviewed, validated, and, where appropriate, extensively tested by the authors prior to inclusion in this work. For example, AI-generated code was manually inspected for correctness and verified through unit and integration testing, while benchmark instances underwent automated and independent validation procedures. The authors take full responsibility for the accuracy, validity, and integrity of all content presented in this paper.

ETHICS STATEMENT

The human study described in Section 4 involved four annotators who compared pairs of machine-generated responses for logical coherence and provided free-text justifications for their judgments. No personal data was collected, and no such data is released. The evaluated material consisted exclusively of LLM-generated expository text on neutral technical topics and contained no content about identifiable individuals. The primary ethical risk is misinterpretation: coherence is not factuality. A response may be internally coherent yet entirely false, and a highly factual response may still be incoherent. For this reason, our framework treats coherence and factuality as distinct dimensions and we discourage the use of LCS as a standalone factuality metric. Indeed, a fluent, internally consistent fabrication may receive a high coherence score. Finally, LOCOBENCH intentionally contains false claims as part of its controlled evaluation design; these claims are explicitly labeled and the benchmark should not be redistributed without the accompanying annotations.

REPRODUCIBILITY STATEMENT

We include in the supplementary material our implementation and benchmark datasets. The code and LOCOBENCH benchmark dataset will be open-sourced if the paper is accepted.

REFERENCES

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer (eds.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 141–148, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219858. URL https://aclanthology.org/P05-1018/.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyi, Samira Khorshidi, Fei Wu, Ihab F. Ilyas, and Yunyao Li. Fleek: Factual error detection and correction with evidence retrieved from external knowledge, 2023. URL https://arxiv.org/abs/2310.17119.

Debarun Bhattacharjya, Balaji Ganesan, Junkyu Lee, Radu Marinescu, Katsiaryna Mirylenka, Michael Glass, and Xiao Shou. Simba uq: Similarity-based aggregation for uncertainty quantification in large language models, 2025. URL https://arxiv.org/abs/2510.13836.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning, 2023. URL https://arxiv.org/abs/2212.07919.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Jiwei Li and Dan Jurafsky. Neural net models of open-domain discourse coherence. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 198–209, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1019. URL https://aclanthology.org/D17-1019/.

Qiang Liu and Alexander Ihler. Bounding the partition function using Hölder's inequality. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

Wei Liu and Michael Strube. Joint modeling of entities and discourse relations for coherence assessment. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21910–21926, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1113. URL https://aclanthology.org/2025.emnlp-main.1113/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988. doi: doi:10.1515/text.1.1988.8.3.243. URL https://doi.org/10.1515/text.1.1988.8.3.243.

Radu Marinescu, Debarun Bhattacharjya, Junkyu Lee, Tigran Tchrakian, Javier Carnerero Cano, Yufang Hou, Elizabeth Daly, and Alessandra Pascale. Factreasoner: A probabilistic approach to long-form factuality assessment for large language models, 2025. URL https://arxiv.org/abs/2502.18573.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*,

pp. 1754–1768, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.115. URL https://aclanthology.org/2022.emnlp-main.115/.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Prasad Rashmi, Webber Bonnie, Lee Alan, and Joshi Aravind. Penn discourse treebank version 3.0. LDC2019T05. Web Download. Philadelphia: Linguistic Data Consortium, 2019.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL https://aclanthology.org/2024.findings-emnlp.552/.

Julius Steen and Katja Markert. How to find strong summary coherence measures? a toolbox and a comparative study for summary coherence measure evaluation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6035–6049, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.527/.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14199–14230, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.830. URL https://aclanthology.org/2024.findings-emnlp.830/.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Qiming Xie, Wenjie Zheng, Xiangqing Shen, and Rui Xia. LoReFact: Bridging the logic gap in fact-checking. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6974–6996, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.346. URL https://aclanthology.org/2026.findings-acl.346/.

Wei Zhao, Michael Strube, and Steffen Eger. DiscoScore: Evaluating text generation with BERT and discourse coherence. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3865–3883, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.278. URL https://aclanthology.org/2023.eacl-main.278/.

A Prompts A and B Used for Relation Extraction

In this section, we provide the prompts A and B used for relation extraction.

A.1 Prompt A — discourse sense and coupling

6,148 characters. Placeholders: `{{sense_menu}}`, `{{response}}`, `{{atom_a}}`, `{{atom_b}}`, where `sense_menu` is:

```
Cause-Effect, Effect-Cause, Evidence, Condition, Restatement, Instantiation, Contrast, Concession, Alternative, Disjunction, Precedence, Succession, None
```

```


Instructions:
You are given the full RESPONSE a model produced, and two atomic claims, A and B, both taken FROM THAT RESPONSE. Decide the discourse/logical relation FROM A TO B.

Most pairs of claims in a text are NOT related. They are simply near each other. Your default answer is None, and you should depart from it only when the response itself draws a dependence between A and B.

Three things that are NOT relations, however plausible they look:
  - Topical adjacency. A and B being about the same subject, or sitting in neighbouring sentences, is not a relation.
  - World knowledge. If the link is true in general but the response does not draw it, the answer is None.
  - Transitivity. If the response links A to some third claim C, and C to B, that does NOT make A related to B. Report only the links the text draws directly.

1. First ask the discriminating question: does the response ITSELF assert a dependence from A to B -- as written, or as an explicit step in the author's argument? Identify the words in the response that do it. If you cannot point to such words, the answer is None.

2. If and only if you can, name the DISCOURSE SENSE, one of: {{sense_menu}}.
   - None: the response draws no dependence between A and B. This is the common case: unrelated, merely adjacent, related only by world knowledge, or related only through a third claim.
   - Cause-Effect: the response says A causes/leads to B. Effect-Cause: A is the effect, B its cause.
   - Evidence: the response offers A as evidence/support for B.
   - Restatement: A and B say the same thing -- a paraphrase, a converse, or a restatement the text marks ("in other words", "equivalently", "that is").
   - Contrast: A and B are opposed but NOT exhaustive (both could be false).
   - Concession: A and B are in tension and the text concedes or resolves it ("although A, still B"; a holding settles it).
   - Alternative: A and B are EXHAUSTIVE competing options -- exactly one holds (not both, not neither).
   - Disjunction: AT LEAST ONE of A and B holds; both may.
   - Precedence: A and B are ordered in time with no truth dependence.

3. Map the sense to a COUPLING, one of: entailment, contradiction, equivalence, exclusive, co_necessity, none.
   - Cause-Effect, Effect-Cause, Evidence -> entailment
   - Restatement -> equivalence
   - Contrast, Concession -> contradiction
   - Alternative -> exclusive       (exactly one of A, B is true)
   - Disjunction -> co_necessity    (at least one of A, B is true)
```

```
   - Precedence, None -> none
   Prefer "exclusive" when the two claims are incompatible AND exhaustive; "contradiction"
   ↪  when they merely cannot both hold but could both be false.

4. Answer on ONE line, with three bracketed tags. The evidence must be a SHORT VERBATIM
↪  QUOTE COPIED FROM THE RESPONSE -- the words that draw the link. Do not quote claim A or
↪  claim B; quote the response. Use [evidence=none] when the sense is None.
[sense=Evidence] [coupling=entailment] [evidence="which is why the panel held"]

Use the following examples to better understand your task.

Example 1 (adjacent and on-topic, but the response draws no link -> None):
RESPONSE: The survey was administered in March to a sample of 400 households. Response rates
↪  in rural districts have declined over the past decade. The analysis weighted results by
↪  district population.
A: The survey was administered in March to a sample of 400 households.
B: Response rates in rural districts have declined over the past decade.
1. The two sentences are adjacent and both concern the survey, but the response never says
↪  the March administration bears on rural response rates. There are no words that draw a
↪  dependence.
2. Discourse sense: None.
3. Coupling: none.
4. [sense=None] [coupling=none] [evidence=none]

Example 2 (the response makes the causal link explicitly):
RESPONSE: The company launched a flawed product last quarter. Reviewers panned it, returns
↪  spiked, and as a direct result the company's stock price fell 15 percent over the same
↪  period.
A: The company launched a flawed product last quarter.
B: The company's stock price fell 15 percent last quarter.
1. The response presents the launch as the head of a chain ending in the stock decline, and
↪  marks it: "as a direct result".
2. Discourse sense: Cause-Effect.
3. Coupling: entailment.
4. [sense=Cause-Effect] [coupling=entailment] [evidence="as a direct result"]

Example 3 (true in general, but this response does not draw it -> None):
RESPONSE: Higher interest rates raise the cost of corporate borrowing. The firm opened three
↪  distribution centres last year. Its logistics costs fell by eight percent.
A: Higher interest rates raise the cost of corporate borrowing.
B: The firm opened three distribution centres last year.
1. One could argue borrowing costs affect expansion decisions, but that is world knowledge.
↪  This response states the two facts side by side and links neither to the other.
2. Discourse sense: None.
3. Coupling: none.
4. [sense=None] [coupling=none] [evidence=none]

Example 4 (the response states an exhaustive alternative -> exclusive):
RESPONSE: The two readings are mutually exclusive and exactly one must be right: either no
↪  one was harmed in the incident, or three people died in it.
A: No one was harmed in the incident.
B: Three people died in the incident.
1. The response sets A and B against each other and says exactly one must be right:
↪  "mutually exclusive and exactly one must be right: either ... or".
2. Discourse sense: Alternative.
3. Coupling: exclusive.
4. [sense=Alternative] [coupling=exclusive] [evidence="mutually exclusive and exactly one
↪  must be right"]

Example 5 (linked only through a third claim -> None):
RESPONSE: The alloy is chemically identical to the certified reference. It therefore meets
↪  the reference specification. Meeting that specification is a precondition for airframe
↪  use.
A: The alloy is chemically identical to the certified reference.
B: Meeting that specification is a precondition for airframe use.
1. The response links A to the middle claim (it meets the specification), and the middle
↪  claim to B. It does not link A to B directly; that would be transitivity, which is not a
↪  relation to report.
2. Discourse sense: None.
3. Coupling: none.
4. [sense=None] [coupling=none] [evidence=none]

Your task:
RESPONSE: {{response}}
A: {{atom_a}}
B: {{atom_b}}
```

A.2 PROMPT B — SURROGATE-TOKEN CONDITIONAL STRENGTH

2,182 characters. Placeholders: `{{response}}`, `{{atom_a}}`, `{{atom_b}}`, `{{coupling}}` (twice). This is the one prompt that does not end in a newline: it terminates with `Answer:` including the trailing space, so that the model's next token is the surrogate whose logprobs are read. Reproducing it without that trailing space changes the tokenization of the answer position and therefore the estimate.

```

Instructions:
You are given the full RESPONSE, two claims A and B drawn from it, and a {{coupling}} relation that holds from A to B. Assuming A is TRUE, judge -- IN THE CONTEXT OF THIS RESPONSE -- whether the relation's implication about B is credible: at least plausible / more likely than not, NOT whether it is certain.

- entailment or equivalence: given A and how the response uses it, is B at least plausibly TRUE (more likely than not)?
- contradiction or exclusive: given A and how the response uses it, is B at least plausibly FALSE (more likely than not)? (For "exclusive", A and B are exhaustive alternatives, so A being true makes B false.)
- co_necessity: A and B are a pair of which at least one holds. Given the response rules out "neither", is it at least plausible that B holds when A does NOT?

Answer with a SINGLE WORD, the very first word of your reply: Yes or No.
- Answer "Yes" if the implication is credible/plausible (even if not certain).
- Answer "No" only if the implication is implausible or the response does not actually support it.
Do not output anything before the word Yes or No.

Example 1 (response supports a near-certain entailment):
RESPONSE: The new alloy is chemically identical to the certified reference alloy, so it meets the certified reference specification.
A: The new alloy is chemically identical to the certified reference alloy.
B: The new alloy meets the certified reference specification.
coupling: entailment
Answer: Yes

Example 2 (weak but plausible, and the response draws the link -- still Yes):
RESPONSE: The company launched a flawed product last quarter, and its stock price fell 15 percent over the same period.
A: The company launched a flawed product last quarter.
B: The company's stock price fell 15 percent last quarter.
coupling: entailment
Answer: Yes

Example 3 (clear contradiction stated by the response):
RESPONSE: The official statement said no one was harmed, but three people died in the incident.
A: No one was harmed in the incident.
B: Three people died in the incident.
coupling: contradiction
Answer: Yes

Your task:
RESPONSE: {{response}}
A: {{atom_a}}
B: {{atom_b}}
coupling: {{coupling}}
Answer:
```

B THE RELATION TAXONOMY

B.1 THE INFERENTIAL COUPLINGS WITH EXAMPLES

Table 3 gives a response excerpt for each coupling, set against the world that coupling denies; since a coupling *is* its penalized set, reading the two together is the quickest way to see what each one commits the response to. The first three couplings are exactly the NLI edge types Bowman et al. (2015). The last two are additions required by argumentative prose, and the distinction between them and CONTRADICTION is the one the taxonomy exists to draw.

Why exhaustiveness needs its own coupling. CONTRADICTION models incompatibility: two claims cannot both be true, but they may both be false. Exhaustive alternatives are stronger. For

Table 3: One realistic response excerpt per Level-1 coupling of Def. 2. As in Table 4, each excerpt is prose containing both claims — A is the source, underlined, B **the target, bold**, the *italic* words drawing the link — and $V(\tau)$ is repeated from Def. 2 so the excerpt can be read against the world it rules out. The *Why* column names the world the response denies, which is what distinguishes each coupling from the one above it; the three conflict-adjacent couplings CONTRADICTION, EXCLUSIVE and CO_NECESSITY are illustrated on one two-part weld failure so that only exhaustiveness varies. NONE produces no factor, so its excerpt is one where the text relates the claims without constraining their truth.

Coupling	Reading	$V(\tau)$	Response excerpt, and the world it denies
ENTAILMENT	$a_s \Rightarrow a_t$	$\{(1,0)\}$	"The alloy is chemically identical to the certified reference, *and therefore* **it meets the airframe specification**." *Why:* the response rules out the world where the source holds and the target fails — $(1,0)$ — and says nothing about the case where the alloy differs, which is why the a_s=0 row returns the target to its prior.
CONTRADICTION	$a_s \Rightarrow \neg a_t$	$\{(1,1)\}$	"The weld passed the tensile test, *yet* **the same joint was porous along its full length**." *Why:* only the both-true world $(1,1)$ is denied. Both can still fail together — if the wrong joint was tested, neither claim holds — so $(0,0)$ must stay available.
EXCLUSIVE	exactly one holds	$\{(0,0),(1,1)\}$	"The joint was destructively examined, so *exactly one of* the weld met the tensile minimum *and* **the weld fell short of it** *must be the finding — there is no third outcome*." *Why:* both same-value worlds are denied. The added $(0,0)$ is the whole difference from CONTRADICTION: once the joint has actually been examined the pair cannot both fail, so leaving that world open — as CONTRADICTION does — would under-penalize the conflict.
CO_NECESSITY	at least one holds	$\{(0,0)\}$	"The fracture began at the joint, so *at least one of* the weld *and* **the fastener** *must have given way* — conceivably both." *Why:* the dual of EXCLUSIVE. Only the both-false world is denied, so asserting the disjunction is evidence *for* both endpoints (Example 1).
EQUIVALENCE	$a_s \Leftrightarrow a_t$	$\{(0,1),(1,0)\}$	"The shipment was certified to specification — *that is to say*, **every contractual requirement had been met**." *Why:* the two disagreeing worlds are denied, in both directions; unlike ENTAILMENT there is no vacuous row, since a false source now also forces a false target.
NONE	no coupling	$\varnothing$	"The replacement units were installed in March, *some three months before* **the first in-flight failures were reported**." *Why:* no world is denied. The response genuinely relates the claims, but ordering them in time constrains neither one's truth, so the pair contributes **no factor** and both claims stay at their priors.

example, "no one was harmed" and "three people died" can neither both hold nor both fail, requiring the additional constraint encoded by EXCLUSIVE. Treating such pairs as mere contradiction leaves an impossible world available and therefore under-penalizes incoherence. Conversely, CO_NECESSITY is the dual coupling, capturing disjunctive statements that require at least one claim to hold. Together, CONTRADICTION, EXCLUSIVE, and CO_NECESSITY represent three distinct ways of constraining compatibility between claims.

Example 1 (A CO_NECESSITY factor in isolation)**.** *Take a single* CO_NECESSITY *relation with* $p = 0.9$ *over two claims with priors* $\pi = 0.5$*, and no other factors. The pairwise table is* $[1 - p,\ \pi_s,\ \pi_s,\ p] = [0.1, 0.5, 0.5, 0.9]$*, and multiplying in the unaries and normalizing gives world masses*

$$\mathbb{P}(0,0) = 0.05, \quad \mathbb{P}(0,1) = 0.25, \quad \mathbb{P}(1,0) = 0.25, \quad \mathbb{P}(1,1) = 0.45.$$

The both-false world is suppressed from the 0.25 *it would receive under independent fair priors to* 0.05*, while the three worlds satisfying "at least one holds" share the recovered mass in proportion to how many endpoints they assert. Each marginal rises from* 0.5 *to* 0.7*: asserting that at least one of two claims holds is evidence for both, which is the intended reading.*

B.2 THE DISCOURSE SENSES WITH EXAMPLES

Table 4 illustrates each discourse sense with a response excerpt containing both the relevant claims and the linguistic cue that licenses the relation. This distinction matters because coherence is expressed through discourse relations, not directly through logical couplings. Some design choices warrant discussion. Restatement is assigned a fixed strength prior of 0.90 because explicit paraphrase markers such as "in other words" or "equivalently" already signal near-equivalence, leaving little uncertainty for the estimator to resolve. Five discourse senses compile to the same ENTAILMENT coupling, reflecting an intentional abstraction: while the model distinguishes causation, evidence, justification, and related discourse functions at the linguistic level, a pairwise potential over truth values can represent only their shared inferential effect. The discourse sense is therefore retained for interpretability and error analysis, while the coherence model operates on the compiled coupling.

The distinction between Contrast, Alternative, and Disjunction illustrates why multiple couplings are needed. These relations may connect the same pair of claims yet impose different logical constraints. Contrast expresses opposition (CONTRADICTION), Alternative expresses an exhaustive choice (EXCLUSIVE), and Disjunction requires that at least one claim hold (CO_NECESSITY). They differ only in which joint truth assignments remain admissible. Finally, Precedence and Succession capture ordering relations that organize the narrative without constraining claim truth values. They represent genuine discourse structure, but compile to no coupling because they carry no inferential content.

B.3 THE COMPILE MAP

Couplings are the right vocabulary for a factor table and the wrong one for annotation. Asked whether two claims stand in ENTAILMENT, an annotator — human or model — has been handed a question about logic, when the fact actually observable in the text is a question about discourse: does this sentence give a cause, offer evidence, concede a tension, or lay out alternatives? We therefore mine an interpretable layer and compile it down. Table 5 shows the compile map we use.

C THE PAIRWISE COHERENCE POTENTIALS, AND WHY FACTUALITY'S TABLES ARE WRONG

An inferential coupling specifies *which* joint assignments are penalized; the factor table determines *how strongly* they are penalized. Choosing these penalties completes Definition 1 and marks the key point at which coherence modeling diverges from the factuality framework.

The distinction arises from the semantics of the source node. In factuality assessment, the source is a retrieved evidence passage whose reliability is itself uncertain, so a factor may legitimately revise belief in the source. In coherence assessment, the source is another claim within the same response. The goal is not to assess that claim's reliability, but to model the inferential constraints it imposes on other claims. Consequently, factor tables designed for claim–evidence reasoning are inappropriate

Table 4: One realistic response excerpt per Level-2 sense of Table 5, in the same row order and grouped by the coupling each compiles to. Each excerpt is prose of the kind the estimator actually sees and *contains both claims*: A is underlined, B **is bold**, and the *italic* words are the ones that draw the link. The Contrast, Alternative and Disjunction rows describe one two-part weld failure three ways, separated only by exhaustiveness; they reuse the scenario of Table 3 deliberately, so that a discourse sense and the coupling it compiles to can be compared on the same prose.

Level-2 sense	$\mathcal{C}$	Response excerpt, and why this sense rather than a neighbour
Cause-Effect*	ENTAILMENT	"The company shipped a batch it knew to be out of tolerance, and *as a direct result* **its share price fell** 15% **that quarter**." *Why:* the response itself presents A as bringing B about; the marker names the direction.
Effect-Cause*	ENTAILMENT	"The entire fleet was grounded on 14 March. The order came *because* **inspectors had found a cracked wing spar on two airframes**." *Why:* the same support runs from the stated effect to its cause. Direction of *explanation* is reversed; direction of inferential support is not.
Evidence*	ENTAILMENT	"The maintenance log records no inspection between January and June, *which shows that* **the mandatory 500-hour check was skipped**." *Why:* A is offered in support of B without any claim that it caused it — a record is evidence, not a mechanism.
Condition	ENTAILMENT	"**A part is cleared for airframe use** *only if* its alloy meets the reference specification." *Why:* B is asserted contingent on A. Note how weakly the cue constrains: nearly any dependence can be reread as a condition, which is why the sense is inadmissible.
Instantiation	ENTAILMENT	"Several suppliers failed the 2023 audit — *for instance*, **the Tucson plant failed on two counts**." *Why:* B is one case falling under A's generalization. Like Condition, broad enough to fit almost any pair that shares a topic.
Restatement*	EQUIVALENCE	"The shipment was certified to specification — *in other words*, **every requirement in the contract had been met**." *Why:* a *marked* paraphrase asserts identity of content, which is why the strength prior is fixed at 0.90 rather than estimated.
Contrast*	CONTRADICTION	"The weld passed the tensile test, *but* **the same joint was porous along its full length**." *Why:* opposed and *not* exhaustive. A third possibility — the wrong joint was tested — leaves both false, so the both-false world must stay available.
Concession*	CONTRADICTION	"*Although* the airline argued throughout that pilot error was to blame, **the final report identified a metallurgical defect**; the tribunal *ultimately held* for the latter." *Why:* a tension the text itself settles. The resolving holding triggers the discount of Eq. **??** rather than counting as a live conflict.
Alternative*	EXCLUSIVE	"The two accounts cannot be reconciled: *either* no one aboard was harmed, *or* **three passengers died**, and *exactly one of them is true*." *Why:* exhaustive. Unlike Contrast the pair can neither both hold nor both fail, so the both-false world is penalized too.
Disjunction*	CO_NECESSITY	"Metallurgy confirms the fracture began at the joint, so *at least one of* the weld *and* **the fastener** *must have failed* — possibly both." *Why:* at least one holds and both may. The dual of Alternative: only the both-false world is penalized.
Precedence*	NONE	"The replacement units were installed in March, *some three months before* **the first in-flight failures were reported in June**." *Why:* the text genuinely relates the two claims, and says nothing about either one's truth — so the sense induces **no factor**.
Succession	NONE	"**The first in-flight failures were reported in June**, *fully three months after* the replacement units had been installed." *Why:* the mirror of Precedence — the same two claims, the same absence of truth-functional content. Offering both senses invites an arbitrary choice on a relation that produces no factor either way.

Table 5: The compile map $\mathcal{C}$. *Ordering-only* senses record assertion order but induce **no factor**, which is what makes an order-only edit provably score-invariant. Concession is a conflict the text may itself resolve.

Level-2 sense	$\mathcal{C}$ coupling	prior	dir.	flag	glossary
Cause-Effect*	ENTAILMENT	$\perp$	yes		source brings target about
Effect-Cause*	ENTAILMENT	$\perp$	yes		source is the effect, target the cause
Evidence*	ENTAILMENT	$\perp$	yes		source is evidence for target
Condition	ENTAILMENT	$\perp$	yes		source conditions target
Instantiation	ENTAILMENT	$\perp$	yes		target instantiates source
Restatement*	EQUIVALENCE	0.90	no		definitional identity
Contrast*	CONTRADICTION	$\perp$	no		opposed, *not* exhaustive
Concession*	CONTRADICTION	$\perp$	yes	conc.	a tension the text resolves
Alternative*	EXCLUSIVE	$\perp$	no		exhaustive: exactly one holds
Disjunction*	CO_NECESSITY	$\perp$	no		at least one holds
Precedence*	NONE	—	yes	ord.	temporal order only
Succession	NONE	—	yes	ord.	temporal order only

for coherence: they treat claims as evidence to be evaluated rather than premises from which the response reasons.

The factor tables. Each potential $\psi_{\tau,p}$ is a 2×2 table over the joint assignments of a source claim a_s and a target claim a_t. For compactness, we write the table as a list of four entries ordered as $(a_s, a_t) \in \{(0,0),(0,1),(1,0),(1,1)\}$ with the assignments in the penalized set $V(\tau)$ shown in bold. Equivalently, the table may be read row-wise: the first two entries correspond to the case $a_s = 0$ (the source claim does not hold), and the last two to the case $a_s = 1$ (the source claim holds).

The parameter π_s is a fixed constant of the factor family and should not be confused with the prior probability of the source claim itself. In the coherence model, π_s is determined by the relation type and remains the same for all claim–claim edges; throughout this work we use $\pi_s = \frac{1}{2}$. The actual prior belief associated with a claim remains the unary factor ϕ_i of Definition 1 and is independent of π_s. Thus, even if a source claim has prior probability $P(a_s = 1) = 0.9$, the row corresponding to $a_s = 0$ is still $[1 - \pi_s, \pi_s] = [\frac{1}{2}, \frac{1}{2}]$. This choice has an important consequence. When the source claim is false, the corresponding row is uniform and therefore contributes the same multiplicative constant to every state of the target claim. After normalization, this constant disappears, leaving the target governed solely by its own prior and the remaining relations in the graph. In other words, a false source neither supports nor opposes the target. This is precisely the sense in which Proposition 1 states that a false source "returns the target to its prior": the prior being recovered is the target claim's own unary prior, not the source claim's prior.

ENTAILMENT	CONTRADICTION	CO_NECESSITY	EQUIVALENCE	EXCLUSIVE
$[1-\pi_s, \pi_s, \mathbf{1-p}, p]$	$[1-\pi_s, \pi_s, p, \mathbf{1-p}]$	$[\mathbf{1-p}, \pi_s, \pi_s, p]$	$[p, \mathbf{1-p}, \mathbf{1-p}, p]$	$[\mathbf{1-p}, p, p, \mathbf{1-p}]$

Proposition 1 (Row-stochasticity of the asymmetric couplings)**.** *For* $\tau \in \{\text{ENTAILMENT}, \text{CONTRADICTION}\}$ *every row of* $\psi_{\tau,p}$ *sums to one, so the row indexed by* a_s is *the conditional* $\mathbb{P}(a_t \mid a_s)$*:* $\mathbb{P}(a_t{=}1 \mid a_s{=}0) = \pi_s$*, and* $\mathbb{P}(a_t{=}1 \mid a_s{=}1) = p$ *for* ENTAILMENT *and* $1 - p$ *for* CONTRADICTION*. Such a factor therefore constrains its target only when the source is true; when the source is false it contributes the constant row* $[1 - \pi_s, \pi_s]$*, which at the* $\pi_s = \frac{1}{2}$ *coherence uses is flat and so cancels under normalization, handing the target back to its* own *prior. This fails for the symmetric couplings, and is not claimed of them: the rows of* CO_NECESSITY $= [1{-}p, \pi_s, \pi_s, p]$ *sum to* $(1 - p) + \pi_s$ *and* $\pi_s + p$*, which are both one only in the degenerate case* $p = \pi_s = \frac{1}{2}$*.*

Proof. The $a_s{=}0$ row is $[1 - \pi_s, \pi_s]$ in both tables, and the $a_s{=}1$ row is $[1 - p, p]$ for ENTAILMENT and $[p, 1 - p]$ for CONTRADICTION; each sums to one, so dividing by the row sum is the identity. For CO_NECESSITY, row-stochasticity needs $\pi_s = p$ and $\pi_s = 1 - p$ at once, hence $p = 1 - p$. □

One important clarification is that Proposition 1 is a statement about the *factor table*, not about the posterior marginals. Row normalization implies that $\psi(a_s, \cdot)$ can be interpreted directly as a

conditional distribution $P(a_t \mid a_s)$. In particular, the uniform row for $a_s = 0$ formalizes the familiar intuition that an implication conveys no information when its antecedent is false. This does *not* imply that the source claim itself remains unchanged. Once all factors in the MRF are multiplied together and the resulting distribution is normalized, the beliefs of both source and target may shift.

For example, with $p = 0.9$ and $\pi_s = 0.5$, the asymmetric couplings are

$$\text{ENTAILMENT} = [0.5, 0.5, 0.1, 0.9], \qquad \text{CONTRADICTION} = [0.5, 0.5, 0.9, 0.1].$$

Both tables are row-normalized, yielding $P(a_t = 1 \mid a_s = 0) = 0.5 = \pi_s$ for both couplings, while $P(a_t = 1 \mid a_s = 1) = 0.9$ for ENTAILMENT and $P(a_t = 1 \mid a_s = 1) = 0.1$ for CONTRADICTION. Thus, the source influences the target only when it holds.

The symmetric couplings behave differently, which is why Proposition 1 applies only to $\{\text{ENTAILMENT}, \text{CONTRADICTION}\}$. Consider CO_NECESSITY, whose table is $[0.1, 0.5, 0.5, 0.9]$. Its rows sum to 0.6 and 1.4, so they cannot be interpreted as conditional distributions. More importantly, the informative case is now the opposite one: when the source is false, the probability of the target holding is $0.5/0.6 \approx 0.833$, whereas when the source is true it is $0.9/1.4 \approx 0.643$. This reflects the semantics of disjunction. A disjunctive statement constrains its target most strongly when the source fails, a dependency that cannot be expressed through a source-conditioned normalization. As shown in Example 1, the effect is symmetric: both endpoints receive support from the relation and their marginals rise together (e.g., from 0.5 to 0.7).

Where the factuality tables differ. Factuality models typically employ a *no-prior* factor family, whose ENTAILMENT table takes the form $[p, p, \mathbf{1} - \mathbf{p}, p]$. The difference from our coherence tables is concentrated in a single row. Both families treat a *true* source identically, penalizing the world in which the source holds but the target does not. They diverge only when the source is false. In the coherence model, the corresponding row is $[\frac{1}{2}, \frac{1}{2}]$, a uniform distribution that contributes no information about the target. In the factuality model, the row is $[p, p]$: still uniform, but scaled by an additional factor of p.

This scaling has an important semantic consequence. Because it increases the weight of worlds in which the source claim holds, the factor acts as evidence *for* the source itself. That behavior is appropriate in factuality assessment, where the source is an external evidence passage whose reliability is uncertain. In coherence assessment, however, the source is another claim within the same response. The role of the factor is to model how the response reasons *from* that claim, not to provide additional support *for* it. Consequently, a factor table designed for claim–evidence reasoning introduces the wrong inductive bias when applied to coherence.

What that costs: a single entailment edge. The effect of the two factor families can be seen in the smallest possible example. Consider a graph containing only a single entailment edge $a_1 \to a_2$ with strength $p = 0.9$, where both claims have prior probability $\pi_1 = \pi_2 = 0.9$. Two distinct quantities appear in this construction and should not be confused. The claim priors π_1 and π_2 enter through the unary factors $\phi_i = [0.1, 0.9]$. The parameter $\pi_s = \frac{1}{2}$ belongs to the factor family itself and determines the row corresponding to a false source. Consequently, the source claim's prior of 0.9 does *not* appear in the factor table: the row for $a_1 = 0$ remains $[1 - \pi_s, \pi_s] = [0.5, 0.5]$ regardless of how likely a_1 is believed to be.

The resulting MRF contains four worlds, each weighted by two unary factors and one pairwise factor, $w(a_1, a_2) = \phi_1(a_1)\,\phi_2(a_2)\,\psi(a_1, a_2)$:

(a_1, a_2)	$\phi_1\phi_2$ (from π_i=0.9)	ψ (with / no priors) (from $\pi_s = \frac{1}{2}, p$)	w with	w no
$(0, 0)$	$0.1 \times 0.1 = 0.01$	$1 - \pi_s = 0.5$ / $p = 0.9$	0.005	0.009
$(0, 1)$	$0.1 \times 0.9 = 0.09$	$\pi_s = 0.5$ / $p = 0.9$	0.045	0.081
$(1, 0)$	$0.9 \times 0.1 = 0.09$	$\mathbf{1} - \mathbf{p} = 0.1$ (both)	0.009	0.009
$(1, 1)$	$0.9 \times 0.9 = 0.81$	$p = 0.9$ (both)	0.729	0.729
$Z = \sum w$			0.788	0.828

For the worlds in which the source claim is true, the coherence and factuality tables are identical. The only difference is in the two worlds where the source is false. Under the coherence table these

worlds receive weights 0.005 and 0.045, whereas under the factuality table they receive 0.009 and 0.081. Nothing else changes.

This seemingly minor difference has a substantial effect on the posterior. The numerator of $P(a_1 = 1) = \frac{w(1,0)+w(1,1)}{Z}$ is exactly the same under both factor families: $w(1,0) + w(1,1) = 0.009 + 0.729 = 0.738$. The only quantity that changes is the partition function Z. By assigning additional mass to worlds in which the source claim is false, the factuality table increases the denominator from 0.788 to 0.828, yielding $P(a_1 = 1) = 0.9365$ (coherence) versus $P(a_1 = 1) = 0.8913$ (factuality)

The interpretation is revealing. Under the coherence table, the source claim becomes slightly more credible than its prior because its conclusion is likely true and the entailment relation is satisfied. Under the factuality table, the same claim becomes *less* credible than its prior, despite facing no contradictory evidence and participating only in a satisfied entailment. The reduction arises solely because the factor assigns additional probability mass to worlds in which the source is false. In effect, the source claim loses credibility simply for being used as a premise.

This behavior is appropriate in factuality assessment, where the source is an evidence passage whose reliability is itself uncertain. It is undesirable in coherence assessment, where the source is another claim in the same argument. A coherence factor should model how the response reasons *from* a claim, not revise belief *in* that claim merely because it serves as a premise.

What that costs on the running example. A single edge is only a toy example; the same effect persists in larger networks. Consider again the five-claim Voyager 1 example of Figure 1, analyzed formally in §E.2. Everything is held fixed: the same claims, the same priors $(0.9, 0.9, 0.9, 0.1, 0.1)$, and the same relation graphs (four edges for Response A and six for Response B). The only change is the factor family. Because the network contains only five binary variables, all quantities can be computed exactly by enumerating the $2^5 = 32$ worlds of Eq. 1, isolating the effect of the factor tables from every other component of the model.

The difference is substantial. Under the coherence tables, Response A receives $\mathrm{LCS}_{\mathrm{mm}} = 0.590$, whereas the factuality tables reduce it to 0.492. Response B drops from 0.494 to 0.409. More revealing than the aggregate scores is the behavior of the individual claims. In Response A, the central conclusion a_3 ("Voyager 1 is in interstellar space") attains posterior probability 0.992 under the coherence model but only 0.889 under the factuality model, despite receiving strong support and facing no contradiction. The same pattern appears throughout the graph: claims that function as premises systematically lose posterior mass because the no-priors tables reward worlds in which those premises are false.

This behavior is not a side effect of inference but a direct consequence of the factor semantics. The factuality tables treat a source node as uncertain evidence whose reliability should be revised by the network. In the Voyager example, however, the source nodes are claims within the response itself. As a result, the model effectively penalizes a claim for participating in the argument. The coherence tables avoid this pathology: a claim may gain or lose support through the relations it enters, but it is never discounted merely for being used as a premise. The effect is small on a single edge and accumulates over larger graphs, producing the substantial reduction in dynamic range observed empirically.

	tables	a_1	a_2	a_3	a_4	a_5	$\mathrm{LCS}_{\mathrm{mm}}$	below own prior
A *(coherent)*	with priors	0.939	0.989	0.992	0.013	0.020	**0.590**	a_4, a_5
	no priors	0.890	0.976	0.975	0.014	0.021	0.575	$\mathbf{a_1}, a_4, a_5$
B *(incoherent)*	with priors	0.960	0.973	0.513	0.006	0.018	**0.494**	a_3, a_4, a_5
	no priors	0.855	0.890	0.545	0.004	0.025	0.464	all five

The *ranking* survives: the coherent response A still beats the incoherent B, 0.575 against 0.464. What breaks is the per-claim reading, which is where a coherence score earns its keep. Response A asserts three true claims in an unbroken support chain and contradicts none of them, yet its chain root a_1 falls from 0.939 to 0.890 — below its own 0.9 prior — by the same mechanism the four-world table above makes explicit, since a_1 is the source of A's $a_1 \to a_2$ edge. Under our tables the only claims A leaves below their priors are a_4 and a_5, the two it explicitly refuses, which is the right answer.

That matters because we can use the count of below-prior claims to *localize* an incoherence. The no-priors tables report a support chain's own root as damage in a response with no conflict in it at all. In B they flag all three true claims instead of the single broken one, so B's real defect — the true a_3 collapsing because its only support is a premise the text itself denies — no longer stands out from two artefacts of the table.

D THE LOGICAL COHERENCE SCORES

Definition 1 specifies a probability distribution over claim truth assignments. To evaluate the coherence of a response, we must derive a scalar score from this distribution. Since several such scores are possible, we first establish the properties a coherence score should satisfy, then introduce four candidate constructions and compare them empirically.

Definition 3 (Desiderata for a coherence score)**.** *A coherence score should:* ***(R1)*** *depend on the joint distribution* $\mathbb{P}(\mathbf{a})$ *rather than the relation graph alone;* ***(R2)*** *be bounded in* $[0, 1]$ *and readily interpretable;* ***(R3)*** *be* monotonic*, so that resolving or removing a conflict cannot decrease the score;* ***(R4)*** *capture logical conflict without introducing ad hoc parameters; and* ***(R5)*** *be computable using standard probabilistic inference.*

Requirement R1 is particularly important. A score that depends only on the relation set $\mathcal{R}$ measures properties of the graph itself, such as the number or type of relations, rather than the coherence induced by those relations. Coherence should instead be evaluated through their effect on the joint distribution over claims. We define next four logical coherence scores.

D.1 (A) MEAN POSTERIOR MARGINAL

$$\mathrm{LCS}_{\mathrm{mm}}(y) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(a_i = 1)\,. \tag{3}$$

$\mathrm{LCS}_{\mathrm{mm}}(y)$ is the expected fraction of claims upheld by the response's argument as a whole and can be computed directly from a single marginal-inference (MAR) query. Conflicts reduce posterior support for the claims they involve, entailment propagates support through the graph, and isolated claims remain at their priors. In addition to the scalar score, the same inference query yields three useful diagnostics: the posterior marginals themselves, which localize incoherence to specific claims; the number of claims whose posterior falls below their prior, a strong indicator of active conflict; and the mean normalized binary entropy, $\mathcal{E} = \frac{1}{n}\sum_i H_2(q_i)$, which quantifies the decisiveness of the model's beliefs. The normalized binary entropy of a posterior marginal $q_i = \mathbb{P}(a_i{=}1)$ is:

$$H_2(q_i) = -\frac{q_i \log q_i + (1-q_i)\log(1-q_i)}{\log 2} \tag{4}$$

which maps entropy to the interval $[0, 1]$. Intuitively, $H_2(q_i) = 0$ when the model is certain ($q_i{=}0$ or $q_i{=}1$), and $H_2(q_i) = 1$ when the model is maximally uncertain ($q_i{=}0.5$). Intermediate values quantify how ambiguous the truth status of the claim remains after inference. Therefore $\mathcal{E}$ measures the average uncertainty of the coherence MRF about the claims. Lower values indicate that the argument structure leads to decisive posterior beliefs, whereas higher values indicate unresolved tensions, conflicting evidence, or weak inferential support.

D.2 (B) CONSISTENCY: A TWO-TERM PROBABILITY

A natural alternative to averaging posterior marginals is to ask whether the response is globally consistent. We operationalize this idea through a two-part score:

$$\mathrm{LCS}_{\mathrm{cn}} = \frac{1}{2}\big(\mathcal{K} + \mathcal{S}\big), \quad \mathcal{K} = P\left(\bigwedge_{r:\tau_r=\text{CONTRADICTION}} \neg(a_{s_r} \wedge a_{t_r})\right), \quad \mathcal{S} = \frac{\sum_{r\in\mathcal{R}_\mathcal{S}} p_r U_r}{\sum_{r\in\mathcal{R}_\mathcal{S}} p_r}, \tag{5}$$

where $\mathcal{K}$ is the probability that all contradiction constraints are respected, and $\mathcal{S}$ measures how strongly the response upholds its supportive relations. The set $\mathcal{R}_\mathcal{S}$ contains entailment, equivalence,

and exclusivity relations, and

$U_r = \mathbb{P}(a_{s_r}{=}1 \wedge a_{t_r}{=}1)$ for ENTAILMENT, EQUIVALENCE; $U_r = \mathbb{P}(a_{s_r} \neq a_{t_r})$ for EXCLUSIVE.

The distinction between the two terms $\mathcal{K}$ and $\mathcal{S}$ is intentional. The consistency term $\mathcal{K}$ focuses on explicit contradictions, whereas the support term $\mathcal{S}$ rewards relations that are actively realized by the argument. In particular, a relation receives credit only in its *informative* states. For example, an entailment contributes when both premise and conclusion are supported, not when the premise is simply false and the entailment is vacuously satisfied.

Clearly a score based solely on the probability that all relations are satisfied would assign the maximum value to a response with no relations at all, since an empty set of constraints is trivially satisfied. This is formalized by Proposition 2. Such a score therefore measures properties of the relation graph rather than the coherence of the response, violating Requirement R1. By rewarding only actively upheld relations, $\mathcal{S}$ avoids this vacuous-truth pathology: a disconnected collection of claims receives no coherence credit.

Both $\mathcal{K}$ and $\mathcal{S}$ can be computed exactly using standard marginal inference with a small number of auxiliary variables. We average rather than multiply the two terms so that neither can dominate the other: a single contradiction should not erase an otherwise well-supported argument, and a relation-free response should not inherit a perfect score from an empty conflict set.

Proposition 2 (The vacuous-truth trap)**.** *Let $\mathcal{K}' = \mathbb{P}($every relation in $\mathcal{R}$ is satisfied$)$, the apparently natural extension of $\mathcal{K}$ to all couplings. Then a response with no mined relations satisfies $\mathcal{K}' = 1$, the maximum attainable value.*

Proof. $\mathcal{K}'$ is the probability of an intersection of $|\mathcal{R}|$ events; for $\mathcal{R} = \varnothing$ the intersection is the whole space. □

D.3 (C) A REIFIED COHERENCE NODE

The most direct interpretation of "is this response coherent?" as a single probability is to introduce a variable for it. Add a binary R with Bernoulli prior ρ and, for each relation, a ternary *vote* factor

$$h_r(R, a_s, a_t) = \begin{cases} 1 & R = 0 \quad \text{(the branch is flat)} \\ 1 & R = 1,\ (a_s, a_t) \notin V(\tau_r) \\ 1 - p_r & R = 1,\ (a_s, a_t) \in V(\tau_r), \end{cases} \qquad \mathrm{LCS}_{\mathrm{re}}(y) = \mathbb{P}(R = 1)\,. \tag{6}$$

Under $R = 1$, each violated relation contributes a penalty of $1 - p_r$, so coherence is supported only when the relations are jointly respected. Under $R = 0$, all vote factors are neutral, making incoherence a catch-all explanation. If no relations are present, R becomes independent of the graph and $\mathrm{LCS}_{\mathrm{re}} = \rho$, correctly indicating that coherence cannot be assessed from the response structure alone. This construction has an intuitive interpretation but inherits a limitation of relation-satisfaction scores. Because entailments are treated as satisfied whenever their premise is false, vacuous satisfaction contributes positively to coherence. As a result, $\mathrm{LCS}_{\mathrm{re}}(y)$ is influenced not only by actively supported relations but also by relations that hold trivially.

D.4 (D) NORMALIZED LOG-PARTITION

The partition function measures how much probability mass the constraints leave standing, which is a global coherence signal requiring no marginals at all. It is unnormalized, so we grade it between two references built from the same edge skeleton:

$$\mathrm{LCS}_{\mathrm{lp}} \;=\; \frac{\log Z - \log Z_{\min}}{\log Z_{\max} - \log Z_{\min}}. \tag{7}$$

For the ceiling, $Z_{\max}$ is the partition function of the same network with every conflict edge removed: deleting constraint factors only adds mass, so $\log Z \leq \log Z_{\max}$ always. For the floor we take the *MAP world mass* of the base network, $Z_{\min} = \max_{\mathbf{a}} \prod_i \phi_i \prod_r \psi_r$.

Lemma 1 (The MAP world mass is a valid, coherence-graded floor)**.** $\log Z_{\min} \leq \log Z$ *for every network, and $Z_{\min}$ is non-increasing in conflict strength.*

Table 6: The four scores against the criteria of Def. 3, with the inference queries each requires.

Readout	Queries	R1	R2	R3	R4	R5	Note
(a) mean marginal ($\mathrm{LCS}_{\mathrm{mm}}$)	MAR	✓	✓	✓	✓	✓	**selected**
(b) consistency ($\mathrm{LCS}_{\mathrm{cn}}$)	MAR×2	✓	✓	×	✓	✓	concession dip
(c) reified ($\mathrm{LCS}_{\mathrm{re}}$)	MAR	✓	✓	×	×	✓	free ρ; dip
(d) norm. $\log Z$ ($\mathrm{LCS}_{\mathrm{lp}}$)	PR, MAP	✓	✓	✓	✓	✓	needs 2 references

Algorithm 1 COMPUTELCS(y) — the end-to-end coherence score.

Require: response y; relation estimator $\mathcal{M}$; discount λ; priors π
Ensure: LCS $\in [0,1]$, marginals $\{q_i\}$, diagnostics

$\mathcal{A} \leftarrow$ ATOMIZEANDDECONTEXTUALIZE(y)
$P \leftarrow$ SELECTCANDIDATEPAIRS$(\mathcal{A}, y, \dots)$ ▷ Algorithm 2
$\mathcal{R} \leftarrow \varnothing$
for all $(a_i, a_j) \in P$ **do** ▷ concurrent, rate-limited
  $(\text{sense}, \tau, c) \leftarrow$ PROMPTA(a_i, a_j, y) ▷ $c = \mathbb{P}(\tau \mid \cdot)$
  if $\tau =$ NONE **then continue**
  $\sigma \leftarrow$ PROMPTB(a_i, a_j, τ, y) ▷ surrogate-token strength
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{(i, j, \tau,\ \mathrm{clip}_{[0,1]}(c \cdot \sigma))\}$
for all $r \in \mathcal{R}$ with $\tau_r \in \mathcal{T}_{\mathrm{conf}}$ sensed as Concession **do**
  if a resolving holding h exists **then** $p_r \leftarrow p_r(1 - \lambda\pi_h)$
  else mark r unresolved
$\mathcal{N} \leftarrow$ network of Eq. 1 over $(\mathcal{A}, \mathcal{R}, \pi)$ ▷ with-priors tables
$\{q_i\} \leftarrow$ INFER$(\mathcal{N}, \mathrm{MAR})$
return $\frac{1}{n}\sum_i q_i$, $\{q_i\}$, $\big(\#\{q_i < \pi_i\},\ \mathcal{E},\ \log Z\big)$

Proof. $Z = \sum_{\mathbf{a}} \mathrm{mass}(\mathbf{a})$ is a sum of non-negative terms, so its single largest term is a lower bound. For the second claim, increasing p_r on a conflict edge multiplies the mass of every world in $V(\tau_r)$ by a factor below one and leaves other worlds unchanged, so the maximum over worlds cannot increase. □

D.5 COMPUTING THE SELECTED COHERENCE SCORE

Table 6 summarizes the criteria satisfied by the proposed scores. We can see that both $\mathrm{LCS}_{\mathrm{mm}}$ and $\mathrm{LCS}_{\mathrm{lp}}$ are monotone. Algorithm 1 computes the selected logical coherence score.

E DETAILS ON THE TWO-STAGE COHERENCE PIPELINE

Figure 3 illustrates the two-stage architecture that combines factuality assessment with logical coherence. The coherence model consumes claim-level factuality estimates produced by FactReasoner (Marinescu et al., 2025); these estimates become the priors of the coherence MRF. A response y is first decomposed into atomic claims $\mathcal{A} = \{a_1, \dots, a_n\}$, each representing a single independently verifiable proposition. Claims are then decontextualized to resolve anaphora and remove near-duplicates.

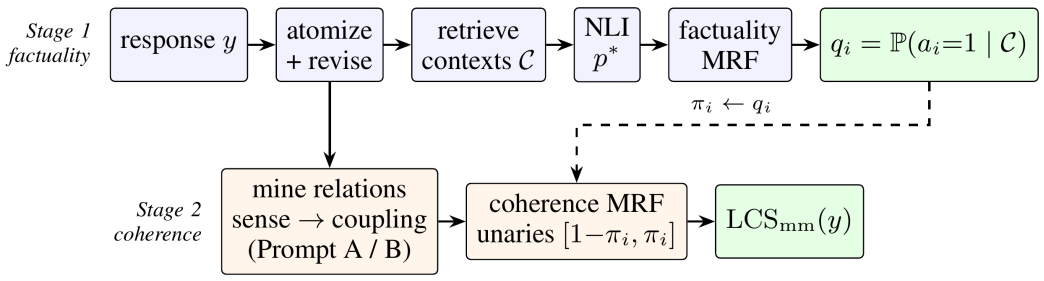


Figure 3: The two-stage pipeline for computing the logical coherence score.

Algorithm 2 SELECTCANDIDATEPAIRS: response-anchored candidate selection.

Require: claims $\mathcal{A} = \langle a_1, \ldots, a_n \rangle$ in assertion order; response y; policy; window w; gate threshold θ_g; discourse thresholds θ_d, δ
Ensure: ordered candidate pairs P, plus a coverage record
1: **if** policy = ALL_PAIRS **then return** every ordered pair (a_i, a_j), $i \neq j$
2: $W \leftarrow \{(a_i, a_j) : 0 < j - i \leq w\}$ ▷ forward order window
3: **if** policy = BIDIRECTIONAL **then** $W \leftarrow W \cup \{(a_j, a_i) : 0 < j - i \leq w\}$
4: $G \leftarrow \varnothing$
5: **if** policy = SEMSIM **then**
6: $G \leftarrow \{(a_i, a_j) : j - i > w,\ \mathrm{sim}(a_i, a_j) \geq \theta_g\}$
7: align each claim to its best-matching sentence of y
8: $Prom \leftarrow \{(a_i, a_j) \notin W \cup G : \textsc{DiscourseAdjacent}(i, j)\}$
9: $Dem \leftarrow \{(a_i, a_j) \in W : \neg\, \textsc{DiscourseAdjacent}(i, j)\}$
10: **return** $\big(\textsc{Dedup}(W \cup G \cup Prom) \setminus Dem\big),\ (|W|, |G|, |Prom|, |Dem|)$

For each claim, a retrieval component returns a set of supporting passages $\mathcal{C} = \{c_1, \ldots, c_m\}$. An LLM-based NLI extractor labels claim–evidence and evidence–evidence pairs as `entailment`, `contradiction`, or `neutral`; mutual entailment between evidence passages is relabeled as *equivalence*. Neutral relations are discarded, yielding a graph containing only informative dependencies.

Each retained relation is associated with a confidence score p^*. Rather than eliciting a separate confidence estimate, we derive p^* from the probability assigned by the LLM to the emitted relation label. If T denotes the tokens spanning the predicted label, we define

$$p^* = \exp\Big(\frac{1}{|T|} \sum_{t \in T} \log p_t\Big), \tag{8}$$

that is, the geometric mean probability of the label tokens. When token-level probabilities are unavailable, we estimate confidence using SIMBA-UQ (Bhattacharjya et al., 2025); degenerate cases default to a neutral confidence of 0.5.

The resulting factuality graph is represented as a pairwise MRF containing one binary variable per claim and per evidence passage. Unary factors encode prior beliefs, while each retained relation contributes a pairwise factor parameterized by its label and confidence. Probabilistic inference yields posterior claim probabilities $P(A_i = 1)$, which are passed directly to the coherence model. We consider three graph variants: FR1, which connects claims only to their retrieved evidence; FR2, which scores every claim against all deduplicated evidence passages; and FR3, which additionally introduces evidence–evidence relations. Throughout this work we use FR2. Unless otherwise stated, all experiments employ the with-priors factor tables with $\pi_C = 0.9$ for evidence passages and $\pi_A = 0.5$ for claims.

E.1 CANDIDATE PAIR SELECTION FOR COHERENCE

Algorithm 2 defines the candidate-pair selection procedure used by the relation miner. We consider four policies: ALL_PAIRS, WINDOWED, SEMSIM, and BIDIRECTIONAL. Each claim is aligned to its originating sentence. A pair of claims is considered *discourse-adjacent* if the corresponding sentences are nearby, share sufficient lexical overlap (measured by Jaccard similarity), or the target sentence begins with a discourse connective. Candidate selection is then refined using these cues: out-of-window discourse-adjacent pairs are promoted into the candidate set, while in-window pairs lacking such evidence are removed. This refinement is not a neutral preprocessing step but part of the extraction policy itself, as it biases candidate generation toward the same discourse signals later used by the relation classifier. In our experiments, we use a Jaccard threshold of 0.2 to determine discourse adjacency and a similarity threshold of 0.3 for the SEMSIM candidate-selection policy.

E.2 EXAMPLE

We revisit the motivating pair from Figure 1, now expressed in the language of discourse relations and coherence MRFs. The example also makes the orthogonality of factuality and coherence concrete:

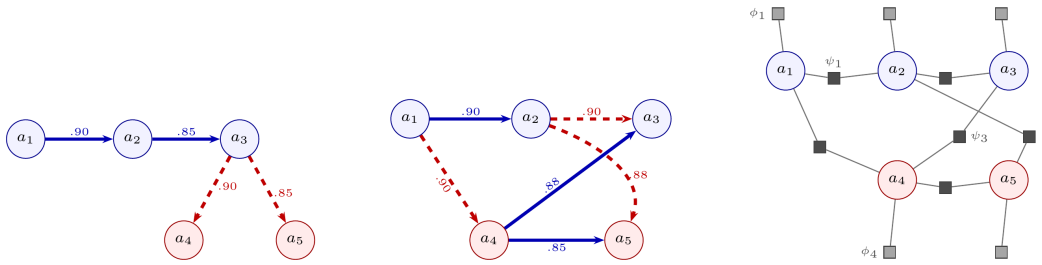


Figure 4: The relation graphs and coherence MRF for motivating example.

both responses contain exactly the same claims and therefore receive identical claim-level factuality judgments, differing only in how those claims are connected.

Consider five claims about Voyager 1: three true claims (a_1, a_2, a_3; prior 0.9) and two false ones (a_4, a_5; prior 0.1). The true claims state that Voyager 1 crossed the heliopause in 2012, that its instruments observed a sharp rise in plasma density, and therefore that it entered interstellar space. The false claims assert that its plasma wave instrument had failed before the crossing and that the probe has left the Solar System entirely.

Both responses assert these same five claims. In **Response A**, the rise in plasma density supports the conclusion that Voyager 1 is in interstellar space, while the two false claims are explicitly rejected. In **Response B**, the false claim a_4 is instead used as support for a_3, and a_3 is simultaneously contradicted elsewhere in the response before both false claims are ultimately rejected. Thus, the difference lies entirely in the relations between claims rather than in the claims themselves.

Figure 4 shows the resulting relation graphs and the coherence MRF induced by Response B. Two relation changes account for the entire difference. First, the support relation $a_2 \rightarrow a_3$ in Response A becomes a contradiction in Response B, causing the response to undermine its own conclusion. Second, the edge between a_3 and a_4 reverses direction: what Response A uses to reject a false claim, Response B uses as evidence for a conclusion. Mentioning a falsehood is not what harms coherence; reasoning from it is.

Because the model contains only five binary variables, the coherence score can be computed exactly by enumerating all 2^5 assignments. Response A receives $\mathrm{LCS}_{\mathrm{mm}} = 0.590$, whereas Response B receives $\mathrm{LCS}_{\mathrm{mm}} = 0.494$. In Response A, the true claims attain posteriors of 0.939, 0.989, and 0.992, all exceeding their prior of 0.9 because the entailment structure reinforces them. In Response B, however, the true claim a_3 drops to 0.513: its only supporting premise is a claim that the response itself later rejects, preventing support from propagating through the graph. Since both responses share identical priors, the entire difference in score is attributable to the arrangement of relations alone.

F THE LOCOBENCH BENCHMARK

Evaluating a coherence measure requires responses that vary in coherence while holding claim content fixed, together with gold relation graphs specifying the intended argument structure. Natural corpora provide neither: coherence is confounded with differences in content, topic, and length, and the absence of gold relations prevents scoring errors from being disentangled from relation-extraction errors. We therefore introduce LOCOBENCH, a synthetic benchmark designed to isolate coherence as the sole varying factor and to provide gold relation graphs for evaluation. Because the benchmark is generated, its construction must be specified explicitly. The remainder of this section describes the coherence-ladder framework (Section F.1), the generation pipeline (Section F.2), the derivation of gold relation graphs (Section F.3), the validation procedure (Section F.4), and the composition of the released corpus (Section F.5).

F.1 COHERENCE LADDERS

Definition 4 (Coherence ladder)**.** *A* coherence ladder *is a tuple* $(\mathcal{P}, \langle \rho_0, \ldots, \rho_4 \rangle, \Gamma)$*, where* $\mathcal{P}$ *is a base relation plan consisting of a claim set and a relation graph,* $\rho_0, \ldots, \rho_4$ *are perturbed realizations*

Table 7: The four family types. Each holds the claim set, the topic and every unperturbed relation fixed, varying one structural dimension. ORDER and CONTROL apply only meaning-preserving edits and so assert *invariance*: CONTROL is the false-positive check, since a measure showing any monotone trend across it is tracking *having been edited* rather than coherence. The final column is the count in the released corpus (§F.5).

Type	Dimension varied	Rungs, $\rho_0 \rightarrow \rho_4$	Required pattern	n
CONFLICT	conflict structure	worse, base, concession_resolved, fix_one_conflict, coherent	strictly increasing	8
CHAIN	entailment-chain integrity	break_3, break_2, break_1, base, coherent	strictly increasing	2
ORDER	assertion order	shuffle_full, shuffle_block, shuffle_adjacent, base, ordering_only	invariant	3
CONTROL	nothing (meaning-preserving)	ordering_only, direction_reversal, base, direction_reversal, ordering_only	flat throughout	1

of $\mathcal{P}$, each with its own response text and gold relation graph, and Γ is a set of expected ordering constraints over their coherence scores. A ladder is admitted only if all five rungs satisfy the validation criteria of §F.4.

The ladder construction isolates coherence from content. All five rungs share the same topic and claim set, so any score difference must arise from changes in relational structure rather than factual content. At the same time, each rung carries its own gold relation graph, allowing perturbations to alter the intended argument structure while keeping the underlying claims fixed. This makes ladders suitable for evaluating coherence scoring independently of relation extraction.

Ladders are admitted or rejected as a whole. Since the coherence constraints in Γ are defined over the relative ordering of all five rungs, a missing rung invalidates the comparison. Consequently, if any rung fails validation, the entire ladder is discarded.

Table 7 summarizes the four ladder families used in LOCOBENCH. Each family holds the claim set, topic, and all unaffected relations fixed while varying a single aspect of argument structure. CONFLICT ladders progressively resolve contradictions, testing sensitivity to explicit inconsistency. CHAIN ladders break and repair entailment links, testing whether a measure can detect missing support rather than overt conflict. In both cases, coherence is expected to increase monotonically across the ladder.

The remaining two families serve as negative controls. ORDER ladders apply only meaning-preserving reorderings, while CONTROL ladders modify the text without changing the underlying relation graph. Both therefore require invariance: a coherence measure that changes across these ladders is responding to surface edits rather than argument structure. As shown in our experiments, these controls are essential for distinguishing genuine coherence measures from metrics that track editing artifacts or sentence order.

Each ladder induces three classes of evaluation constraints. **C1** requires monotonic improvement between consecutive rungs, **C2** specifies expected decreases or invariances for particular readouts, and **C3** requires separation between the first and last rung. Together, these constraints define the expected behavior of a coherence measure across each ladder and form the basis of our evaluation protocol.

F.2 THE GENERATION PIPELINE

The benchmark is generated through a five-stage pipeline followed by independent validation (Figure 5). The stages are intentionally modular, allowing generation failures to be localized and corrected at their source. All prompts are versioned and specified explicitly (Table 9) to ensure reproducibility.

P1: Topic → Question. Given a topic drawn from a pool of 36 topics spanning four domains, generate an open-ended question. Topics are selected to support the target discourse relations, with a preference for domains such as investigations, litigation, and diagnostic reasoning where contrast, exclusivity, and resolution naturally arise.

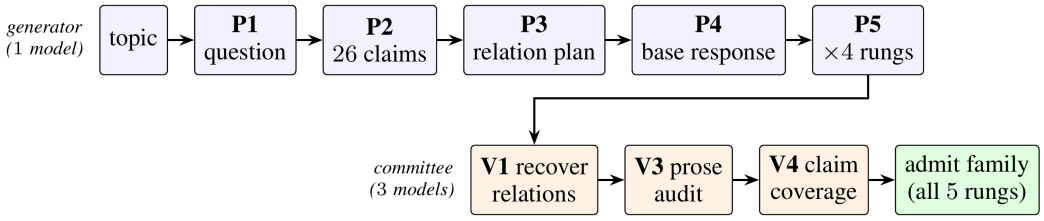


Figure 5: The LOCOBENCH benchmark generation pipeline.

Table 8: The six perturbation operators and the ten calls realizing them. Each call has a declared effect on the rung's *gold relation graph* as well as on its text, which is what makes §F.3's derivation well defined. The two O6 calls are factor-invariant by design: they are what the ORDER and CONTROL ladders exist to check.

Op	Calls	Effect on the text
O1 relabel	wrong_sense, direction_reversal, exhaustiveness_flip	rewrite the connective so the text draws a different sense, reverses the relation, or flips Contrast↔Alternative
O2 add conflict	inject_contradiction, spurious_relation	add one sentence directly contradicting a named claim, at least two sentences later; or assert a link between two claims the response kept separate
O3 break chain	break_chain	remove or negate the link realizing an edge so the chain no longer connects, keeping both endpoints
O4 drop relation	drop_relation	remove the connective realizing an edge, leaving the two claims as adjacent assertions
O5 resolve	add_resolution, remove_resolution	add or delete the holding sentence that settles a conceded tension
O6 reorder	shuffle_order, ordering_only	permute sentences at full, block or adjacent granularity; or exchange Precedence↔Succession preserving chronology — *no coupling changes*

P2: Question → Claims. Generate 26 atomic claims answering the question, each labeled with a truth value. Producing a surplus of claims allows subsequent stages to select those best suited to the desired relation structure.

P3: Claims → Relation Plan. Select 14–16 claims and construct a relation plan containing approximately 10 discourse relations, 4–6 explicitly unrelated claim pairs, relation strengths, and validity labels. Relations are chosen using the discourse taxonomy of Section 2.3 and are preferentially local, making short-range extraction policies realistic baselines. Explicit non-relations provide labeled negative examples for evaluating precision.

P4: Relation Plan → Response. Realize the plan as a coherent response of 550–650 words that asserts all selected claims and expresses the intended relations in natural prose.

P5: Response → Ladder Rungs. Generate ladder variants by applying controlled perturbations to the base response. Each perturbation modifies only the targeted relation structure while preserving the underlying claims. In addition to the modified response, the stage produces a structured record of the relations added, removed, or relabeled. The available perturbation operators are summarized in Table 8.

The generation pipeline employs four failure actions, applied at different stages. A **retry** addresses transient parsing failures by reissuing the same prompt at progressively higher sampling temperatures. A **regenerate** replaces an artifact whose output is invalid despite valid inputs. A **drop edge** applies only during relation validation: if a planned relation cannot be recovered from the generated text, it is removed from the gold graph rather than retained as an unsupported annotation. Finally, a **reject** discards an entire ladder when it can no longer support its intended ordering constraints. Rejected ladders are retained together with their failure reasons, ensuring that rejection rates remain transparent and auditable.

Table 9: Prompt specification. Each prompt's numbered-instruction count and required placeholder set are declared here and asserted by a test against the prompt strings, so a prompt cannot drift from the document describing it without the test failing. Instructions are counted by matching lines beginning at column one with a digit, period and space, so indented sub-items — P3's per-sense glosses, P5's per-call definitions — do not inflate the count. The validators carry no numbered instructions because their prompts specify an output schema rather than a procedure.

Stage	Role	Instructions	Required placeholders
P1	topic → question	9	`TOPIC`
P2	question → claims	7	`QUESTION`
P3	claims → relation plan	10	`CLAIMS`, `QUESTION`
P4	plan → base response	8	`PLAN`, `QUESTION`
P5	response → rung	5	`OPERATOR`, `PLAN`, `RESPONSE`
V1	recover relations	0	`ATOMS`, `RESPONSE`
V3	prose audit	0	`RESPONSE`
V4	claim coverage	0	`ATOMS`, `RESPONSE`

Table 10: Every acceptance threshold. V1 rates are computed over the *valid* planned relations only, since an invalid planned relation is one the prose is not required to draw soundly. The rightmost column is the committee decision rule for that gate.

Gate	Quantity	Threshold	Rule
V1	planned relations recovered with correct coupling	≥ 0.80	majority
V1	… and with correct sense	≥ 0.70	majority
V3	fluency, formality, organization (each, of 5)	≥ 4	majority
V3	leakage and hedging span lists	must be empty	majority
V4	claim coverage over gating statuses	$= 1.00$	majority
P3	claims selected	14–16	—
P3	relations planned (accepted band; 10 instructed)	8–12	—
P3	declared non-relations	4–6	—
P3	fraction of relations valid	0.55 ± 0.15	—
P3	false claims selected	≥ 2	—
P5	length drift	$\leq 15\%$	—

F.3 Deriving each rung's gold relation graph

Each ladder rung is associated with its own gold relation graph, obtained by applying the rung's perturbations directly to the base relation plan. This derivation ensures that the gold graph reflects the intended argument structure of the generated response rather than merely inheriting the structure of the original plan.

Three mechanisms guarantee consistency between perturbations and gold annotations. First, every perturbation is assigned a distinct eligible target relation. For example, two drop_relation operations must affect two different edges. Eligibility depends on the perturbation type: a drop_relation targets an unresolved conflict, whereas an add_resolution targets a conflict that can be resolved. Plans that cannot support the required perturbations are rejected during generation.

Second, whenever a perturbation changes a discourse sense, all derived properties of the relation are recomputed through the compilation map $\mathcal{C}$. In particular, the coupling type, directionality, exhaustiveness, and ordering-only status are re-derived rather than inherited, ensuring that the resulting graph remains consistent with the discourse taxonomy.

Finally, an adjacency check prevents redundant rungs. A ladder is rejected if two consecutive rungs produce identical relation graphs despite a perturbation claiming to modify the argument structure. The only exception is the pair of factor-invariant ordering operators, which are designed to leave the graph unchanged and therefore serve as explicit invariance tests.

Table 11: Composition of the released corpus. Every figure is computed from the shipped items rather than from the generation targets. *Edge-producing* excludes the 67 ordering-only relations, which compile to NONE and contribute no factor; *backward* counts edge-producing relations whose target precedes its source in assertion order, the relations a forward-only candidate policy cannot reach.

Quantity	Value	Quantity	Value
Families	14	Gold relations, total	670
Items (rungs)	70	edge-producing	603
Claims	1,115	valid	408 (0.61)
Claims per item	15–16	backward	105
Words per item	609–698	Declared non-relations	342
Distinct topics	14	Ordering constraints per source	202
CONFLICT / CHAIN	8 / 2	C1 / C2 / C3 checks	112 / 30 / 45 + 15
ORDER / CONTROL	3 / 1	Generator / committee	1 / 3 models

F.4 VALIDATION

Every ladder is admitted only after passing an independent validation process. To avoid self-certification, the model that generated a ladder is excluded from the validation committee, and all decisions are made by majority vote among the remaining models. This design prevents generator-specific artifacts from being accepted as ground truth and reduces the influence of any single overly strict or overly permissive validator. All acceptance thresholds are defined explicitly in Table 10, ensuring that the benchmark can be reproduced and re-evaluated under different operating points. Ladders that fail validation are rejected rather than repaired, preserving a transparent correspondence between the generation process and the resulting gold annotations.

Validation consists of three complementary checks. **V1** verifies that the generated response realizes the intended relation structure by asking an independent model to recover discourse relations directly from the text. Admission requires at least 80% of valid planned relations to be recovered with the correct coupling and at least 70% with the correct discourse sense. This is the key validation step, as it certifies that the gold relation graph reflects the response itself rather than the underlying generation plan. **V3** evaluates linguistic quality, including fluency, formality, and organization, and rejects responses containing annotation leakage or perturbation cues that could reveal a ladder rung without reasoning about coherence. **V4** verifies claim coverage, ensuring that every selected claim is explicitly asserted in the response. Together, these checks guarantee that admitted ladders are coherent linguistic artifacts whose gold relation graphs faithfully describe the discourse structure expressed in the text.

F.5 THE RELEASED CORPUS AND LIMITATIONS

Table 11 summarizes the released LOCOBENCH corpus. The benchmark contains 14 coherence ladders (70 responses) spanning 14 distinct topics across four domains: Natural Sciences, Engineering & Technology, Humanities, and Social Sciences. The corpus was generated using `claude-opus-5` and validated by an independent cross-vendor committee consisting of `gpt-5.6-terra`, `gemini-3.1-pro`, and `claude-sonnet-5`, none of which participated in generation. The benchmark includes all four ladder families, yielding 202 declared ordering constraints: 112 from CONFLICT, 30 from CHAIN, 45 from ORDER, and 15 from CONTROL. This diversity is essential, as it enables separate evaluation of ordering and invariance properties.

Two considerations are important when interpreting the results. First, the corpus is entirely LLM-generated, so systematic biases shared by both the generator and validators may remain undetected. Second, although the benchmark contains 70 responses, evaluation is ultimately based on 202 ordering constraints. This is sufficient to distinguish methods that differ substantially in performance, but not to reliably resolve small differences between closely matched approaches.

F.6 ONE FAMILY END TO END: `F001`

So far, we have described in detail the benchmark construction process. We now explain in detail a single shipped family, namely `f001` — a CONFLICT ladder on Anthropology (Social Sciences) — to

better understand its components: the five rungs, the relation graph each carries, the constraints the family declares, and the scores those constraints are checked against. Every number below is read from the released corpus (`families.json`, `items.jsonl`) and from the GOLD relation graphs.

The family's framing question is: *how do anthropologists adjudicate between rival interpretations when skeletal remains at a burial site show signs of trauma, and what evidence establishes whether the injuries were inflicted by human violence or by post-mortem processes?* All five rungs assert the same 16 claims about that question, in the same order (i.e., the order of the claims is not varied).

The five rungs. Table 12 gives the ladder. Each non-base rung is a sequence of operator calls applied to the base plan ρ_1 — *not* to the rung below it, so a defect in one rung cannot propagate up the ladder. The relation counts show the mechanism directly: the number of CONTRADICTION edges falls $4 \to 3 \to 3 \to 2 \to 2$ as conflicts are repaired, the count of resolved concessions rises $1 \to 2$ at ρ_2 where the concession is discharged, and ρ_4 additionally drops the lone EXCLUSIVE edge. Because the claim set never changes, no per-claim factuality verdict differs across the five rows, and so none of them can explain any score difference.

Table 12: The five rungs of `f001`. *Calls* are the perturbation operators applied to the base plan; *parent* is the rung the perturbation is applied to, which is the base for all four. *Rel.* counts all declared relations and *edge* only the edge-producing ones (the difference is the ordering-only sense, which compiles to NONE). *Cfl.* counts CONTRADICTION edges and *Res.* resolved concessions — the two quantities the ladder is built to vary.

	Rung	Calls applied to the base	Par.	Rel.	Edge	Cfl.	Res.
ρ_0	worse	spurious_relation	ρ_1	11	10	4	1
ρ_1	base	— (the base response)	—	10	9	3	1
ρ_2	concession_resolved	add_resolution	ρ_1	10	9	3	2
ρ_3	fix_one_conflict	add_resolution, drop_relation	ρ_1	9	8	2	2
ρ_4	coherent	add_resolution, drop_relation $\times 2$	ρ_1	8	7	2	2

The coherence scores. Table 13 gives all four coherence scores (or readouts) on all five rungs, from the GOLD relation source. Three of them are graded; re is computed and stored but no constraint is stated over it, which is why the multiplier behind the corpus's 202 assertions is three rather than four. Read down the mm column and the ladder is visible as intended, $0.750 \to 0.823$ end to end; read down cn and it is not monotone, which is the point of the next paragraph rather than a defect.

Table 13: The four readouts on `f001`, GOLD source. mm rises across the ladder; the cn dip at ρ_2 is predicted, not incidental. re is stored but ungraded. These are the values every constraint in Table 14 is checked against.

	Rung	mm	cn	lp	re
ρ_0	worse	0.750400	0.499618	0.342596	0.150678
ρ_1	base	0.760539	0.649468	0.374712	0.221853
ρ_2	concession_resolved	0.789580	0.494457	0.374268	0.252912
ρ_3	fix_one_conflict	0.789239	0.500444	0.425367	0.261813
ρ_4	coherent	0.822792	0.523759	0.488651	0.320921

The constraints. The declaration grid is three graded readouts $\times$ four adjacent rung pairs, plus one endpoint assertion per readout: $12 + 3 = 15$ cells. A CONFLICT family declares only 14, because one cell is deliberately left undeclared — cn on $\rho_2 \to \rho_3$, where the readout is *unconstrained*: the rung both discharges a concession and drops a conflict, and the two move the conflict term in opposite directions by amounts the construction does not fix. Asserting either direction there would be asserting something the design does not predict, so the cell appears in neither C1 nor C2. Over the eight CONFLICT families this single omission is exactly what separates the corpus's 202 assertions from the 210 a uniform 15 per family would give.

The $1 \rightarrow 2$ pair is where the three readouts are asked for three *different* things, and it is the clearest illustration of why the constraint classes are per readout rather than per rung. The edit is add_resolution: a concession is discharged. Then mm must *rise* (C1), because resolving a conflict raises belief in the claims it suppressed; cn must *fall* (C2), because it counts conflict activity and the discharge raises the both-true cell mass; and lp must be *exactly flat* (C2), because the edit is weight-only and the normalized log-partition is invariant to it. A measure that is monotone across all three has not implemented the concession discount, and the C2 class is what detects that.

Table 14: All 14 declared constraints of `f001`, checked against Table 13. 12 hold. A constraint is one assertion about one readout on one pair of rungs; Δ is the observed difference, classified against expected with a tie tolerance of 10^{-6}. The undeclared fifteenth cell (cn on $2 \rightarrow 3$) is absent by construction, not omitted. Both failures are discussed below; note that both are smaller in magnitude than every passing assertion.

Class	Readout	Rungs	Expected	Observed	Δ	OK
C1	mm	$0 \rightarrow 1$	increase	increase	$+0.010139$	✓
C1	cn	$0 \rightarrow 1$	increase	increase	$+0.149850$	✓
C1	lp	$0 \rightarrow 1$	increase	increase	$+0.032116$	✓
C1	mm	$1 \rightarrow 2$	increase	increase	$+0.029040$	✓
C2	cn	$1 \rightarrow 2$	decrease	decrease	-0.155011	✓
C2	lp	$1 \rightarrow 2$	invariant	decrease	-0.000444	×
C1	mm	$2 \rightarrow 3$	increase	decrease	-0.000340	×
C1	lp	$2 \rightarrow 3$	increase	increase	$+0.051099$	✓
cn *on* $2 \rightarrow 3$*: undeclared (unconstrained)*						
C1	mm	$3 \rightarrow 4$	increase	increase	$+0.033553$	✓
C1	cn	$3 \rightarrow 4$	increase	increase	$+0.023315$	✓
C1	lp	$3 \rightarrow 4$	increase	increase	$+0.063284$	✓
C3	mm	$0 \rightarrow 4$	increase	increase	$+0.072392$	✓
C3	cn	$0 \rightarrow 4$	increase	increase	$+0.024142$	✓
C3	lp	$0 \rightarrow 4$	increase	increase	$+0.146055$	✓

What the two failures show. Both are at the tie floor, and both are the same phenomenon rather than two. mm on $2 \rightarrow 3$ misses by -3.4×10^{-4}: the two rungs differ by one dropped CONTRADICTION edge against one added resolution, and on this family those nearly cancel on the belief readout, so the declared strict increase is not delivered. lp on $1 \rightarrow 2$ misses by -4.4×10^{-4} against a prediction of *exact* flatness. The prediction is sound — the edit is weight-only — but it is checked at a tie tolerance of 10^{-6} against values from weighted mini-bucket at i-bound 6, whose own approximation error is larger than that. The failure is therefore attributable to inference precision rather than to the model: the ORDER and CONTROL families pass their flatness assertions $45/45$ and $15/15$ only because their rungs compile to the *same* network, making those ties exact. We report the number as it is scored rather than widening the tolerance to absorb it, and note the consequence: `f001`'s $12/14$ is slightly above the corpus-wide CONFLICT rate of $89/112$, and one of its two failures is a statement about the inference algorithm Liu & Ihler (2011) at i-bound 6, not about the coherence model.

The same family under mined relations. Holding everything but the relation source fixed and re-running `f001` from graphs mined out of its own response text gives $4/14$, $8/14$, $7/14$ and $7/14$ for the four miner–policy combinations of Table 19, against $12/14$ from gold. The claims, the priors, the readouts and every scorer setting are byte-identical across those five numbers, so the gap is caused by the relation graph and nothing else — which is the paper's central empirical claim, visible here on one family.

F.7 THE TWO BASELINE AXES: INCREASE (50) AND INVARIANCE (2020 AXES

Table 1 scores baselines against 50 increase assertions and 20 invariance assertions, not against the 202 constraints of Table 19. The two counts are computed from the same `families.json`, and this subsection derives both, since a reader who assumes one is a subset-by-filtering of the other will find the percentages irreconcilable — $37/50$ is 74% where $167/202$ is 82.7%, and neither is an error.

Why a baseline cannot be scored on the 202**.** A constraint in the corpus sense is an assertion about *one readout* on one pair of rungs, and the three graded readouts are what multiply four adjacent pairs and one endpoint pair into the corpus's per-family 14 or 15 (see Section F.6). A baseline returns a single number per response. It has no mm as against cn as against lp, so the multiplier has nothing to multiply, and only the *readout-independent* content of a constraint transfers to it: the rung pair, and the direction asserted over it. Both axes below are therefore *deduplicated across readouts* — a rung pair is one assertion rather than one per readout — which is the single operation separating 70 from 202.

The increase axis (50)**.** Take the ten increase-type families, eight CONFLICT and two CHAIN. From each, take the C1 consecutive pairs and the C3 endpoint separation, deduplicated. Every one of the ten yields exactly the same five assertions:

$$\text{C1} : 0 \to 1, \quad \text{C1} : 1 \to 2, \quad \text{C1} : 2 \to 3, \quad \text{C1} : 3 \to 4, \quad \text{C3} : 0 \to 4,$$

the four adjacent pairs and the endpoint, for $10 \times 5 = 50$. An assertion is satisfied when the later rung scores strictly higher, by more than 10^{-6}.

Two consequences of the deduplication are worth stating, because they are what make the count uniform rather than ragged. First, **C2 is excluded by construction**, not by an arbitrary filter: it predicts the internal behaviour of a *particular* readout and a single-score baseline makes no claim a C2 constraint could test. Were it included, every baseline would be charged with failing a prediction never made about it. Second, and less obviously, **the undeclared cell that makes a CONFLICT family** 14 **rather than** 15 **disappears here**. That family's C1 block holds nine entries over four distinct pairs ($0 \to 1$ and $3 \to 4$ with three readouts each, $2 \to 3$ with two, $1 \to 2$ with one); cn on $2 \to 3$ is unconstrained, but lp still declares an increase on that pair, so the pair survives deduplication intact. A CONFLICT family declaring 14 and a CHAIN family declaring 15 therefore contribute 5 each, and the asymmetry visible in the 202 is invisible in the 50.

The invariance axis (20)**.** Take the four invariance-type families, three ORDER and one CONTROL, and take *every* rung pair each declares: the four adjacent pairs from C2 plus the C3 invariant endpoint, again deduplicated, for $4 \times 5 = 20$. Here C2 is *included*, which is not an inconsistency with the paragraph above: these families set flat = READOUTS, so every cell they declare is an invariance claim rather than a readout-specific prediction, and the readout-independent content of such a claim is exactly "this pair must not move". A pair is satisfied when the two scores differ by at most 10^{-6}. The pairs are read from each family's own constraints rather than from a hardcoded rung range, so a family with a different rung count would still be scored correctly.

One worked axis, on the example family. On `f001` the increase axis reduces the 14 declared constraints to five, scored here on mm: $0 \to 1$ ($+0.010139$), $1 \to 2$ ($+0.029040$), $2 \to 3$ (-0.000340), $3 \to 4$ ($+0.033553$) and the endpoint $0 \to 4$ ($+0.072392$) — four of five, the failure being the same near-cancelling $2 \to 3$ pair discussed in section F.6. Summing that computation over the ten increase families gives $4 + 3 + 5 + 4 + 3 + 3 + 3 + 3 + 4 + 5 = 37$ of 50, which is the LCS_{mm} row of Table 1. The corresponding per-readout totals are 37, 36 and 34 of 50 for mm, lp and cn, and $20/20$ on invariance for all three; these are asserted as a reproduction check in the reporting script, which refuses to emit any mined number if the gold rows drift from them, on the grounds that an axis definition that no longer reproduces the published gold figures cannot be trusted to score anything else.

What the two counts are and are not. Table 15 sets the two schemes side by side. The 70 assertions are a strict subset of the 202 in the sense that every rung pair scored on an axis is also scored in the corpus count, but they are *not* a subset arrived at by dropping rows: the readout dimension is collapsed, and on the increase axis an entire constraint class is removed. The percentages are therefore not comparable between Table 19 and Table 1, and we report them in separate tables with separate denominators rather than in one normalized column for exactly that reason.

Why two axes rather than one number. The split is the point rather than a presentational choice. The increase axis asks whether a measure can tell a more coherent response from a less coherent one; the invariance axis asks whether it stays put when nothing about coherence has changed. A measure is useful only if it does both, and the two are separately cheap to pass in ways that cancel: a constant

Table 15: The two counting schemes. Both read the same `families.json`; they differ in whether the readout dimension is retained. The right-hand column is what Table 1 uses and the left-hand what Table 19 uses; a percentage from one is not comparable with a percentage from the other. The final row is the one behavioural difference beyond the counting: in the corpus scheme an unscored rung is a *failure*, while on the axes it is dropped from the denominator — which matters only if a measure fails to return a score, and every measure reported here returned all seventy, so both denominators are full.

	Corpus constraints	Baseline axes
Count	202 per relation source	50 increase + 20 invariance
Unit	readout × rung pair	rung pair
Readouts	three, asserted separately	collapsed (single score)
Families	all 14, reported together	10 and 4, reported separately
C1 / C3	included	included
C2	included	increase: excluded; invariance: included
Tie tolerance	10^{-6}	10^{-6}
Used by	Table 19	Table 1
Unscored rung	counts as a failure	dropped from the denominator

function scores $20/20$ on invariance and $0/50$ on increase, while a measure tracking edit count can climb the increase ladder and move on every permutation. Averaging them into one figure would let a vacuous $20/20$ flatter exactly the rows that deserve it least — three of the baselines reach perfect invariance without measuring anything so the two axes are reported side by side and never combined.

F.8 Do readers agree? A human study

Every ordering in this section is declared by construction: the corpus states how consecutive rungs must rank, and a measure is scored against that declaration. A constructed ground truth is only worth having if readers agree it is the truth, so we ran a human study on a subset of the ladders. The falsifiable version of the claim is sharp — if readers cannot reproduce the declared orderings, then 82.7% against those orderings measures agreement with our perturbation recipe rather than with coherence, and every comparison in Table 1 loses its referent.

Design. Four annotators each judged the same 14 screens, a complete design of 56 ratings with no missing data. A screen shows two responses and the query that produced them, and asks which is more logically coherent — A, B, or about equal — with a confidence rating, a free-text reason, and a checkbox for whether the two looked like the same content reordered. The screens cover seven families across all four ladder types and seven topics: 10 *increase* screens, where the corpus declares one side more coherent, and 4 *invariance* screens, where it declares the two equal. Families were chosen on *measured* disagreement between the LCS and the strongest evidence-free baseline rather than on convenience, and one family was kept deliberately because the baseline wins there. Two pairs per family bracket the ends of each ladder: the $(0, 4)$ endpoint assertion, the largest intended gap, and the $(0, 1)$ adjacent step, the hardest. The CONTROL ladder is symmetric by construction, so its $(0, 4)$ pair is byte-identical and would ask a reader to compare a text with itself; it contributes $(0, 2)$ and $(1, 2)$ instead. Annotators saw only the query and the two responses.

Agreement. Krippendorff's α (nominal) is 0.748 over all screens, 0.744 on the increase screens and 0.667 on the invariance screens. Per-annotator accuracy against the declared answer on the increase screens is 9, 9, 7 and 7 of 10. The median time per screen is 15.5 minutes.

Readers reproduce the increase ladders. Eight of the ten increase screens have a majority matching the declared answer, and 32 of the 40 individual ratings agree with it. At the rating level that is far from chance ($p = 1.8 \times 10^{-9}$ against a uniform choice over the three options, $p = 9.1 \times 10^{-5}$ against a coin flip between A and B). At the screen level, $8/10$ gives $p = 0.003$ against the uniform null but only $p = 0.055$ against a coin flip, which does not clear 0.05: with ten screens the majority-level test is underpowered, and we state that rather than quote the favourable null alone. The free-text reasons say what readers were responding to. Of the 55 ratings of 56 that carry a written reason, 32 name a relational defect — a contradiction, a presupposition used before it is introduced, a broken

Table 16: The four invariance screens, split by ladder type. Both ladders are built from operators that provably move no factor, so the LCS is exactly invariant across every rung of both — to six decimals, $\mathrm{LCS}_{\mathrm{mm}} = 0.784182$ on all five ORDER rungs and 0.788987 on all five CONTROL rungs. Readers separate the two cases sharply. Fisher's exact test on equal-versus-not by ladder gives $p = 0.007$; the ratings within a ladder are four annotators over two screens of one family rather than eight independent trials, so this is a descriptive contrast rather than a family-level significance claim.

Ladder	Screens	"Equal" ratings	LCS	What readers said
CONTROL	s013, s014	$6/8$	invariant	the pair is the same content
ORDER	s011, s012	$0/8$	invariant	the shuffled response is worse

chain of support — and only 2 mention surface fluency or style. Readers were reasoning about structure, not readability.

One invariance requirement is validated and the other is not. Pooled, the invariance screens look like a failure: 1 of 4 majorities match. Split by ladder type they are two different results, and the pooled fraction is an average of them (Table 16).

On CONTROL, which exchanges Precedence↔Succession while preserving chronology, six of eight ratings say the pair is equal and one screen is unanimous. That invariance requirement is human-validated, and the LCS's exact flatness there is a property readers endorse. On ORDER, which permutes sentences, *not one* of the eight ratings called the pair equal, and all eight preferred the same side — the less aggressively shuffled one. On both screens the rejected side was the shuffle_full rung, and on one of them readers preferred a mere Precedence swap over it, so they were tracking the *degree* of scrambling that the shuffle levels encode.

The reasons name a specific and checkable defect: dangling anaphora. Two annotators independently flagged that the shuffled response "relies on *that deficit* and *that stress* before those concepts are introduced" and "contains summary-level language to concepts before they have been introduced". This is verifiable in the corpus text. In the base response the phrase *that stress* occurs after the sentence introducing drought-induced stress; in shuffle_full it occurs roughly $1{,}500$ characters earlier, with no antecedent anywhere before it. The claim set is identical, every relation is identical, and the perturbation record says so in as many words: *sentence order only; edge set unchanged*.

So the LCS is invariant here for a principled reason and readers disagree for an equally principled one. Atoms are order-free by construction — nothing in Eq. 1 references claim position — which is what makes that determinism provable; the cost is that referential cohesion between sentences is not representable at all, and it is a real component of what readers call coherence. The ORDER ladder's $20/20$ invariance is therefore not a human-referenced success but a measured blind spot, and we count it as a limitation rather than as evidence. The CONTROL result is the human-referenced success, and it is the one the invariance claim should rest on.

Measures against the human majority. Scoring each measure the way a reader was scored — mapping the sign of its two rung scores onto the side that held the higher rung, with the ladder scorer's own 10^{-6} tie tolerance so that float noise reads as "equal" rather than as a preference — gives the only comparison in this paper where a measure is graded against human judgement rather than against a construction. The twelve screens with a majority are the denominator; a measure cannot be credited for agreeing with a majority that does not exist.

Three readings, in descending order of how comfortable they are. First, on a correct graph the LCS leads both columns, at $9/12$ against the best baseline's $8/12$ and $14/14$ against the best baseline's $9/14$; and its three misses are exactly the cases already discussed — the two ORDER screens, plus one screen where four readers unanimously called a single-conflict repair too subtle to matter while the model separated the pair. Second, the baseline ranking is not the ladder ranking: the length control reaches $7/12$, which is a reminder that fourteen screens cannot separate measures finely, and the direct judge at $8/12$ does far better against readers than against the declared orderings ($5/14$) — a judge and a human agree partly because both read prose. Third and least comfortable, the mined arms do not beat the best baseline here. At $7/12$ and $3/12$ they are well below the gold row, which is the same gold-versus-mined gap as the findings of our experimental section, now measured against readers instead of against the corpus. The human study localizes the bottleneck to relation extraction

Table 17: Agreement with the human majority on the twelve screens that have one, beside agreement with the declared ordering on all fourteen. LCS rows are the mm readout. The gold row is the model given a correct graph; the mined rows are the end-to-end pipeline. Only measures covering all fourteen screens are listed. Generated by `scripts/analyze_human_study.py`.

Measure	vs. human majority	vs. declared
$\mathrm{LCS}_{\mathrm{mm}}$, gold graph	$\mathbf{9}/12$	$\mathbf{14}/14$
$\mathrm{LCS}_{\mathrm{cn}}$, gold graph	$8/12$	$13/14$
$\mathrm{LCS}_{\mathrm{mm}}$, mined (`gpt-oss`, bidir.)	$7/12$	$7/14$
$\mathrm{LCS}_{\mathrm{mm}}$, mined (`llama`, windowed)	$3/12$	$4/14$
JUDGE direct (`gpt-oss`)	$8/12$	$5/14$
LENGTH control	$7/12$	$9/14$
JUDGE G-Eval (`gpt-oss`)	$6/12$	$7/14$
ROSCOE-SC mean (`gpt-oss`)	$6/12$	$6/14$
RC cohesion	$4/12$	$8/14$
CLAIM COUNT control	$2/12$	$4/14$
CONTRADICTION RATE (`gpt-oss`)	$1/12$	$2/14$

rather than to the scoring model, and it does not show that the deployable configuration tracks human judgement.

Limitations. Our human study establishes that the increase ladders track human judgement, so the declared orderings are not an artefact of our recipe; that the CONTROL invariance requirement is one human endorse; and that the ORDER invariance requirement is not. The study does not establish human correlation at scale. Fourteen screens over seven of the fourteen families, four annotators and a single judgement per screen is a *validation of the corpus*, not a benchmark of measures against human coherence ratings; the comparison above should be read as a consistency check with a known answer, and the wide confidence intervals implied by $n = 12$ are why we draw no ranking conclusion from the one-screen differences among the baselines.

G EXPERIMENTS

In this section, we present additional experimental results on LOCOBENCH and five long-form factuality benchmarks: BIO, ASKHIST, BOOKS, ELI5, and LFOBJ. Unless otherwise stated, all probabilistic inference is performed exactly using the open-source Merlin inference engine (https://github.com/radum2275/merlin), which provides the MRF inference backend for all reported coherence scores.

G.1 RESULTS ON THE LOCOBENCH BENCHMARK

We begin with the estimator, since every mined arm inherits its errors. Measuring it against gold separates a model that scores a graph badly from an extractor that supplies the wrong graph.

Relation Estimation is the Bottleneck Table 18 evaluates relation extraction by pairing each miner with a candidate-pair policy while holding the downstream coherence model fixed. The WINDOWED policy considers only forward claim pairs within a radius of four claims, whereas BIDIRECTIONAL considers both directions for all pairs, nearly doubling the candidate set. Each candidate is processed by Prompt A (relation type) and, if a relation is detected, Prompt B (relation strength). Extracted relations are compared against the 603 gold relations at three levels: *pair* (detecting that a relation exists), *coupling* (predicting the correct inferential coupling), and *sense* (predicting the correct discourse sense). We report precision (P), recall (R) and the F1 score, respectively. The results reveal that extraction errors arise primarily from relation typing rather than relation discovery: e.g., `llama-3.3-70b` under WINDOWED recovers most related claim pairs but loses many of them through incorrect coupling or sense assignments. Precision is equally important, since spurious relations introduce additional factors into the MRF and actively distort posterior beliefs rather than merely adding noise. This explains the central trade-off in the table: although BIDIRECTIONAL substantially improves recall of backward-directed relations, the accompanying increase in false

Table 18: MINED vs GOLD agreement, micro-averaged over the 70 items of LOCOBENCH.

		Pair			Coupling			Sense	Recall by dir.		Errors	
Miner/Policy	Edges	P	R	F1	P	R	F1	F1	fwd	bwd	non-rel	dup
mined by `llama-3.3-70b`												
WINDOWED	1,106	44.8	82.3	58.0	34.1	62.5	44.1	37.6	71.1	21.9	34	0
BIDIRECTIONAL	3,172	28.5	91.7	43.5	17.9	79.3	29.3	24.7	79.9	76.2	119	1,231
mined by `gpt-oss-120b`												
WINDOWED	1,094	46.0	83.4	59.3	40.2	73.0	**51.9**	**47.0**	82.1	29.5	38	0
BIDIRECTIONAL	2,973	29.8	96.5	45.5	22.4	88.4	35.8	32.3	89.0	85.7	147	1,020

Table 19: Left: declared ordering constraints satisfied, grouped by ladder type (CONFLICT, CHAIN, ORDER, CONTROL). Right: mean and standard deviation of LCS scores (over 70 items).

	Constraints satisfied					Mean LCS value ($\pm$sd)		
Miner / policy	Conflict	Chain	Order	Control	%	$\mathrm{LCS}_{\mathrm{mm}}$	$\mathrm{LCS}_{\mathrm{cn}}$	$\mathrm{LCS}_{\mathrm{lp}}$
GOLD	89/112	18/30	45/45	15/15	**82.7**	0.795 $\pm$0.022	0.515 $\pm$0.135	0.410 $\pm$0.090
mined by `llama-3.3-70b`								
WINDOWED	53/112	17/30	0/45	0/15	34.7	0.796 $\pm$0.089	0.676 $\pm$0.239	0.224 $\pm$0.158
BIDIRECTIONAL	57/112	18/30	0/45	0/15	37.1	0.723 $\pm$0.171	0.542 $\pm$0.196	0.037 $\pm$0.026
mined by `gpt-oss-120b`								
WINDOWED	59/112	12/30	0/45	0/15	35.1	0.825 $\pm$0.074	0.500 $\pm$0.138	0.129 $\pm$0.096
BIDIRECTIONAL	53/112	17/30	0/45	0/15	34.7	0.808 $\pm$0.128	0.462 $\pm$0.095	0.033 $\pm$0.059

positives reduces overall coupling F1. Across all settings, `gpt-oss-120b` achieves the highest pair, coupling, and sense F1 scores, but its strength estimates are heavily concentrated at the extremes, providing substantially less graded uncertainty than `llama-3.3-70b`. Overall, the results identify relation extraction, and in particular accurate relation typing, as the primary bottleneck rather than the coherence model itself.

The Model Orders Correctly when the Graph is Correct Table 19 evaluates whether coherence scores satisfy the ordering constraints defined by LOCOBENCH. Each miner uses the same claims, priors, readouts, and scoring procedure; only the source of the relation graph varies. The left half of the table reports the fraction of the 202 declared constraints satisfied, while the right half reports the mean and standard deviation of the resulting coherence scores over the 70 benchmark responses.

The results reveal a clear separation between probabilistic scoring and relation extraction. With GOLD relation graphs, the three LCS variants satisfy 167 of 202 constraints (82.7%), including perfect performance on the ORDER (45/45) and CONTROL (15/15) ladders. Performance is strongest on CONFLICT ladders (89/112) and weaker on CHAIN ladders (18/30), indicating that missing-support failures are more challenging than explicit contradictions. In contrast, all four mined configurations satisfy only 34.7–37.1% of constraints despite using the same scoring functions. Since the graph source is the only difference between arms, this 46–48 point gap is attributable entirely to relation extraction.

The per-ladder breakdown localizes the failure. Mined graphs remain moderately effective on CONFLICT (53–59 of 112) and CHAIN (12–18 of 30) ladders, but fail completely on ORDER and CONTROL (0/45 and 0/15 in every configuration). These ladders preserve the claim set and intended argument structure while applying only meaning-preserving edits, so a coherence measure operating on a correct graph should remain unchanged. The fact that every mined configuration fails all invariance tests shows that current LLM-based relation extractors are highly sensitive to surface form and sentence order. Consequently, the main limitation of the end-to-end pipeline is not probabilistic scoring but the reliability of the extracted relation graph. The central takeaway is therefore that the proposed coherence model orders responses correctly when provided with the intended argument structure, while accurate and invariant relation extraction remains the primary open challenge.

Table 20: The mined relation inventory on the LoCoBench corpus (70 items), as a percentage of all relations emitted, at both levels of the taxonomy. *Gold* is the corpus's own distribution over its 603 edge-producing relations. Level-1 couplings select the factor table; Level-2 senses compile to them through $\mathcal{C}$, and the sense rows are ordered by their gold share so a mined column can be read against it down the page. The final block is gold-referenced recall of CONTRADICTION specifically: of the 164 gold contradictions, the fraction each source recovered with the correct coupling. WIN is the forward-only WINDOWED policy, BID the BIDIRECTIONAL one.

	`llama-3.3-70b`		`gpt-oss-120b`		
	WIN	BID	WIN	BID	Gold
Relations mined per item	15.8	45.3	15.6	42.5	8.6
Level-1 coupling (% of relations mined)					
ENTAILMENT	47.2	49.9	43.0	46.9	32.8
CONTRADICTION	22.5	22.4	**27.1**	**28.3**	27.2
EXCLUSIVE	15.1	12.6	12.4	10.2	12.9
CO_NECESSITY	8.0	10.2	6.2	6.6	11.6
EQUIVALENCE	7.1	5.0	11.2	8.0	15.4
Level-2 sense (% of relations mined), ordered by gold share					
Restatement	7.1	4.9	11.2	8.0	15.4
Contrast	14.7	17.1	22.8	24.6	14.9
Evidence	32.6	35.5	21.9	22.2	13.6
Alternative	15.0	12.6	12.4	10.2	12.9
Concession	7.9	5.5	4.4	3.8	12.3
Disjunction	8.0	10.1	6.2	6.6	11.6
Cause-Effect	2.6	2.6	6.3	8.2	11.4
Instantiation	9.9	9.6	7.6	6.9	6.3
Effect-Cause	0.0	0.3	1.4	3.6	0.8
Condition	0.9	1.1	5.8	6.1	0.7
ordering-only†	1.3	0.9	0.0	0.0	0.0
Gold-referenced CONTRADICTION *recall (%)*					
of the 164 gold contradictions	48.2	56.7	**70.1**	**76.2**	—

† Precedence, Succession and None, folded into one line. These compile to Level-1 NONE and contribute no factor; the ladder pipeline records their pre-compile coupling, so they are counted under ENTAILMENT in the block above, inflating it by at most 1.3 points for `llama` and not at all for `gpt-oss`, which emits none of them.

Table 21: The factuality corpus: 50 responses from each of five long-form datasets, drawn with a fixed seed and stratified so that no dataset dominates, for a total of 250 responses. Atom counts are per response.

Dataset	Domain	Items	Atoms/item	Relations/item*
BIO	biographies (Min et al., 2023)	50	31.1	85.3
ASKHIST	history questions (Marinescu et al., 2025)	50	23.2	87.5
BOOKS	book questions (Marinescu et al., 2025)	50	24.0	89.0
ELI5	explanatory answers (Marinescu et al., 2025)	50	21.1	76.0
LFOBJ	long-form objective (Wei et al., 2024)	50	25.9	111.2
Total		**250**	25.1	89.8

* mined relations per response, `llama-3.3-70b` under the WINDOWED policy.

MINED relation inventory on LOCOBENCH. Table 20 summarizes the distribution of relations extracted by the two miners at both levels of the taxonomy. Both miners substantially over-generate relations relative to the gold graphs, producing on average 15.6–15.8 relations per item under WINDOWED extraction and 42.5–45.3 under BIDIRECTIONAL, compared with 8.6 gold relations. At the coupling level, entailment is consistently over-produced (43.0–49.9% versus 32.8% in gold), whereas equivalence and co-necessity are under-represented. Contradictions constitute 27.2% of gold relations; `gpt-oss-120b` closely matches this distribution (27.1–28.3%), while `llama-3.3-70b` under-produces contradictions (22.4–22.5%). This difference is reflected in contradiction recall: `gpt-oss` recovers 70.1% and 76.2% of the 164 gold contradictions under the WINDOWED and BIDIRECTIONAL policies, compared with 48.2% and 56.7% for `llama`.

The Level-2 inventory reveals where these discrepancies originate. Both miners over-predict generic support relations, particularly Evidence, which accounts for 21.9–35.5% of extracted relations but only 13.6% of gold relations. In contrast, Concession (3.8–7.9% versus 12.3% in gold) and Cause-Effect (2.6–8.2% versus 11.4%) are systematically under-produced. These are among the most informative discourse relations for coherence assessment, suggesting that current extractors preferentially recover generic support links while missing more nuanced argumentative structure. Neither miner dominates across all senses: `gpt-oss` is more accurate overall at the coupling level and substantially better at contradiction recovery, whereas `llama` more closely matches the gold frequency of rare senses such as Condition (0.9–1.1% versus 0.7% in gold, compared with 5.8–6.1% for `gpt-oss`). Finally, bidirectional extraction consistently increases recovery of backward-directed relations, such as Effect-Cause, but often does so at the cost of substantially higher relation density. Overall, the inventory suggests that current LLM miners recover the broad coupling distribution reasonably well but struggle to reproduce the finer-grained discourse structure required for robust coherence assessment.

G.2 RESULTS ON THE FACTUALITY CORPUS

Corpus stats. Table 21 reports summary statistics for the five factuality datasets, including the average number of atomic claims and extracted discourse relations per response. The relation counts are produced by the relation mining pipeline described in Section 2.6, using `llama-3.3-70b` to identify and label relations between claims.

Table 22 summarizes the claim-level factuality estimates produced by Stage 1, which serve as priors for the coherence MRF. Across all 6,266 claims in the evaluation corpus, the mean posterior factuality is high for both backends (0.912 for `llama-3.3-70b` and 0.853 for `gpt-oss-120b`), indicating that most claims are judged plausible given the retrieved evidence. To assess whether these scores carry meaningful signal, we exploit the BIO subset, the only dataset with human claim-level labels. On the 1,555 labeled claims, both backends clearly separate supported from unsupported claims: `llama` assigns mean scores of 0.933 versus 0.576 (separation +0.357), while `gpt-oss` assigns 0.887 versus 0.523 (separation +0.364). Although the two models differ in calibration, they achieve nearly identical class separation, suggesting that the factuality priors used by the coherence model capture genuine evidence-based distinctions rather than merely reflecting overall confidence.

Factuality and LCS scores. Table 23 reports the mean and standard deviation of the factuality score together with the four coherence scores across the five datasets and two relation-mining backends

Table 22: Stage-one factuality marginals, per backend. Every cell is a mean of *per-atom posterior marginals* $\pi_i = \mathbb{P}(a_i{=}1)$ — the values handed to the coherence MRF as its priors — and means are taken over atoms, pooling across responses, so a 40-atom response contributes forty times what a one-atom response does. **The two blocks have different denominators.** *All atoms* covers all 6,266 atoms of the 250-response sample, and **sd** is the spread of π_i across those atoms rather than any between-response or run-to-run variation. The right-hand block is the external check and covers only the BIO subset: 1,555 atoms, a quarter of the sample, being the 50 BIO responses, which are the only ones in either corpus carrying human atom-level labels. **sep.** is the supported mean minus the not-supported mean. Both backends separate the two classes by nearly the same margin while disagreeing on the absolute level.

	All atoms (6,266)		BIO only (1,555): vs. human labels		
Backend	mean	sd	supported	not supported	sep.
`llama-3.3-70b`	0.912	0.228	0.933	0.576	**+0.357**
`gpt-oss-120b`	0.853	0.289	0.887	0.523	**+0.364**

The BIO block partitions its 1,555 labelled atoms exactly: 992 supported and 563 not-supported. The remaining 4,711 atoms of the sample carry no human label, so they enter the left block and not the right.

Table 23: Mean and standard deviation: factuality score and four LCS scores per dataset.

Backend	Dataset	factuality	$\mathrm{LCS_{mm}}$	$\mathrm{LCS_{cn}}$	$\mathrm{LCS_{re}}$	$\mathrm{LCS_{lp}}$
`llama-3.3-70b`	BIO	0.665 ±0.211	0.945 ±0.103	0.922 ±0.171	0.498 ±0.008	0.929 ±0.232
	ASKHIST	0.897 ±0.112	0.967 ±0.078	**0.589** ±0.198	0.364 ±0.205	**0.287** ±0.431
	BOOKS	0.891 ±0.153	0.982 ±0.069	0.820 ±0.250	0.442 ±0.145	0.721 ±0.431
	ELI5	0.963 ±0.065	0.995 ±0.013	0.757 ±0.252	0.405 ±0.180	0.603 ±0.472
	LFOBJ	0.974 ±0.033	0.990 ±0.020	0.927 ±0.174	0.469 ±0.107	0.885 ±0.312
`gpt-oss-120b`	BIO	0.602 ±0.194	0.877 ±0.101	0.776 ±0.238	0.472 ±0.100	0.902 ±0.234
	ASKHIST	0.769 ±0.151	0.942 ±0.091	**0.522** ±0.136	0.488 ±0.046	**0.377** ±0.432
	BOOKS	0.789 ±0.196	0.959 ±0.110	0.729 ±0.276	0.482 ±0.075	0.658 ±0.432
	ELI5	0.878 ±0.138	0.982 ±0.042	0.668 ±0.238	0.481 ±0.084	0.544 ±0.471
	LFOBJ	0.935 ±0.067	0.977 ±0.033	0.859 ±0.219	0.464 ±0.123	0.875 ±0.301

Table 24: Results of the ablation: factuality priors vs. uniform priors on the factuality corpus

			factuality priors			uniform priors		
Miner	Policy	Readout	mean	d	ρ	mean	d	ρ
`llama-3.3-70b`	WIN	$\mathrm{LCS_{mm}}$	0.976	90	+0.665	0.955	**196**	+0.398
		$\mathrm{LCS_{cn}}$	0.803	74	+0.347	0.815	**163**	+0.212
		$\mathrm{LCS_{lp}}$	0.685	45	+0.265	0.772	**86**	+0.339
		$\mathrm{LCS_{re}}$	0.436	69	+0.393	0.451	**86**	+0.020
	BID	$\mathrm{LCS_{mm}}$	0.980	79	+0.626	0.948	**171**	+0.369
		$\mathrm{LCS_{cn}}$	0.804	59	+0.296	0.805	**147**	+0.184
		$\mathrm{LCS_{lp}}$	0.661	35	+0.287	0.735	**58**	+0.297
		$\mathrm{LCS_{re}}$	0.427	51	+0.381	0.441	**63**	+0.204
`gpt-oss-120b`	WIN	$\mathrm{LCS_{mm}}$	0.948	148	+0.674	0.908	**224**	+0.169
		$\mathrm{LCS_{cn}}$	0.711	134	+0.361	0.704	**212**	+0.080
		$\mathrm{LCS_{lp}}$	0.671	85	+0.165	0.764	**120**	+0.188
		$\mathrm{LCS_{re}}$	0.478	**37**	+0.407	0.491	32	−0.027
	BID	$\mathrm{LCS_{mm}}$	0.960	111	+0.560	0.909	**195**	+0.103
		$\mathrm{LCS_{cn}}$	0.721	87	+0.266	0.708	**169**	+0.046
		$\mathrm{LCS_{lp}}$	0.620	58	+0.131	0.726	**85**	+0.173
		$\mathrm{LCS_{re}}$	0.464	**24**	+0.380	0.481	15	+0.124

WIN is the forward-only WINDOWED candidate policy, BID the BIDIRECTIONAL one. All ρ are against the same quantity, the stage-one `factuality_score` of the priors-on run. One item of the 250 mines no relations under uniform priors and therefore scores at the prior throughout; it is retained rather than dropped.

(`llama-3.3-70b` and `gpt-oss-120b`). A notable pattern is that coherence does not simply mirror factuality. Under `llama`, factuality increases almost monotonically across datasets, from 0.67 (BIO) to 0.97 (LFOBJ), whereas $\mathrm{LCS_{cn}}$ follows a markedly different trend, taking values 0.92, 0.59, 0.82, 0.76, and 0.93, respectively. The clearest example is ASKHIST, whose responses are highly factual yet among the least coherent according to $\mathrm{LCS_{cn}}$. The same qualitative pattern is observed under `gpt-oss`, suggesting that coherence and factuality capture complementary aspects of response quality.

One possible explanation is that the coherence scores simply inherit the factuality estimates used as priors in the second-stage MRF. The observed divergence argues against a complete collapse of the two measures, but the relationship remains non-trivial. Indeed, coherence scores remain positively correlated with factuality, with Spearman correlations of +0.67 for $\mathrm{LCS_{mm}}$ and +0.35 for $\mathrm{LCS_{cn}}$. The stronger correlation of $\mathrm{LCS_{mm}}$ is expected, since the score averages posterior claim marginals whose priors are supplied by the factuality model. Distinguishing genuine agreement between factuality and coherence from inheritance through the prior therefore requires removing the factuality stage altogether, which we investigate in the ablation study below.

Factuality priors vs. uniform priors ablation Table 24 isolates the effect of the factuality stage by replacing the factuality-derived priors with uniform priors ($\pi_i \equiv \frac{1}{2}$) while keeping the responses, mined relation graphs, and coherence readouts unchanged. We report the mean score, the number of distinct values (d) produced across the 250 responses, and the Spearman correlation (ρ) with the factuality score. Two trends are consistent across almost all configurations. First, removing factuality priors substantially increases discrimination: the number of distinct scores increases in 14 of 16 readout–policy combinations. For example, under `llama`/WIN, $\mathrm{LCS_{mm}}$ increases from 90 to 196 distinct values and $\mathrm{LCS_{cn}}$ from 74 to 163, while under `gpt-oss`/WIN the corresponding increases are from 148 to 224 and from 134 to 212. Thus, factuality priors improve interpretability but compress the dynamic range of the coherence scores.

Second, removing the priors generally weakens the association with factuality. For $\mathrm{LCS_{mm}}$, the correlation drops from 0.665 to 0.398 under `llama`/WIN and from 0.674 to 0.169 under `gpt-oss`/WIN, indicating that part of the observed coherence–factuality relationship is inherited through the prior slot. A similar pattern is observed for $\mathrm{LCS_{re}}$, whose correlation falls from 0.393 to 0.020 under `llama`/WIN and from 0.407 to −0.027 under `gpt-oss`/WIN. The notable exception is $\mathrm{LCS_{lp}}$, whose correlation remains stable or even increases slightly (e.g., $0.265 \rightarrow 0.339$ under `llama`/WIN), consistent with its definition as a normalized partition function rather than a marginal-based score.

Overall, the ablation shows that factuality priors contribute to the observed coherence–factuality correlation, but coherence scores retain substantial independent variation once those priors are removed.

G.3 Summary of the Results

Our experiments support three main conclusions. First, the probabilistic coherence model is effective when provided with the correct argument structure: using gold relation graphs, the proposed scores satisfy 82.7% of the declared LoCoBench constraints and are the only measures that both order responses by coherence ($37/50$) and remain invariant under meaning-preserving edits ($20/20$). Second, relation extraction, rather than probabilistic scoring, is the dominant source of error. Replacing gold graphs with mined ones reduces constraint satisfaction to 34.7–37.1%, and every mined configuration fails all invariance tests despite remaining moderately competitive on ordering. This identifies robust discourse-relation extraction as the primary bottleneck of the end-to-end pipeline. Third, the proposed coherence scores capture information that is distinct from both factuality and existing coherence proxies. On real-world corpora, coherence and factuality rank datasets differently, with AskHist providing the clearest example of high factuality but relatively low coherence. Ablations further show that factuality-derived priors account for part, but not all, of the observed coherence–factuality correlation. At the same time, removing these priors substantially increases score resolution, revealing a trade-off between interpretability and discriminative power. Across both synthetic and real-world evaluations, the proposed scores remain substantially more discriminative than existing LLM judges and contradiction-based baselines, while additionally providing claim-level explanations through their posterior marginals. Taken together, the results suggest that logical coherence is a relational property that cannot be reduced to factuality, cohesion, contradiction detection, or holistic LLM judgments, and that future progress depends primarily on improving discourse-relation extraction rather than developing new scoring functions.