# SPAN: STRUCTURE-PRESERVING ANALOGICAL NARRATIVES FROM SCIENTIFIC TEXTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Scientific explanations encode interacting entities, constraints, and causal dependencies that are difficult to preserve in narrative form. Large language models readily generate analogical stories, but these outputs often trade structural fidelity for fluency, producing surface substitutions, mechanistic retellings, or fictional events whose behavior is insufficiently grounded. We introduce **SPAN (Structure-Preserving Analogical Narratives)**, a five-stage framework that transforms scientific explanations into self-contained fictional narratives through causal rewriting, causal-mechanistic graph construction, functional-role and **causal-affordance grounding**, story-world synthesis, and causally audited realization. SPAN is built around one requirement: important story events must follow from stable properties established within the fictional world, not merely because the source mechanism requires an analogous event. On GPT-4o, SPAN with world-level synthesis (SPAN-W) reaches an analogical-fidelity score of **0.972** and a narrative-quality score of **0.826**, improving over the strongest baselines by **0.264** and **0.262**. In human evaluation, **27.7%** of its stories are accepted without modification and only **7.4%** are rejected, whereas all tested baselines receive **0% Accept**. Ablations identify causal explication and affordance grounding as the clearest contributors. Results on gpt-oss-120B and a 70B model show that the gains remain dependent on model capability, exposing a gap between analogy recognition and analogical generation. We further observe a trade-off between analogical fidelity and narrative quality, where stronger structural preservation can reduce storyness, while greater narrative autonomy can weaken analogical alignment, especially in samller models.

## 1 INTRODUCTION

Scientific understanding rarely depends on isolated facts. Explanations of biological, physical, chemical, and computational processes require learners to coordinate interacting entities, constraints, state changes, and causal dependencies, that is, to build a mental model of *how a system behaves and why*. Analogies address this by relating unfamiliar structures to familiar domains (Duit, 1991; Glynn & Takahashi, 1998; Harrison & Treagust, 1993), but their value depends on the mapping: well-designed analogies support understanding and retention, whereas inappropriate correspondences introduce misconceptions. Structure-Mapping Theory accordingly characterizes strong analogy by correspondence between relational systems rather than surface similarity (Gentner, 1983), and research on narrative explanation argues that scientific ideas become more meaningful when organized as temporally and causally connected events (Norris et al., 2005; Avraamidou & Osborne, 2009). Together, these traditions motivate **analogical narrative**: understanding an unfamiliar mechanism through a different world whose relevant relationships behave in corresponding ways.

Recent NLP work has extended computational analogy to paragraph- and story-level reasoning (Jiayang et al., 2023; Sourati et al., 2024; Sultan et al., 2024; Ye et al., 2024), with substantial progress in *recognizing* structural similarity. Yet recognizing an analogy and generating a useful analogical story are different problems. Fluency does not guarantee that the mechanism survives the transformation; conversely, increasingly explicit structural mappings can preserve the mechanism while destroying the story. In preliminary analyses, direct prompting omitted dependencies or replaced them with thematic resemblance; local relational mappings failed to compose into a globally coher-



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.

ent narrative; role-based representations became procedural, with fictional entities appearing only to execute their assigned operation; and heavy constraint prompting produced worlds whose rules changed exactly when the analogy needed a particular outcome.

These failures expose a limitation of treating analogical generation as entity substitution or relation preservation alone. Mapping a semipermeable membrane to a guarded gate is intuitively reasonable, but it becomes explanatory only if the gate has a stable rule that reproduces the relevant selectivity; a gate that opens selectively merely because the membrane must do so explains nothing. Likewise, a flexible cellular boundary cannot simply become a tent that contracts at the right moment unless some property of the tent, established independently of the plot, makes the contraction follow from changes in its contents or surroundings. The question is therefore not only *which fictional entity corresponds to the scientific entity*, but *what property of that entity makes the corresponding behavior possible*. We call this requirement **causal affordance preservation**. A counterpart may be functionally appropriate yet causally incapable of its role, so we distinguish

$$\text{functional correspondence} \neq \text{causally realizable correspondence}$$

and require each important story event to follow from a prior state, a story-world affordance, and a trigger or constraint. This shifts the problem from structure-preserving renaming to **structure-preserving world construction**, in which readers can reason within the fictional world without continually translating back to the science.

**Motivating example.** On B cells, monomeric Immunoglobulin M (IgM) recognizes antigens; after activation, secreted IgM assembles into pentameric or hexameric complexes that trap pathogens. Because IgM appears early, it signals an acute response that is later superseded by longer-lasting antibodies. An analogical rendering:

> At the gates of the fortress stood a quiet Listener beside the young Bee Archers. The Listener carried no sword; its task was to watch for the first sign of an intruder. One night, strange invaders approached and the warning bells rang. Soon Listeners rode beyond the fortress into the open roads. There, groups of them locked arm in arm into enormous clasping rings, surrounding the invaders before they could spread deeper into the kingdom. Whenever the rings appeared, the people knew the danger was fresh. Yet the Listeners were never meant to remain forever. Once the first assault had been contained, they faded into the background while the kingdom prepared stronger defenses for the longer war.

The Listener does not resemble an antibody. Instead, surface recognition maps to gate surveillance, secretion to deployment beyond the fortress, multi-unit assembly to clasping rings, pathogen trapping to encirclement, and an early, temporary response to evidence of a fresh attack followed by longer-term defenses. Roles, temporal order, and consequences remain aligned while the vocabulary changes entirely.

We introduce **SPAN**, which decomposes generation into five persisted stages: a causally explicit scientific rewrite, a causal-mechanistic graph, functional-role and causal-affordance grounding, story-world synthesis under stable rules, and causally audited narrative realization. Each intermediate representation is inspectable, which also enables ablations that reveal *which* failure emerges when a stage is removed. Because a story can be faithful yet fail as fiction, or compelling yet distort the mechanism, we evaluate analogical fidelity and narrative quality on separate axes, complemented by human pairwise and Accept/Repair/Reject judgments. Our contributions are:

1. **Structure-preserving analogical generation.** SPAN, a five-stage framework that transforms scientific explanations into autonomous fictional narratives via causal rewriting, causal-mechanistic representation, role–affordance grounding, world construction, and audited realization.
2. **Causal affordance grounding.** A distinction between functional and causally realizable correspondence, requiring fictional counterparts to possess stable properties that explain how their mapped behavior can occur.
3. **Joint evaluation of analogy and story.** Separate science-aware analogical and science-blind narrative axes, complemented by human preference and usability judgments.
4. **Empirical analysis.** On GPT-4o, SPAN-W is the only system whose stories humans accept without modification (27.7% vs. 0% for all baselines); ablations isolate causal explication

and affordance grounding, and experiments across model capacities show that representation alone cannot remove generator limitations.

## 2 Related Work

**Analogy and narrative for scientific explanation.** Structure-Mapping Theory holds that useful analogies are defined by correspondences between systems of relations rather than resemblance between entities (Gentner, 1983). In science education, analogies support understanding and retention but introduce misconceptions when mapped relationships are inappropriate or incomplete (Duit, 1991; Glynn & Takahashi, 1998; Harrison & Treagust, 1993). A complementary literature studies narrative as a medium for scientific explanation, organizing information through temporally and causally connected events (Norris et al., 2005; Avraamidou & Osborne, 2009); its effectiveness depends on narrative form, learners, setting, and content (Soares et al., 2023). SPAN asks the computational question of how to automatically construct a fictional narrative that stays structurally faithful to the mechanism it represents.

**Computational analogy over narratives.** StoryAnalogy evaluates identification and generation of analogous stories using structure-mapping principles (Jiayang et al., 2023); ARN separates surface from system mappings between narratives (Sourati et al., 2024); and AnaloBench evaluates abstract and long-context analogical reasoning (Ye et al., 2024). These benchmarks show that deep structural correspondence remains substantially harder for language models than lexical or thematic similarity. Closest to our setting, ParallelPARC elicits explicit relational correspondences before generating analogous paragraphs (Sultan et al., 2024); we include a ParallelPARC-inspired baseline. These approaches, however, do not require the target to become an autonomous fictional world in which mapped behavior follows from internally established causal rules.

**Structured story generation.** SAM uses explicit structure mapping to transfer information from a source story into a target story (Ontañón & Zhu, 2011); MEGATRON-CNTRL combines planning, knowledge retrieval, and conditional generation (Xu et al., 2020); and DOC uses hierarchical outlines for long-form coherence (Yang et al., 2023). These methods control *what happens* in a story; SPAN additionally constrains *why it happens*, because target events must remain aligned with causal dependencies inherited from an external mechanism. Metaphor generation such as MERMAID (Chakrabarty et al., 2021) also transfers across domains, but does not require preserving an extended system of entities, dependencies, and state changes. SPAN combines these threads by treating analogical generation as **causally grounded story-world construction**.

## 3 Method: Structure-Preserving Analogical Narratives

SPAN transforms a scientific explanation into a self-contained fictional narrative while preserving its salient functional, relational, and causal organization. Its premise is that analogical generation should preserve not only *what happens* in the source mechanism but also the conditions that make those events possible: if a source component behaves in some way because it has property $p$ under condition $c$, its fictional counterpart should possess an analogous affordance that makes the mapped behavior possible under a corresponding condition. SPAN uses a fixed pretrained model (no training) and five sequential, persisted transformations:

$$x \xrightarrow{f_1} z_1 \xrightarrow{f_2} G_s \xrightarrow{f_3} G_a \xrightarrow{f_4} W \xrightarrow{f_5} y, \tag{1}$$

where $x$ is the scientific explanation, $z_1$ a causally explicit rewrite, $G_s$ a causal and constraint-aware graph, $G_a$ the graph augmented with domain-neutral roles and causal affordances, $W$ a fictional world model, and $y$ the final narrative. Persisting each artifact removes the need for a single generation step to simultaneously interpret the science, discover an analogy, build a world, and write prose, and makes each decision inspectable and ablatable (Section 6).

### 3.1 Problem Formulation

Given $x \in \mathcal{X}$, we seek a narrative $y \in \mathcal{Y}$ satisfying

$$\mathrm{Struct}(x, y) \ \wedge \ \mathrm{Causal}(x, y) \ \wedge \ \mathrm{Story}(y), \tag{2}$$

i.e., preservation of salient functional and relational structure, preservation and target-world grounding of important causal dependencies, and functioning as a narrative rather than a renamed process description. The target world must also satisfy *causal closure*. Representing a story as world states $s_0, \ldots, s_T$, every major transition must be explained by a previously established property or rule $\lambda$ and a trigger or changed condition $a_t$:

$$s_t = \mathcal{T}(s_{t-1}, a_t; \lambda). \tag{3}$$

An important event should thus be explainable from the fictional world itself, not merely because its scientific counterpart happens. We summarize this constraint as

$$\boxed{\text{Story Event} \leftarrow \text{Prior State} + \text{Affordance} + \text{Trigger/Constraint}} \tag{4}$$

where $+$ denotes causal composition rather than arithmetic. Because a lexical transformation can preserve structure while being narratively poor, and fluent fiction can distort the mechanism, we evaluate these properties separately (Section 5).

### 3.2 Stage 1: Causally Explicit Scientific Rewriting

Scientific explanations often encode causal information implicitly: relations are distributed across sentences, compressed into equations, or described as sequences without saying why transitions occur. Stage 1 restructures the source strictly within the scientific domain, without generating any analogy. For each important transition it exposes the tuple (prior state, enabling property, trigger, response, consequence), recovering dependencies of the form $(P_A, C) \Rightarrow A \rightarrow B$ rather than only $A \rightarrow B$, where $P_A$ is a relevant property of $A$ and $C$ the enabling condition. For example, "hypertonic environment $\rightarrow$ plasmolysis" is expanded to external hypertonicity $\rightarrow$ net water loss $\rightarrow$ reduced internal volume $\rightarrow$ contraction of the flexible interior $\rightarrow$ separation from the rigid wall. Where present, thresholds, bottlenecks, equilibria, opposing forces, reversals, and failure or control conditions are also made explicit. Stage 1 is constrained not to introduce external knowledge: if the source does not explain a mechanism, SPAN preserves that gap rather than inventing one.

### 3.3 Stage 2: Causal and Constraint-Aware Graph Construction

Stage 2 converts $z_1$ into a graph $G_s = (V_s, E_s)$ whose nodes are entities, properties, conditions, states, and outcomes, and whose edges are directed relations $(v_i, r_{ij}, v_j)$, e.g., (NetWaterLoss, `reduces`, ProtoplastVolume). Wherever the source supports it, SPAN seeks causal completeness: if $v_i \rightarrow v_j$ depends on property $p_i$, condition $c_i$, and action $a_i$, the graph exposes $p_i \rightarrow c_i \rightarrow a_i \rightarrow v_j$. The graph can thus represent selectivity, rigidity, accumulation, pressure, thresholds, opposing forces, equilibria, and flows of resources or information. Its role is global: individually correct local mappings need not preserve a mechanism's topology, so Stage 2 retains branches, multi-step dependencies, and interacting constraints before any fictional realization.

### 3.4 Stage 3: Functional Roles and Causal Affordances

Stage 3 performs the key abstraction from scientific identity to target-realizable function. For each structurally important node $v_i$, SPAN assigns a pair $(r_i, p_i)$: a domain-neutral *functional role* $r_i$ (*what does this component do?*) and a *causal affordance* $p_i$ (*what property would make a fictional counterpart naturally capable of doing it?*). These are not equivalent. Mapping a selectively permeable boundary to a *gate* supplies a counterpart but not selectivity; specifying a gate whose openings admit lightly burdened travelers while blocking loaded carts provides a property from which selective passage follows. Similarly, "rigid outer boundary" is a role, whereas "thick stone blocks whose dimensions do not change when the contents inside move" is an affordance capable of sustaining it. Affordances preferentially use intuitive physical or social properties such as size, rigidity, elasticity, porosity, weight, capacity, crowding, friction, or scarcity. When no ordinary property suffices, a fictional one may be introduced, but it must obey a precise, stable rule, be established before its causal use, and never acquire new capabilities to satisfy later events. Stage 3 changes node descriptions while preserving graph topology: scientific vocabulary disappears, but the relational skeleton remains.

### 3.5 STAGE 4: CAUSAL STORY-WORLD SYNTHESIS

Individually plausible mappings need not form a coherent story. Stage 4 therefore searches for a *single* target world $W = (V_w, R_w, L_w, S_w, I_w)$, with fictional entities $V_w$, relations $R_w$, stable laws $L_w$, mutable state variables $S_w$, and invariants $I_w$, in which the grounded roles coexist. The premise (a kingdom, marketplace, harbor, caravan, etc.) must be sufficiently removed from the source domain. We favor *causal economy*, minimizing $|L_w|$ subject to sufficient coverage of the source mechanism, so that one law can support several relations; e.g., a fixed admission rule can explain selective passage, accumulation on one side of a barrier, and a later change of state. Separating state variables from invariants blocks a damaging failure: changing world rules whenever the science needs a different outcome. Pressure or population may change; gate width or an established admission rule may not. Where appropriate, a protagonist observes an abnormal state, infers a governing rule, and manipulates a condition to resolve the conflict, interacting with the mechanism rather than reciting it. The decisive check is: *why does event $e$ occur in this world?* If the only answer is "because the scientific event occurs," the mapping is revised.

### 3.6 STAGE 5: CAUSALLY AUDITED NARRATIVE REALIZATION

Stage 5 has two coupled responsibilities. For *realization*, SPAN identifies the world's deep dynamics (initiation, accumulation, resistance, bottleneck, reversal, release, stabilization) and compresses related dependencies into scenes; the graph is hidden infrastructure, not a sentence-level plan. The text prioritizes characters, goals, visible change, tension, and resolution, and excludes scientific terminology, explicit mappings, and "this represents X" explanations, so it remains intelligible without the source. For *verification*, every major transition is audited for an explanation $s_{t-1} + p + c \Rightarrow s_t$ from an established affordance $p$ and changed condition $c$. Changes in direction receive special attention: if movement starts, stops, or reverses, a corresponding condition must change, and a protagonist cannot simply command a reversal that the source attributes to a threshold or gradient. A counterfactual audit requires that removing a fictional property $p_w$ removes the event it enables ($\neg p_w \Rightarrow \neg e_w$), and that source control states (equilibrium, damage, absence) have fictional contrasts for the same functional reason. Finally, the system checks for source-domain leakage, unexplained special abilities, and appended mapping explanations.

## 4 EXPERIMENTS

**Data.** We curate scientific and technical explanations from secondary-school and early undergraduate material spanning biology, chemistry, physics, biotechnology, data structures and algorithms, medical science, materials science, and reinforcement learning, focusing on explanations with interacting entities, causal dependencies, constraints, thresholds, or multi-step mechanisms. The procedural collection contains 120 examples (Appendix A); the main controlled human evaluation uses the 47 sources for which all compared systems produced valid outputs.

**Systems.** We compare four progressively structured baselines: B1, a ParallelPARC-inspired baseline that builds explicit source–target correspondences and then realizes them as a story; B2, local-to-global generation that composes local analogical mappings into a narrative; B3, explicit functional/role mapping before generation; and B4, a single heavily constrained prompt containing all structural and narrative requirements. **SPAN** ($\text{FULL}_0$) is the five-stage pipeline of Section 3. **SPAN-W** (FULL) keeps the same stages but strengthens Stage 4 with whole-world mechanism search: it first seeks one target setting whose native rules jointly realize multiple source relations, preferring few stable rules, before assigning individual counterparts (Appendix K).

**Models and controls.** GPT-4o is the primary generator and gpt-oss-120B an open-weight comparison; a llama3.3:70b 70B model tests whether mediation or fine-tuning compensates for lower capability. All systems receive the same source, output budget, and sampling configuration ($T = 0.6$, top-$p = 0.9$). Generation and evaluation calls are separate, and candidate order is randomized in pairwise evaluation.Automatic evaluation was conducted using a heterogeneous panel of LLM-as-a-judge models comprising Claude Sonnet 5.

Table 1: **Main results on GPT-4o** ($n = 47$ sources). $\alpha_A$/$\alpha_N$: rubric composites (95% CIs in Table 6). Acc./Rep./Rej.: pooled human usability (%). Right: human science-aware analogy preference, SPAN-W / Tie / Comparator (%), for annotators H1 and H2.

| Method | $\alpha_A \uparrow$ | $\alpha_N \uparrow$ | Acc.↑ | Rep. | Rej.↓ | H1 | H2 |
|---|---|---|---|---|---|---|---|
| B1 | .658 | .563 | 0.0 | 14.9 | 85.1 | 72.3/27.7/0.0 | 78.7/21.3/0.0 |
| B2 | .708 | .499 | 0.0 | 31.9 | 68.1 | 17.0/83.0/0.0 | 14.9/83.0/2.1 |
| B3 | .697 | .564 | 0.0 | 48.9 | 51.1 | 12.8/87.2/0.0 | 29.8/68.1/2.1 |
| B4 | .606 | .315 | 0.0 | 22.3 | 77.7 | 19.1/80.9/0.0 | 31.9/66.0/2.1 |
| SPAN ($\text{FULL}_0$) | **.974** | .746 | 25.5 | 64.9 | 9.6 | 23.4/76.6/0.0 | 8.5/91.5/0.0 |
| SPAN-W (FULL) | .972 | **.826** | **27.7** | 64.9 | **7.4** | – | – |

**Evaluation.** Each story is scored on two axes (Appendix B). Analogical fidelity $\alpha_A$ averages entity–role fit, relation fidelity, structural fidelity, causal fidelity, and coverage (each 0–3, normalized) and divides by a penalty for entity-level surface similarity. Narrative quality $\alpha_N$ averages narrative arc, story causality, world coherence, and narrative integration, divided by a cognitive-load penalty (element load, terminology, translation burden, extraneous complexity). Composites are computed per story and then averaged, and are never collapsed into one number, since a fluent but causally faulty story and a faithful but dry retelling are different failures. Humans additionally provide science-aware pairwise analogy judgments, science-blind story judgments, and absolute ACCEPT/REPAIR/REJECT (ARR) labels; for Repair cases, annotators make the minimum revision and we measure normalized word-level edit distance. Ablations remove the causal rewrite (A1), graph (A2), affordance grounding (A3), and world synthesis (A4).

# 5 RESULTS

## 5.1 MAIN COMPARISON

Table 1 shows that both proposed variants substantially exceed all baselines. Original SPAN obtains $\alpha_A = 0.974$ and $\alpha_N = 0.746$, gains of $+0.266$ over the strongest baseline in analogical fidelity (B2) and $+0.182$ in narrative quality (B3). SPAN-W leaves analogical fidelity essentially unchanged ($0.974 \rightarrow 0.972$) while raising narrative quality to $0.826$ ($+0.080$): the world-level search does not buy narrative quality by weakening the analogy, but changes how an already preserved causal structure is realized as fiction. This addresses the central baseline failure pattern. Structure-oriented baselines preserve scientific steps but read like process diagrams rendered in prose, whereas more natural fiction yields "nice story, not science," in which a causal dependency is weakened or replaced. Other recurrent failures include surface relabeling, unexplained fictional capabilities, abrupt entity intrusion, and volitional shortcuts in place of mechanisms (Appendix J).

## 5.2 HUMAN EVALUATION

**Pairwise analogy.** SPAN-W is clearly preferred over B1. Against B2–B4, most judgments are ties, and preference for the baseline never exceeds $2.1\%$: the stronger baselines often retain a recognizable analogy, but this alone does not produce a usable analogical story. Against original SPAN, annotators choose SPAN-W in $23.4\%$ and $8.5\%$ of cases and otherwise report ties; neither ever prefers SPAN. Together with the near-identical $\alpha_A$, this indicates that SPAN-W improves narrative realization while keeping SPAN's analogy.

**Absolute usability.** ARR separates systems much more sharply. No baseline yields a single story accepted without modification, and their rejection rates range from $51.1\%$ to $85.1\%$. SPAN reaches $25.5\%$ Accept and $9.6\%$ Reject, and SPAN-W $27.7\%$ Accept and $7.4\%$ Reject, with $93.6\%$ inter-annotator agreement for both. The large Repair category mostly reflects local problems: typical repairs involve one to five local corrections and about 10–15% normalized word-level edit distance, after which $64.9\%$ of SPAN-W and $62.8\%$ of SPAN stories are revised-Accept (Table 8).

Table 2: **Stage-wise ablation** ($n = 47$ per ablation; FULL pools five realizations per source, $n = 235$). Human FULL pref.: % of judgments preferring FULL (H1/H2); the rest are almost entirely ties.

| Condition | $\alpha_A$ [95% CI] | $\alpha_N$ [95% CI] | FULL pref. | Dominant failure |
|---|---|---|---|---|
| FULL | .838 [.786,.880] | .983 [.957,.993] | – | some magical events need grounding |
| A1: −S1 | .000 [0,.076] | .000 [0,.076] | 89.0/90.0 | Generic causal relations |
| A2: −S2 | .745 [.605,.847] | 1.000 [.924,1] | 80.9/78.7 | Branches, chains collapse |
| A3: −Affordance | .723 [.582,.831] | .596 [.453,.724] | 59.6/63.8 | Behavior asserted |
| A4: −S4 | .957 [.858,.988] | .936 [.924,1] | 19.1/21.3 | Occasional slips |

### 5.3 Reliability of LLM-Based Evaluation

We tested whether holistic LLM pairwise judgments reproduce human decisions (Table 12). For analogy, humans agree with each other on $83.9\%$ of comparisons ($\kappa = 0.679$), but the LLM judge agrees with individual humans on only $37.6$–$38.3\%$ and with human consensus on $39.2\%$; we therefore do *not* use holistic LLM analogy preference as evidence. For storyness, LLM–consensus agreement reaches $81.9\%$, although skewed labels make $\kappa$ unstable. The decomposed $\alpha_A$/$\alpha_N$ rubric is therefore our primary automatic measure, and human ARR the final usability criterion. Calibrating against the independent human labels gives the regions Reject if $\alpha_A < 0.35$ or $\alpha_N < 0.65$, Accept if $\alpha_A \geq 0.80$ and $\alpha_N \geq 0.70$, and Repair otherwise; these are used only to interpret automatic scores.

### 5.4 Effect of Model Capacity

The gains depend on the generator. With gpt-oss-120B, SPAN-W's analogical score ($0.422$) falls below the strongest baseline ($0.695$, B3), but it obtains the best narrative score ($0.549$ vs. at most $0.511$), is the only condition producing accepted stories ($16.9\%$), and cuts rejection to $50.6\%$ from roughly $85$–$88\%$. Structural mediation thus improves narrative realization and usability for the smaller model without recovering the analogical fidelity of the stronger one.

## 6 Ablations and Analysis

### 6.1 Which Representations Matter?

Table 2 tests whether gains come from a longer prompting procedure or from specific representations; its scores are interpreted within this experiment rather than compared with Table 1. Removing the causal rewrite (**A1**) produces the clearest collapse: both axes fall to zero and humans prefer FULL in about $90\%$ of comparisons. Entities may still receive plausible counterparts, but mechanism-specific dependencies degrade into generic relations such as "$X$ set off $Y$"; Stage 1 makes explicit the causal preconditions that structural transfer requires. Removing the graph (**A2**) affects local mappings less but drops secondary branches and compresses multi-step chains (FULL preferred in $78.7$–$80.9\%$), indicating that the graph mainly contributes global relational coverage. Removing affordance grounding (**A3**) lowers $\alpha_N$ from $0.983$ to $0.596$ (FULL preferred in $59.6$–$63.8\%$); outputs contain structurally convenient but causally unsupported events, as objects acquire unexplained abilities, source vocabulary leaks in, or transitions occur simply because the analogy needs them. This supports our central claim that functional correspondence does not imply causally realizable correspondence. Finally, **A4** is a useful negative result: removing explicit world synthesis leaves high scores and mostly tied human judgments, so not every stage contributes equally. Given well-grounded mappings, the generator recovers much of the world composition itself, which motivated SPAN-W's more targeted world-level search for shared mechanisms.(Appendix M)

### 6.2 Where Does SPAN-W Still Fail?

Criterion-level scores for the 47 SPAN-W stories (Table 14) explain why residual errors are mostly Repair rather than Reject. Relation preservation, structural consistency, and causal coherence all reach $1.000$ (normalized), causal fidelity $0.957$, and world coherence $0.993$. Residual difficulty

lies in cognitive mediation: translation burden ($0.426$) and element load ($0.340$), while terminology burden is $0.000$. Even when the world is structurally faithful, readers may track too many correspondences or need to translate back to the concept. These are repairable presentation burdens rather than causal failures.

### 6.3 Recognition Does Not Imply Generation

A zero-shot 70B model achieves $76.8\%$ mean accuracy on four analogy-recognition benchmarks (AnaloBench, StoryAnalogy, ParallelPARC, ARN; Table 15), yet its generation is far weaker: $(\alpha_A, \alpha_N) = (0.13, 0.64)$ for few-shot generation, $(0.23, 0.35)$ with SPAN, $(0.35, 0.24)$ with supervised fine-tuning, and $(0.38, 0.41)$ with an InfoNCE variant. None reaches the Accept region. This *recognition–generation gap* shows that identifying analogical relations does not imply the ability to construct an extended world in which they coexist causally.

**Overall finding.** The results separate three effects. The SPAN decomposition substantially improves both analogical fidelity and narrative quality over the tested baselines; SPAN-W improves the *realization* of that structure ($\alpha_N$ $0.746 \rightarrow 0.826$ at unchanged $\alpha_A$), with no human preference for original SPAN; and both effects remain model-dependent. Explicit causal structure and affordance grounding prevent many characteristic failures, but do not make generator capability irrelevant.

## 7 Conclusion

We introduced SPAN, a five-stage framework for generating structure-preserving analogical narratives from scientific explanations. Its premise is that long-form analogy generation requires more than mapping scientific entities to fictional counterparts: the target world must preserve the source's functional and causal dependencies while working as an autonomous story, which SPAN achieves by mediating between causal structure, functional roles, causal affordances, world construction, and narrative realization. On GPT-4o, SPAN substantially improves both analogical fidelity and narrative quality over structure-oriented baselines; SPAN-W further improves narrative realization without sacrificing structure and is the only condition whose stories humans accept unmodified. Ablations identify causal explication and affordance grounding as the key components.

Structured mediation does not, however, remove dependence on the generator: gains on gpt-oss-120B and a 70B model do not reach the human usability region, and strong analogy recognition does not guarantee long-form analogy generation. Analogical story generation is thus a joint problem of structural representation, causal grounding, narrative realization, and model capability. We hope SPAN supports analogies whose worlds genuinely behave like the systems they explain.

### AI use statement

We used generative AI tools as automated judges (LLM-as-a-judge) to score model outputs, as described in Section 4. We did not use generative AI tools for research ideation, experimental design, code implementation, or data collection. We also used generative AI tools to assist with the literature search for related work and to polish text originally drafted by the authors. For evaluation, prompts and the scoring rubric are fully specified in our code repository; we validated the judge's scores against human annotations on a subset of 47 samples and manually inspected a sample of judgments to assess their reliability. Every paper suggested by AI tools was manually located, read, and verified by the authors before being cited. All AI-polished text was checked by the authors to confirm it preserves the intended meaning and does not introduce unsupported claims. We take responsibility for the final content of this work.

## References

Lucy Avraamidou and Jonathan Osborne. The role of narrative in communicating science. *International Journal of Science Education*, 31(12):1683–1707, 2009. doi: 10.1080/09500690802380695.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. Mermaid: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of NAACL-HLT*, 2021. doi: 10.18653/v1/2021.naacl-main.336.

Reinders Duit. On the role of analogies and metaphors in learning science. *Science Education*, 75 (6):649–672, 1991. doi: 10.1002/sce.3730750606.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170, 1983. doi: 10.1207/s15516709cog0702_3.

Shawn M. Glynn and Tomone Takahashi. Learning from analogy-enhanced science text. *Journal of Research in Science Teaching*, 35(10):1129–1149, 1998. doi: 10.1002/(SICI)1098-2736(199812)35:10⟨1129::AID-TEA5⟩3.0.CO;2-2.

Allan G. Harrison and David F. Treagust. Teaching with analogies: A case study in grade-10 optics. *Journal of Research in Science Teaching*, 30(10):1291–1307, 1993. doi: 10.1002/tea.3660301010.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of EMNLP*, pp. 11518–11537, 2023. doi: 10.18653/v1/2023.emnlp-main.706.

Stephen P. Norris, Sandra M. Guilbert, Martha L. Smith, Shahram Hakimelahi, and Linda M. Phillips. A theoretical framework for narrative explanation in science. *Science Education*, 89 (4):535–563, 2005. doi: 10.1002/sce.20063.

Santiago Ontañón and Jichen Zhu. The sam algorithm for analogy-based story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 7, pp. 67–72, 2011. doi: 10.1609/aiide.v7i1.12426.

Sara Soares, Matilde Gonçalves, Rita Jerónimo, and Régine Kolinsky. Narrating science: Can it benefit science learning, and how? a theoretical review. *Journal of Research in Science Teaching*, 60(9):2042–2075, 2023. doi: 10.1002/tea.21851.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. Arn: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086, 2024. doi: 10.1162/tacl_a_00688.

Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. Parallelparc: A scalable pipeline for generating natural-language analogies. In *Proceedings of NAACL-HLT*, pp. 5900–5924, 2024. doi: 10.18653/v1/2024.naacl-long.329.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of EMNLP*, pp. 2831–2845, 2020. doi: 10.18653/v1/2020.emnlp-main.226.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. Doc: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 3378–3465, 2023. doi: 10.18653/v1/2023.acl-long.190.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. Analobench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of EMNLP*, pp. 13060–13082, 2024. doi: 10.18653/v1/2024.emnlp-main.725.

# A Additional Experimental Details

## A.1 Source Collection

Our primary source collection consists of procedural scientific and technical explanations selected to contain non-trivial relational or causal structure. The recorded procedural pool contains 120 examples distributed across the domains shown in Table 3. Purely descriptive, mathematical derivations, taxonomic, or symbolic material is excluded from the primary generation experiment when it does not provide a meaningful process structure to transfer.

Table 3: **Recorded composition of the procedural source pool.**

| Domain | Number of examples |
|---|---|
| Data Structures and Algorithms | 31 |
| Medical science | 25 |
| Physics / Chemistry | 19 |
| Biotechnology | 17 |
| Materials Science | 14 |
| Reinforcement Learning | 14 |
| Total | 120 |

The main detailed human study uses 47 source explanations for which valid outputs are available across the compared systems. Unless stated otherwise, pairwise human results in this appendix therefore refer to these 47 source items.

## A.2 Compared Generation Methods

Table 4 summarizes the principal generation conditions.

SPAN-W should therefore be interpreted as a targeted extension of SPAN's world-construction mechanism rather than a different overall pipeline.

## A.3 Generation Controls

For controlled comparisons, systems receive the same source explanation, maximum output budget, and sampling configuration wherever technically applicable. The recorded generation settings are

$$T = 0.6, \qquad p_{\text{top}} = 0.9.$$

Candidate presentation order is randomized in pairwise evaluations. Generation and evaluation are executed as separate inference calls to reduce direct self-evaluation effects.

The main experiments use GPT-4o and gpt-oss-120B as generation models. The auxiliary capacity study uses a 70B Llama-family model (Llama 3.3 70B, run via Ollama as `llama3.3:70b`).

# B Evaluation Protocol

## B.1 Two-Axis Evaluation

Every generated story is represented by two independent composite scores,

$$(\alpha_A, \alpha_N),$$

where $\alpha_A$ measures analogical fidelity and $\alpha_N$ measures standalone narrative quality. Each constituent rubric dimension is scored on a 0–3 scale.

The analogical composite is

Table 4: **Generation methods used in the main comparison.**

| Method | Short description | Purpose |
|---|---|---|
| B1 | ParallelPARC-inspired relational generation | Constructs explicit base-target relational correspondences before realizing them as a story. Tests whether explicit relational alignment alone is sufficient. |
| B2 | Local-to-Global generation | Constructs local or sentence-level analogical mappings before attempting to integrate them into one global narrative. Tests whether locally plausible mappings compose naturally. |
| B3 | Role-Mapping / RMT-style generation | Maps source entities into functional target roles before story generation. Tests whether explicit correspondence planning without the complete causal mediation pipeline is sufficient. |
| B4 | Heavy single-prompt structural generation | Places scientific, structural, and narrative constraints into one generation prompt. Tests whether explicit instruction can substitute for persisted intermediate representations. |
| SPAN ($\text{FULL}_0$) | Original five-stage framework | Causal scientific rewrite $\rightarrow$ causal-mechanistic graph $\rightarrow$ functional role and affordance grounding $\rightarrow$ causal story-world synthesis $\rightarrow$ causally audited narrative. |
| SPAN-W (FULL) | SPAN with strengthened world-level synthesis | Retains the five stages of SPAN but strengthens Stage 4 through whole-world mechanism search, global coexistence constraints, shared-mechanism preference, causal economy, and world selection before node-level mapping. |

$$\alpha_A = \frac{\text{mean}(ERF, RF, SF, CF, CC)/3}{1 + \lambda_E ES/3}, \tag{5}$$

where $ERF$ is entity-role fit, $RF$ and $SF$ capture relational and structural correspondence, $CF$ measures causal fidelity, $CC$ measures coverage/consistency, and $ES$ is an entity-similarity penalty.

The narrative composite is

$$\alpha_N = \frac{\text{mean}(NA, SC, WC, NI)/3}{1 + \lambda_C (CL_1 + CL_2 + CL_3 + CL_4)/12}, \tag{6}$$

where $NA$ is narrative arc, $SC$ is story/event causality, $WC$ is world coherence, and $NI$ is narrative integration. $CL_1$–$CL_4$ correspond to element load, terminology burden, translation burden, and extraneous complexity.

All main results use equal positive-dimension weights and

$$\lambda_E = \lambda_C = 1.$$

The two composites are not combined into a single overall score. Composite scores are computed per story and then averaged; because the penalty terms are nonlinear, composites reconstructed from aggregate criterion means need not exactly equal the reported mean composites.

### B.2 Criterion Definitions

### B.3 Human Evaluation Tasks

Humans perform three independent tasks.

**Science-aware analogy comparison.** Annotators receive the scientific explanation and two generated stories and select which better preserves its relational and causal structure. Ties are allowed.

**Science-blind story comparison.** Annotators receive only the two stories and judge which functions better as an autonomous narrative, independently of scientific correctness.

Table 5: **Diagnostic dimensions used to construct the two evaluation axes.**

| Axis | Criterion | Operational question |
|---|---|---|
| Analogy | Entity-Role Fit (ERF) | Does the target counterpart perform the corresponding functional role? |
| | Entity Similarity (ES; penalty) | Is the mapping relying excessively on source-target surface similarity? |
| | Relation Fidelity (RF) | Are important first-order relations preserved? |
| | Structural Fidelity (SF) | Are higher-order dependencies and system structure preserved? |
| | Causal Fidelity (CF) | Do target-world causes produce corresponding consequences for analogous reasons? |
| | Coverage / Consistency (CC) | Are the important parts of the mechanism represented without contradiction? |
| Narrative | Narrative Arc (NA) | Does the text exhibit setup, development/change, and resolution? |
| | Story Causality (SC) | Do earlier events create or constrain later events? |
| | World Coherence (WC) | Are the fictional world's rules stable and understandable? |
| | Narrative Integration (NI) | Do characters, objects, stakes, and mechanism form one integrated story? |
| Cognitive penalty | Element Load ($CL_1$) | Must the reader track too many distinct symbolic elements? |
| | Terminology Burden ($CL_2$) | Does difficult or disguised technical vocabulary create avoidable burden? |
| | Translation Burden ($CL_3$) | Must the reader repeatedly translate story objects back into scientific concepts? |
| | Extraneous Complexity ($CL_4$) | Do unnecessary devices, subplots, metaphors, or decorations increase mental effort? |

**Absolute usability.** Each story independently receives one of three labels:

- ACCEPT: usable without substantive modification;
- REPAIR: the underlying analogy/story is salvageable through localized correction;
- REJECT: substantial reconstruction is required.

The detailed rubric scores do not determine the human label during annotation.

### B.4 HUMAN-CALIBRATED SCORE REGIONS

Using equal weights and $\lambda_E = \lambda_C = 1$, the empirical regions aligned to human usability judgments are

$$\text{REJECT} \quad \text{if} \quad \alpha_A < 0.35 \ \vee \ \alpha_N < 0.65, \tag{7}$$

$$\text{ACCEPT} \quad \text{if} \quad \alpha_A \geq 0.80 \ \wedge \ \alpha_N \geq 0.70, \tag{8}$$

with all remaining cases falling into the REPAIR region. These regions are used for interpretation of automatic scores and do not replace the independently collected human ARR judgments.

### B.5 REPAIR EFFORT

For Repair-labelled stories, experts make the minimum changes necessary to obtain a usable story. Repair burden is measured using normalized word-level Levenshtein distance,

$$d_{\text{repair}} = \frac{N_{\text{insert}} + N_{\text{delete}} + N_{\text{substitute}}}{N_{\text{words}}}. \tag{9}$$

Typical successful repairs involve one to five local additions or corrections, corresponding to approximately 10–15% of an approximately 1,000-word story.

# C Additional Main Results

## C.1 Complete Main Rubric Comparison

Table 6: **Complete GPT-4o rubric comparison.** Brackets indicate 95% confidence intervals.

| Method | $\alpha_A \uparrow$ | $\alpha_N \uparrow$ |
|---|---|---|
| B1 | .658 [.631,.683] | .563 [.556,.569] |
| B2 | .708 [.681,.731] | .499 [.492,.505] |
| B3 | .697 [.671,.720] | .564 [.561,.569] |
| B4 | .606 [.576,.636] | .315 [.311,.321] |
| SPAN | **.974 [.958,.989]** | .746 [.733,.758] |
| SPAN-W | .972 [.953,.989] | **.826 [.809,.839]** |

SPAN-W changes analogical fidelity only marginally relative to SPAN ($0.974 \rightarrow 0.972$) while increasing narrative quality by $0.080$ ($0.746 \rightarrow 0.826$).

## C.2 Complete Human Accept / Repair / Reject Counts

Pooling the two annotators gives

$$\text{SPAN-W} : 27.7\%/64.9\%/7.4\%$$

and

$$\text{SPAN} : 25.5\%/64.9\%/9.6\%$$

for Accept / Repair / Reject respectively.

## C.3 Usability After Minimal Repair

Pooling annotations gives a revised-Accept rate of

$$\frac{28+33}{94} = 64.9\%$$

for SPAN-W and

$$\frac{31+28}{94} = 62.8\%$$

for SPAN, with $27.7\%$ remaining Repair for both and $7.4\%$ and $9.6\%$ Reject, respectively. Thus, a large proportion of the initial Repair cases are recoverable with localized rather than structural revision.

# D Complete Pairwise Evaluation

## D.1 Science-Blind Standalone Story Quality

Table 9 gives all pairwise science-blind narrative comparisons. "Pooled decisive" excludes *Both* judgments from the winner denominator, while "Both" is calculated over all valid judgments.

These comparisons suggest an approximate direct-pairwise narrative ordering

Table 7: **Independent human Accept/Repair/Reject judgments.** Each cell reports Accept / Repair / Reject counts.

| Story set | Annotator 1 | Annotator 2 | Raw human agreement |
|---|---|---|---|
| SPAN-W (FULL) | 12 / 32 / 3 | 14 / 29 / 4 | 93.6% |
| SPAN ($\text{FULL}_0$) | 11 / 30 / 6 | 13 / 31 / 3 | 93.6% |
| B1 | 0 / 6 / 41 | 0 / 8 / 39 | 95.7% |
| B2 | 0 / 17 / 30 | 0 / 13 / 34 | 91.5% |
| B3 | 0 / 22 / 25 | 0 / 24 / 23 | 95.7% |
| B4 | 0 / 11 / 36 | 0 / 10 / 37 | 97.9% |
| A1 | 0 / 9 / 38 | 0 / 8 / 39 | 97.9% |
| A2 | 0 / 3 / 44 | 0 / 5 / 42 | 95.7% |
| A3 | 0 / 6 / 41 | 0 / 10 / 37 | 91.5% |
| A4 | 10 / 30 / 7 | 10 / 28 / 9 | 95.7% |

Table 8: **Human labels after minimal expert repair.** Counts are revised-Accept / remaining Repair / Reject.

| Method | Annotator 1 | Annotator 2 | Agreement |
|---|---|---|---|
| SPAN-W | 28 / 16 / 3 | 33 / 10 / 4 | 87.2% |
| SPAN | 31 / 10 / 6 | 28 / 16 / 3 | 87.2% |

$$\text{SPAN-W} > B3 > B2 > B4 > B1,$$

but we do not interpret this as a guaranteed transitive global ranking. In particular, SPAN-W's advantage over B3 is narrow and accompanied by a substantial tie rate.

### D.2 Holistic Science-Aware Analogy Judgments

For completeness, Table 10 reports the corresponding holistic LLM analogy comparisons. As shown in Appendix F, these judgments align poorly with humans and are therefore treated as diagnostic rather than primary evidence.

## E Human Pairwise Analogy Results

Table 11 reports all human analogy comparisons between SPAN-W and the baselines, ablations, and original SPAN.

The SPAN-W versus SPAN comparison is dominated by ties and contains no human preference for original SPAN. This supports the interpretation that SPAN-W's largest effect is improved narrative realization rather than replacement of the underlying analogy.

## F LLM Judge Validation

Human-human analogy agreement is high relative to LLM-human analogy agreement. The poor agreement of the holistic analogy judge motivates the decomposed criterion-level evaluation used for $\alpha_A$. The storyness judge aligns more closely with human judgments, although label imbalance makes conventional chance-corrected agreement unstable in some comparisons.

## G Full Ablation Results

### G.1 Automatic Ablation Scores

$$A1: \quad \Delta\alpha_A = -.838,\ p < .001, \qquad \Delta\alpha_N = -.983,\ p < .001,$$

Table 9: **Complete science-blind storyness comparison.**

| Pair | DeepSeek-R1-70B | gpt-oss-120B | Pooled decisive | Both, pooled |
|---|---|---|---|---|
| B1 vs B2 | B2 96.8% | B2 87.5% | B2 92.1% | 1.6% |
| B1 vs B3 | B3 100.0% | B3 97.5% | B3 98.8% | 0.0% |
| B1 vs B4 | B4 86.5% | B4 80.0% | B4 83.1% | 3.8% |
| B1 vs SPAN-W | SPAN-W 100.0% | SPAN-W 97.5% | SPAN-W 98.7% | 3.8% |
| B2 vs B3 | B3 84.6% | B3 80.6% | B3 81.8% | 37.1% |
| B2 vs B4 | B2 87.5% | B2 74.2% | B2 78.7% | 32.9% |
| B2 vs SPAN-W | SPAN-W 70.0% | SPAN-W 68.8% | SPAN-W 69.0% | 40.0% |
| B3 vs B4 | B3 96.4% | B3 83.7% | B3 88.7% | 22.8% |
| B3 vs SPAN-W | SPAN-W 61.1% | SPAN-W 54.8% | SPAN-W 56.7% | 34.8% |
| B4 vs SPAN-W | SPAN-W 96.7% | SPAN-W 95.3% | SPAN-W 95.9% | 20.7% |

Table 10: **Complete holistic science-aware analogy comparison.** These results are diagnostic because the holistic analogy judge exhibits low human agreement.

| Pair | DeepSeek-R1-70B | gpt-oss-120B | Pooled decisive | Both, pooled |
|---|---|---|---|---|
| B1 vs B2 | B2 100.0% | B2 89.5% | B2 94.7% | 5.0% |
| B1 vs B3 | B3 100.0% | B3 84.0% | B3 91.3% | 8.0% |
| B1 vs B4 | B4 100.0% | B4 100.0% | B4 100.0% | 8.0% |
| B1 vs SPAN-W | SPAN-W 87.0% | SPAN-W 72.0% | SPAN-W 79.2% | 2.0% |
| B2 vs B3 | B2 60.0% | B3 55.6% | 50.0% each | 34.9% |
| B2 vs B4 | B2 62.5% | B4 64.3% | B4 54.5% | 50.0% |
| B2 vs SPAN-W | B2 75.0% | B2 75.0% | B2 75.0% | 16.3% |
| B3 vs B4 | B4 66.7% | B4 72.7% | B4 71.4% | 53.3% |
| B3 vs SPAN-W | B3 57.1% | B3 67.9% | B3 64.3% | 25.0% |
| B4 vs SPAN-W | B4 71.4% | B4 86.2% | B4 81.4% | 25.9% |

$$A2: \quad \Delta\alpha_A = -.094,\ p = .24, \qquad \Delta\alpha_N = +.017,\ p = .12,$$

and

$$A3: \quad \Delta\alpha_A = -.115,\ p = .17, \qquad \Delta\alpha_N = -.387,\ p < .001.$$

### G.2 Human Ablation Judgments

Human pairwise results show a complementary pattern. SPAN-W is preferred over A1 by 89.0–90.0% of annotators' judgments, over A2 by 78.7–80.9%, and over A3 by 59.6–63.8%. The A4 comparison is dominated by ties: 78.7–80.9% of judgments rate the two outputs equivalently.

The independently collected ARR judgments are reproduced in Table 7. These human labels should not be conflated with any automatically thresholded Accept/Repair/Reject-like quantities reported in intermediate ablation analyses.

### G.3 Interpretation of Individual Ablations

**A1: causal explication.** Without Stage 1, later stages receive scientific prose in which important dependencies may remain implicit. The observed degeneration is not merely lower fluency: causal links themselves collapse into generic succession.

**A2: global relational topology.** Removing the explicit graph often preserves conspicuous local correspondences while dropping secondary branches or relations that span multiple source sentences. This distinguishes locally plausible analogy construction from global structural preservation.

**A3: causal affordance grounding.** This ablation retains what each counterpart is supposed to do while removing the property that makes the behavior possible. It therefore tests

Table 11: **Complete human pairwise analogy preferences.** Values are SPAN-W / Tie / Comparator (%).

| SPAN-W vs. | Annotator 1 | Annotator 2 |
|---|---|---|
| A1 | 89.0 / 11.0 / 0.0 | 90.0 / 10.0 / 0.0 |
| A2 | 80.9 / 19.1 / 0.0 | 78.7 / 21.3 / 0.0 |
| A3 | 59.6 / 40.4 / 0.0 | 63.8 / 36.2 / 0.0 |
| A4 | 19.1 / 80.9 / 0.0 | 21.3 / 78.7 / 0.0 |
| B1 | 72.3 / 27.7 / 0.0 | 78.7 / 21.3 / 0.0 |
| B2 | 17.0 / 83.0 / 0.0 | 14.9 / 83.0 / 2.1 |
| B3 | 12.8 / 87.2 / 0.0 | 29.8 / 68.1 / 2.1 |
| B4 | 19.1 / 80.9 / 0.0 | 31.9 / 66.0 / 2.1 |
| SPAN | 23.4 / 76.6 / 0.0 | 8.5 / 91.5 / 0.0 |

Table 12: **Complete human and LLM pairwise agreement analysis.**

| Axis | Comparison | Raw agreement | Cohen's $\kappa$ | Gwet's AC1 |
|---|---|---|---|---|
| Analogy | Annotator 1 vs Annotator 2 | 83.9% | 0.679 | 0.786 |
| Analogy | Annotator 1 vs LLM | 38.3% | 0.096 | 0.098 |
| Analogy | Annotator 2 vs LLM | 37.6% | 0.085 | 0.086 |
| Analogy | All three | 32.9% all agree | Fleiss $\kappa = 0.221$ | – |
| Analogy | LLM vs human consensus ($n = 355$) | 39.2% | – | – |
| Storyness | Annotator 1 vs Annotator 2 | 88.7% | 0.000[†] | 0.873 |
| Storyness | Annotator 1 vs LLM | 78.7% | 0.000[†] | 0.764 |
| Storyness | Annotator 2 vs LLM | 75.2% | 0.141 | 0.711 |
| Storyness | All three | 72.6% all agree | Fleiss $\kappa = 0.037$ | – |
| Storyness | LLM vs human consensus ($n = 375$) | 81.9% | – | – |

[†]Annotator 1 gave the same label on all relevant Axis-2 items, so Cohen's $\kappa$ is not informative for those comparisons.

$$\text{functional role} \neq \text{causal affordance}.$$

Typical failures include arbitrary magical capability, target-world behavior introduced only when needed, and events that are structurally aligned but not causally entailed by the fictional world.

**A4: explicit world synthesis.** A4 is a useful negative result. Much of the composition can be reconstructed from sufficiently strong role and affordance representations. This motivates the more targeted contribution of SPAN-W: world synthesis is strengthened not by adding more independent mappings, but by explicitly searching for shared world-level mechanisms and causal economy.

## H Criterion-Level Diagnostic Analysis

Table 14 reports criterion-level means for the 47 SPAN-W stories. Raw values use the original 0–3 rubric; normalized values divide each dimension by three.

The mean positive analogical component is approximately

$$P_A = 0.977,$$

while the entity-similarity penalty is very small. Computing the nonlinear composite per story and then averaging gives

$$\alpha_A = 0.972.$$

For narrative quality, the mean positive component is

Table 13: **Detailed ablation results.** Each ablation contains $n = 47$ items; the FULL reference aggregates five samples per source ($n = 235$).

| Condition | $\alpha_A$ [95% CI] | $\alpha_N$ [95% CI] | Question tested | Observed failure signature |
|---|---|---|---|---|
| FULL | .838 [.786,.880] | .983 [.957,.993] | Complete system | Occasional omission of secondary sub-mechanisms. |
| A1: $-S1$ | .000 [0,.076] | .000 [0,.076] | Is explicit causal rewriting necessary? | Objects may remain grounded, but mechanism-specific relations collapse into generic event chains. |
| A2: $-S2$ | .745 [.605,.847] | 1.000 [.924,1] | Is global relational structure necessary? | Core mechanism survives, but secondary branches are dropped and multi-stage chains are compressed. |
| A3: −Affordance | .723 [.582,.831] | .596 [.453,.724] | Is functional correspondence alone sufficient? | Causal steps are asserted rather than entailed by target-world properties; source vocabulary and narrative fiat may appear. |
| A4: $-S4$ | .957 [.858,.988] | .930 [.924,1] | Is an explicit world-planning stage necessary? | Occasional composition slips; otherwise many outputs remain close to FULL. |

Table 14: **Criterion-level SPAN-W diagnostic scores ($n = 47$).**

| Criterion | Raw mean | Normalized |
|---|---|---|
| Entity-role fit | 2.894 | .965 |
| Entity similarity penalty | 0.021 | .007 |
| Relation preservation | 3.000 | 1.000 |
| Structural consistency | 3.000 | 1.000 |
| Causal fidelity | 2.872 | .957 |
| Mapping completeness | 2.894 | .965 |
| Narrative arc | 2.872 | .957 |
| Causal coherence | 3.000 | 1.000 |
| World coherence | 2.979 | .993 |
| Narrative integration | 2.957 | .986 |
| Element-load penalty | 1.021 | .340 |
| Terminology-burden penalty | 0.000 | .000 |
| Translation-burden penalty | 1.277 | .426 |
| Extraneous-complexity penalty | 0.043 | .014 |

$$P_N = 0.984.$$

The mean cognitive-load term is

$$CL = \frac{1.021 + 0 + 1.277 + 0.043}{12} \approx 0.195. \tag{10}$$

Applying the penalty per story and then averaging gives

$$\alpha_N = 0.826.$$

This decomposition shows that residual SPAN-W difficulty is concentrated more heavily in cognitive mediation than in structural preservation. In particular, translation burden is the largest remaining penalty, followed by element load.

# I SMALLER-MODEL ANALOGY RECOGNITION AND GENERATION

## I.1 ANALOGY-RECOGNITION BENCHMARKS

Table 15: **Zero-shot analogy-recognition performance of the 70B model.**

| Benchmark | Accuracy | Correct / Total |
|---|---|---|
| AnaloBench | 62.1% | 211 / 340 |
| StoryAnalogy | 81.9% | 295 / 360 |
| ParallelPARC | 86.5% | 268 / 310 |
| ARN | 76.7% | 840 / 1095 |
| Mean | 76.8% | – |

## I.2 70B GENERATION AND ADAPTATION

Table 16: **70B generation results under different adaptation regimes.**

| Condition | $\alpha_A \uparrow$ | $\alpha_N \uparrow$ |
|---|---|---|
| Few-shot direct | .13 | .64 |
| 70B + SPAN | .23 | .35 |
| Supervised fine-tuning | .35 | .24 |
| InfoNCE variant | **.38** | .41 |

Despite a mean recognition accuracy of 76.8%, none of these generation conditions reaches the empirical human Accept region

$$\alpha_A \geq 0.80, \qquad \alpha_N \geq 0.70.$$

These results therefore support a distinction between analogy recognition and extended analogy generation. A model can identify relational correspondences without being able to construct an autonomous narrative world in which those relations remain simultaneously causal, coherent, and cognitively manageable.

# J QUALITATIVE FAILURE TAXONOMY

The quantitative metrics are complemented by a qualitative failure taxonomy derived from recurrent baseline, ablation, and repair cases.

These categories capture the recurring tension motivating SPAN. Increasing structural control alone can produce mechanistic analogical narration, whereas increasing narrative freedom alone can produce fluent but causally unfaithful fiction. SPAN targets the region in which fictional events are both structurally corresponding and independently entailed by the target world's rules.

# K PROMPT ARCHITECTURE AND REPRODUCIBILITY

## K.1 SPAN

The original SPAN pipeline contains five persisted transformations:

$$z_1 = f_1(x), \quad \text{causally explicit scientific rewrite,} \tag{11}$$
$$G_s = f_2(z_1), \quad \text{causal-mechanistic knowledge graph,} \tag{12}$$
$$G_a = f_3(G_s), \quad \text{functional-role and affordance grounding,} \tag{13}$$
$$W = f_4(G_a), \quad \text{causal story-world synthesis,} \tag{14}$$
$$y = f_5(W), \quad \text{causally audited narrative realization.} \tag{15}$$

Table 17: **Qualitative failure categories used for diagnostic analysis.**

| Failure category | Description |
|---|---|
| Nice story, not science | The story is fluent or engaging but an important relation, constraint, or causal mechanism from the source has been lost or altered. |
| Mechanistic analogical narration | Source structure is represented, but the output reads like a process diagram or procedural explanation rendered with fictional nouns. |
| Surface rephrasing / label substitution | Scientific entities are replaced with story-like labels without genuine functional or causal transfer. |
| Extended fantasy metaphor | Required behaviors occur through unexplained or overly convenient fictional powers rather than target-world affordances. |
| Abrupt entity intrusion | An object, character, or device appears only when the corresponding source relation becomes necessary. |
| Volitional shortcut | A character's decision substitutes for an outcome that should follow from a physical, structural, or systemic constraint. |
| Constraint violation | The target-world behavior contradicts an important source-world restriction, direction, threshold, or failure condition. |
| Premise collapse | A fictional property or world rule changes during the story to enable a later event. |
| Reader dependency | The target mechanism cannot be understood or reasoned about without already knowing the source science. |
| Symbolic overload | Too many one-to-one mappings create substantial element or translation burden, making the analogy harder to reason through than intended. |

Equivalently,

$$F_{\mathrm{SPAN}} = f_5 \circ f_4 \circ f_3 \circ f_2 \circ f_1. \tag{16}$$

The governing realization principle is

$$\text{Prior State} + \text{Relevant Property/Affordance} + \text{Trigger/Condition} \rightarrow \text{Story Event} \rightarrow \text{Consequence}. \tag{17}$$

No major fictional event is intended to occur merely because an analogous scientific event needs to occur.

### K.2 SPAN-W

SPAN-W retains the same five transformations but modifies the search performed inside Stage 4. Before assigning individual graph nodes to target entities, the model first abstracts the complete causal shape into recurring behavioral demands and searches for a single familiar situation in which those demands coexist.

The search order is

$$\begin{aligned} \text{complete causal shape} &\rightarrow \text{recurring behavioral demands} \\ &\rightarrow \text{candidate whole-world mechanisms} \\ &\rightarrow \text{shared target world} \\ &\rightarrow \text{stable world laws} \\ &\rightarrow \text{individual node mappings}. \end{aligned} \tag{18}$$

Candidate worlds are rejected when major mechanisms require unrelated devices, when later events need newly invented rules, or when the reader must track too many independent symbolic substitutions.

Among viable worlds, SPAN-W prefers causal economy:

$$W^* = \arg\min_W [\text{independent special rules required to realize the source graph}], \tag{19}$$

subject to preservation of the important functional and causal relations.

This world-level intervention is the principal difference between SPAN and SPAN-W.

### K.3 Stage-Wise Ablation Definitions

The ablation conditions used in the reported study are:

$$A1 : x \to f_2 \to f_3 \to f_4 \to f_5, \quad \text{remove causal rewrite;} \tag{20}$$
$$A2 : f_1 \to f_3 \to f_4 \to f_5, \quad \text{remove explicit causal graph;} \tag{21}$$
$$A3 : f_1 \to f_2 \to \text{role-only } f_3 \to f_4 \to f_5, \quad \text{remove causal affordance;} \tag{22}$$
$$A4 : f_1 \to f_2 \to f_3 \to f_5, \quad \text{remove explicit world synthesis.} \tag{23}$$

A3 is particularly important because it isolates the distinction

what the counterpart must do vs. what property makes it capable of doing so.

### K.4 Pairwise Evaluation Formatting

Candidate order is randomized before presentation. For science-aware analogy evaluation, judges are instructed to prioritize functional correspondence, relation preservation, causal direction, constraints, and mechanism coverage, while ignoring prose beauty.

For science-blind story evaluation, the source explanation is hidden. Judges consider narrative progression, causal connection among events, character action, tension/development, coherent world rules, and resolution while ignoring scientific correctness.

The complete prompt templates used for generation and judging are released with the supplementary material.

## L Reporting and Statistical Conventions

We use the following conventions throughout:

- Composite rubric values are computed at the individual-story level before aggregation.
- Reported intervals are 95% confidence intervals.
- Pairwise decisive win rates exclude *Both* judgments from the winner denominator.
- Tie rates use all valid judgments as the denominator.
- Human agreement is reported independently from model-human agreement.
- Cohen's $\kappa$ is supplemented with Gwet's AC1 when prevalence or degenerate marginal distributions make $\kappa$ unstable.
- Accept/Repair/Reject judgments are treated as an ordered human usability outcome but are not used to construct $\alpha_A$ or $\alpha_N$.

For a pair of systems $i$ and $j$, the decisive win rate of $i$ is

$$WR(i,j) = \frac{N(i \text{ preferred})}{N(i \text{ preferred}) + N(j \text{ preferred})}. \tag{24}$$

The pooled rate across evaluators is computed from pooled counts rather than by taking an unweighted average of evaluator-specific percentages:

$$WR_{\text{pooled}}(i,j) = \frac{\sum_e N_e(i \text{ preferred})}{\sum_e [N_e(i \text{ preferred}) + N_e(j \text{ preferred})]}. \tag{25}$$

This convention is used for the complete pairwise tables in Appendix D.

## M NARRATIVE TRADITIONS AS STORY-WORLD RESOURCES

Stage 4 of SPAN searches for a target world whose stable rules can realize the source mechanism (Section 3.5). Established narrative traditions provide ready-made premises in which particular kinds of causal machinery are already familiar to readers, such as cooperating specialists, repeated tests, boundary crossings, nested containers, or tricksters who exploit rules rather than overpower them. Tables 18 and 19 catalogue traditions from many cultures, the narrative machinery each one handles especially well, and the scientific ideas that machinery can naturally personify. The mappings are heuristic starting points, not fixed correspondences. A premise remains subject to the same affordance grounding and causal audit as any other candidate world. Several entries (e.g., Dreaming stories, Yoruba and Mesoamerican narratives, and creation myths) come from living religious or cultural traditions and should be used respectfully, as structural inspiration rather than as material to be rewritten or trivialized.

Table 18: **Narrative traditions and the scientific ideas their narrative machinery can personify (part 1 of 2).**

| Story tradition / character | Region | Narrative machinery it is especially good at | Scientific ideas it can personify |
|---|---|---|---|
| **Panchatantra** | India | Specialized characters cooperating; chains of consequences | Pathways, networks, ecosystems, electron transport, symbiosis |
| **Hitopadesha** | India | Strategy, alliances, information, consequences | Signaling, regulation, game theory, feedback |
| **Jataka Tales** | India / Buddhist Asia | Sacrifice, adaptation, repeated lives, cause and effect | Adaptation, trade-offs, homeostasis, evolutionary fitness |
| **Tenali Rama** | South India | Clever exploitation of existing conditions | Enzymes, catalysis, gradients, optimization |
| **Akbar–Birbal** | India | Puzzles resolved by inference | Scientific method, hypothesis testing, diagnostics |
| **Vikram–Betal** | India | Repeated departure–return cycle | Cyclic pathways, feedback loops, oscillations |
| **Singhasan Battisi** | India | Repeated tests before advancement | Checkpoints, quality control, sequential validation |
| **Kathasaritsagara** | India | Stories nested inside stories | Hierarchical systems, scales within scales |
| **Aesop's Fables** | Greek / Mediterranean | Single-variable moral experiments | Simple mechanisms, competition, thresholds |
| **Anansi stories** | West Africa / Caribbean | Network-building, cunning, information | Webs and networks, information flow, distributed systems |
| **Br'er Rabbit** | African-American folklore | Small agent exploiting constraints | Optimization, niche exploitation, evolutionary strategy |
| **Tortoise tales** | Africa and many traditions | Slow strategy defeating raw speed or power | Efficiency vs. power, rate limiting, long-term optimization |
| **Grimm fairy tales** | Central Europe | Rules, transformations, thresholds | Phase changes, developmental transitions, trigger conditions |
| **Andersen tales** | Denmark | Transformation and identity | Developmental biology, states of matter, emergence |
| **Russian folktales / Baba Yaga** | Slavic world | Boundary crossing, tests, strange rules | Membrane transport, checkpoints, activation barriers |
| **Koschei the Deathless** | Slavic folklore | One thing hidden inside another | Hierarchical containment, cellular organization |
| **Chinese animal fables** | China | Strategy and ecological relationships | Predator–prey systems, competition, adaptation |
| **Journey to the West** | China | Team with radically different abilities | Organ systems, molecular complexes, division of labor |
| **Sun Wukong** | China | Transformations, copies, extreme mobility | Amplification, replication, molecular switching |
| **Zhuangzi stories** | China | Perspective shifts, paradoxes | Relativity of scale, observer effects, classification |
| **Japanese folktales** | Japan | Transformation, reciprocity, hidden identities | Metamorphosis, reversible vs. irreversible change |
| **Momotaro** | Japan | Assembling specialized companions | Molecular complexes, immune responses |
| **Zen koans** | East Asia | Breaking misleading intuitions | Quantum ideas, infinity, counterintuitive reasoning |
| **Korean trickster tales** | Korea | Clever constraint-solving | Optimization, adaptive behavior |
| **Norse myths** | Scandinavia | Cycles, world destruction and rebirth, interlinked realms | Cycles, energy transformations, geological cycles |

Table 19: **Narrative traditions and the scientific ideas their narrative machinery can personify (part 2 of 2).**

| Story tradition / character | Region | Narrative machinery it is especially good at | Scientific ideas it can personify |
|---|---|---|---|
| **Ragnarök** | Norse | System collapse followed by renewal | Ecological succession, extinction and recovery |
| **Yggdrasil** | Norse | Many realms connected through one structure | Networks, phylogenetic trees, transport architecture |
| **Greek myths** | Greece | Personified forces and transformations | Forces, energy, elemental interactions |
| **Sisyphus** | Greece | Repeated uphill work | Cycles requiring continual energy input |
| **Prometheus** | Greece | Acquisition of a powerful resource | Energy capture and use, technological transitions |
| **Theseus and the Labyrinth** | Greece | Navigating branching paths | Algorithms, neural pathways, search problems |
| **Ariadne's thread** | Greece | Persistent trace through complexity | Molecular markers, lineage tracing, memory |
| **Daedalus** | Greece | Engineering under constraints | Design, biomechanics, engineering |
| **Roman myths and fables** | Rome | Civic systems and roles | Division of labor, system organization |
| **Celtic tales** | Ireland / Britain | Transformations, thresholds, seasonal cycles | Developmental change, biological cycles |
| **Native American trickster tales** | Americas | Coyote or Raven reshaping situations | Adaptation, innovation, ecological interactions |
| **Raven stories** | Pacific Northwest | Bringing or redistributing light and resources | Energy and resource distribution |
| **Maya stories / Popol Vuh** | Mesoamerica | Repeated attempts at creation | Trial and error, selection, iterative systems |
| **Aztec mythic cycles** | Mesoamerica | Successive worlds or eras | Geological and ecological transitions |
| **Polynesian Maui stories** | Pacific | Manipulating natural constraints | Mechanics, environmental forces |
| **Aboriginal Australian Dreaming stories** | Australia | Landscape, relationships, routes | Ecological connectivity, spatial systems |
| **Yoruba tales** | West Africa | Interacting divine personalities and forces | Competing regulatory pathways |
| **Eshu stories** | Yoruba | Ambiguity, communication, dual interpretations | Signaling errors, information theory |
| **The blind men and the elephant** | South Asia / widely transmitted | Partial observations of one system | Sampling, incomplete data, model-building |
| **Fables of La Fontaine** | France | Simplified interaction experiments | Behavioral ecology, competition |
| **Tall tales** | Americas / Europe | Exaggerated variables | Scale effects, thought experiments |
| **Detective stories** | Global modern tradition | Evidence $\rightarrow$ hypothesis $\rightarrow$ elimination | Scientific method, diagnosis, forensic science |
| **Quest stories** | Universal | Ordered checkpoints | Metabolic pathways, synthesis pathways |
| **Trickster stories** | Nearly universal | Exploiting rules rather than overpowering them | Enzymes, evolution, optimization |
| **Creation myths** | Nearly universal | Emergence from simpler beginnings | Cosmology and emergence (use with caution) |