STRUCTURAL ABSTRACTION AS AN INDUCTIVE BIAS FOR NON-STATIONARY LANGUAGE MODEL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

A foundational principle in cognitive science holds that intelligent agents do not learn by storing experiences as isolated instances, but by forming abstract schemas that capture relational structure shared across situations. Even though this claim is well supported by behavioral and neuroimaging studies, its role as a computational training signal in language models remains underexplored. We target this gap in the setting of non-stationary language model training, asking whether biasing learning toward structural abstraction reduces catastrophic interference and improves relational generalization as predicted by human results. To study this question, we introduce Abstraction-Augmented Training (AAT), a lightweight loss-level modification that jointly optimizes over concrete instances and their structural abstractions, and two benchmarks, the Relational Cycle Benchmark (RCB) and the Narrative Abstraction Benchmark (NAB). These resources operationalize core cognitive constructs: entity masking as a computational analog of relational alignment, and proverbs as vehicles for implicit abstract meaning that must be inferred across surface-dissimilar situations. Our empirical results demonstrate that AAT improves generalization and retention in a pattern that aligns with cognitive predictions for schema-based learning. Beyond the practical implications for continual learning, these results offer preliminary computational evidence that structural abstraction can serve as a signal for stable learning in non-stationary environments.

1 INTRODUCTION

A central question in cognitive science is how intelligent agents acquire knowledge that transfers across contexts rather than merely memorizing specific experiences. Behavioral and theoretical work suggests that humans form abstract *schemas*: representations that capture relational structure shared across situations and support transfer to unseen but structurally similar cases (Schank & Abelson, 2013; Gick & Holyoak, 1983; Gentner, 1983). Experiences emphasizing relational structure are more likely to be retained and reused across contexts (Clement et al., 1994), while cognitive neuroscience suggests that learning organizes related events around shared components, supporting abstraction and stable generalization (Masís-Obando et al., 2022).

These principles become especially consequential in non-stationary settings, where knowledge must be acquired continuously from evolving data. Because repeatedly retraining language models from scratch is computationally prohibitive, we focus on *online continual learning* (OCL), where data arrives as a stream, task boundaries are absent, and each example is observed only once. Models must therefore acquire new facts while retaining prior knowledge and inducing structure that transfers to unobserved cases. Two central challenges are *catastrophic forgetting*, where previously learned knowledge is overwritten (Goodfellow et al., 2015; Kirkpatrick et al., 2017; Nguyen et al., 2019; Mirzadeh et al., 2022; Tadros et al., 2022; Luo et al., 2025), and *plasticity loss*, where the ability to incorporate new information degrades over time (Lyle et al., 2023; Dohare et al., 2024; Farias & Jozefiak, 2025; Lewandowski et al., 2025; Ahn et al., 2025).

Existing methods address these challenges through regularization (Kirkpatrick et al., 2017; Chaudhry et al., 2018) or experience replay (ER), which revisits stored examples from earlier training to stabilize learning (Rolnick et al., 2019). While ER is often effective (Van de Ven et al., 2020; Buzzega et al., 2020; Wang et al., 2025), it achieves stability by re-exposing the model to past instances requiring a memory buffer, which becomes increasingly costly in strictly online settings where storage is limited

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.

or infeasible. Importantly, such instance-level replay does not address the deeper problem that models trained primarily on surface-level instance signals may successfully recall training examples while losing the shared relational structures necessary for inductive reasoning.

Motivated by the cognitive principle that abstract, schema-consistent representations are more robustly retained and generalized (Hofstadter et al., 2001; Gentner, 2003), we introduce *Abstraction-Augmented Training* (AAT), a loss-level inductive bias. AAT jointly optimizes over a concrete instance and its corresponding structural abstraction at first encounter, biasing the model toward representations invariant to instance-specific surface variation. A small number of local replays then reinforce the concrete instance, consolidating factual details without allowing entity-specific gradients to dominate. AAT requires no storage of past data.

To evaluate abstraction at different levels, we introduce two benchmarks separating *episodic* retention from *structural* generalization. The *Relational Cycle Benchmark* (RCB) derives from knowledge graphs and uses entity masking to expose relational structure, following Gentner (1983)'s Structure-Mapping Theory (SMT). The *Narrative Abstraction Benchmark* (NAB) tests higher-level abstraction by requiring models to align surface-dissimilar narratives through shared proverbs, which serve as proxies for abstract relational meaning (Gentner, 1983; Honeck, 1997).

We also provide a theoretical account in which AAT suppresses high-variance instance-specific gradients while increasing the relative contribution of shared relational structure, paralleling cognitive findings on schema-congruent encoding (Masís-Obando et al., 2022). Across both benchmarks and two model families, AAT improves structural generalization and retention. On RCB, additional multi-seed analysis also shows reduced variability across data orderings.

Our contributions are threefold: (i) AAT, a lightweight training objective that introduces structural abstraction during online learning; (ii) RCB and NAB, two benchmarks that operationalize retention and structural generalization at different abstraction levels; and (iii) theoretical and empirical evidence that abstraction can improve stable relational learning under non-stationarity.

2 METHOD

Cognitive accounts of human learning distinguish between *episodic* representations, which encode specific experienced events, and *schematic* representations, which capture relational patterns shared across events (Schank & Abelson, 2013). Effective long-term learning requires both: episodic detail supports recall of specific facts, while schematic structure supports generalization and reduces interference between related memories (Masís-Obando et al., 2022). We operationalize this distinction computationally within OCL by introducing AAT and analyzing its stabilizing effect.

2.1 ABSTRACTION-AUGMENTED ONLINE LEARNING

A well-documented failure mode in human learning occurs when surface-level details overwhelm relational structure, producing context-bound representations that impair transfer (Gick & Holyoak, 1983; Clement et al., 1994). A similar pattern arises when fine-tuning LLMs on knowledge-graph-derived streams, where entity-specific features can dominate updates and overwrite broader relational dependencies (Jang et al., 2021; Wu et al., 2023). Models may therefore retain individual training instances while losing structure needed for inductive generalization.

Formally, we consider a data stream $D = \{d_1, \ldots, d_m\}$ where each instance is observed once. At step i, model π_i receives batch $b_i = \{d_1^i, \ldots, d_B^i\}$ and must learn b_i while minimizing forgetting over $\bigcup_{k=1}^{i-1} b_k$. Drawing on SMT and theory-based categorization (Murphy & Medin,

Algorithm 1: Abstraction-Augmented Training

Input: Dataset $\mathcal{D}$, initial parameters θ, abstraction ratio α, local replay count n, learning rate η.
for each batch $\mathcal{B} \subset \mathcal{D}$ **do**
  for $j = 0$ **to** $n - 1$ **do**
    $\hat{y}_{inst} \leftarrow \pi(x_{inst}; \theta)$
    $\hat{y}_{abs} \leftarrow \pi(x_{abs}; \theta)$
    if $j = 0$ **then**
      $\mathcal{L}(\theta) \leftarrow \alpha \cdot \text{NLL}(\hat{y}_{abs}, y_{abs}) + (1-\alpha) \cdot \text{NLL}(\hat{y}_{inst}, y_{inst})$ *// Combined loss*
    else
      $\mathcal{L}(\theta) \leftarrow \text{NLL}(\hat{y}_{inst}, y_{inst})$
    end if
    $\theta \leftarrow \text{Adam}(\nabla_\theta \mathcal{L}(\theta), \theta, \eta)$
  end for
end for

1985), we treat abstraction as a gradient-level inductive bias: jointly optimizing a concrete instance and its structural abstraction suppresses entity-specific gradient components while increasing the relative contribution of shared relational structure.

AAT implements this principle with a dual-objective loss and local replay (Algorithm 1). Because each batch is encountered only once in OCL, we use a local replay count $n \in \{1, \ldots, 5\}$ to permit multiple updates per batch. At the first update ($j = 0$), the model receives both the concrete instance (x_{inst}) and its structural abstraction (x_{abs}), constructed by removing instance-specific surface features such as entity identity; the abstraction loss is weighted by α (default 0.5). Subsequent $n - 1$ updates use only the instance loss. This avoids repeatedly optimizing over highly repetitive abstract inputs while allowing later updates to consolidate instance-specific details.

2.2 Theoretical Motivation

We analyze why abstraction can stabilize learning under a simplified model. Although the assumptions are idealized, they isolate the mechanism tested empirically.

Setup. Consider a model that predicts relations from an input paragraph x using an encoder $\phi_\theta(\cdot)$ followed by a linear head W: $f_\theta(x) = W\phi_\theta(x)$. Each example contains relational structure r and entity identities $e_1, \ldots, e_k$. We decompose the representation as $\phi_\theta(x) = \phi_r(r) + \phi_e(e_1, \ldots, e_k)$, where ϕ_r captures structure shared across examples and ϕ_e captures instance-specific entity signals.

Assumptions.

- **A1. (Additive decomposition)** $\phi_\theta(x) = \phi_r(r) + \phi_e(e_1, \ldots, e_k)$, with ϕ_r and ϕ_e operating on disjoint parameter subsets.
- **A2. (Asymmetric gradient variance)** Entity tokens are long-tail and induce higher-variance gradients across sequential examples than relation tokens:
$$\sigma_e^2 = \mathrm{Var}(\nabla_\theta \phi_e^{(i)}) \gg \sigma_r^2 = \mathrm{Var}(\nabla_\theta \phi_r^{(i)}) \geq 0.$$
- **A3. (Entity masking)** Under entity masking, $\phi_\theta(x_{abs}) = \phi_r(r)$, so the abstract forward pass contributes no entity-specific gradient.

Proposition 1 (Abstraction reduces gradient variance)**.** *Under Assumptions **A1–A3**, the AAT combined loss $\mathcal{L}_i^{\text{total}} = \alpha\mathcal{L}_i^{\text{abs}} + (1-\alpha)\mathcal{L}_i^{\text{inst}}$, with $\alpha \in (0,1)$, combines a low-variance relational signal with a higher-variance instance signal. Consequently, $\mathrm{Var}\big(\nabla_\theta \mathcal{L}_i^{\text{total}}\big) < \mathrm{Var}\big(\nabla_\theta \mathcal{L}_i^{\text{inst}}\big)$, and the relational signal-to-noise ratio increases monotonically with α: $\frac{\|\nabla_\theta \phi_r^{(i)}\|}{\|(1-\alpha)\nabla_\theta \phi_e^{(i)}\|} \uparrow \alpha$.*

Remark 1. Assumption **A1** is the strongest idealization. Transformers do not maintain disjoint parameter subsets for relational and entity representations, so the proposition should be read as isolating a mechanism rather than characterizing transformer behavior exactly. The results in Section 5 test whether the predicted effect holds approximately in practice.

The key implication is that abstraction reduces the contribution of high-variance entity-specific gradients while increasing the relative influence of shared relational structure. This parallels cognitive findings that schema-congruent information is preferentially encoded and retained (Masís-Obando et al., 2022). Across many entity-specific instances, the shared abstraction therefore acts as an implicit structural regularizer, reinforcing relational patterns without revisiting past examples. Unlike ER, which stabilizes learning through historical replay, AAT stabilizes learning by reinforcing shared structure at the point of observation. A proof of Proposition 1 is provided in Appendix A.2.

3 Benchmarks

Schema-based learning distinguishes between episodic retention of observed facts and structural generalization to unobserved relations supported by shared patterns (Schank & Abelson, 2013; Gentner, 1983). We introduce two benchmarks that operationalize this distinction at different levels of abstraction. RCB instantiates abstraction through entity masking: removing surface-level identities exposes the relational structure that supports analogical alignment in SMT (Gentner, 1983). NAB instead targets implicit abstraction through proverbs, long used as vehicles for relational meaning across surface-dissimilar situations (Honeck, 1997). Together, the benchmarks test whether language models retain observed information while generalizing over shared structure.

Relational Cycle Benchmark. RCB probes both factual retention and deductive generalization. Figure 1 provides an overview. We construct a curated "Relation Bank" spanning eight semantic domains, including genealogy, sports, and professions, and use it to define 51 relational typologies, small subgraphs with fixed structures containing at least three relations arranged in an undirected cycle. In approximately 60% of typologies, one relation is logically entailed by the others; for example, $\textsc{Mother}(\text{Sarah}, \text{John}), \textsc{Mother}(\text{Sarah}, \text{Jane})$ entails $\textsc{Sibling}(\text{John}, \text{Jane})$. For each typology, we retrieve Wikidata entities (Vrandečić & Krötzsch, 2014) that instantiate the graph, yielding 1,245 instances and 3,975 entity-relation triples, including 680 entailed edges.

For each instance, when an explicit logical dependency exists, the entailed relation is held out and designated as the *unknown edge*, while the remaining relations are treated as *known edges*. Instances without an explicit logical dependency contribute only to known-edge evaluation. Known edges are serialized into a natural-language paragraph with lexical variation and distractor relations, such as $\textsc{Married}(\text{John}, \text{Alice})$ in Figure 1, and used for training. Recovering the unknown edge therefore requires identifying the relevant relational structure rather than direct occurrence in the training instance. This known-unknown distinction operationalizes the episodic-schematic distinction in schema-based accounts (Schank & Abelson, 2013): known edges are directly observed facts, whereas unknown edges must be recovered from shared relational structure. Statistics and zero-shot performances are reported in Appendix A.7.

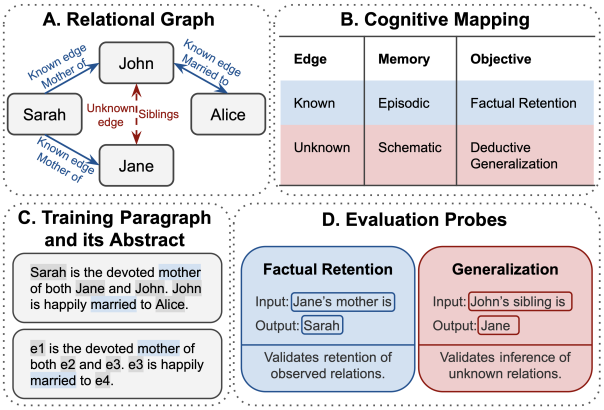


Figure 1: Overview of the Relational Cycle Benchmark.

Narrative Abstraction Benchmark. NAB probes a higher-order form of abstraction by aligning narratives through shared proverb-level relational motifs despite differences in surface form, entities, and event structure. Proverbs have long been used in the analogical reasoning literature as representations of abstract relational meaning that transfer across dissimilar situations (Gentner et al., 2001; Honeck, 1997). Unlike RCB, where abstraction is supplied explicitly through entity masking, NAB requires the model to recover an implicit semantic abstraction.

NAB is constructed from proverb-narrative pairs in ePiC (Ghosh & Srivastava, 2022), which contains 10 narratives per proverb, and expanded with 15 additional narratives per proverb using an SMT-informed analogical generation procedure. Each narrative is converted into a pairwise continuation task with a *correct ending* consistent with the proverb and a length-matched, locally coherent *distractor ending* that violates its relational motif (Table 1). Correct prediction therefore requires sensitivity to the proverb-level abstraction rather than superficial narrative coherence. To test whether the task can be solved from artifacts in the candidate endings alone, we additionally evaluate Qwen-1.5B with the narrative removed. On the validation set, accuracy is 45.7% zero-shot and 41.9% after SFT (seed 42). This Suggests that the endings alone do not provide a reliable shortcut and narrative context contributes substantially to the evaluated performance. Additional generation details, data samples, and zero-shot results are provided in Appendices A.6 and A.7.

Table 1: Example from the Narrative Abstraction Benchmark (NAB). The model must prefer the ending that preserves the proverb-level relational meaning over a locally plausible distractor.

Proverb/Abstract	**It takes two to tango.**
Narrative	Jeff and Jane were getting into a fight. Jane pointed the finger, saying it was all Jeff's fault. Their mother knew better though.
Correct ending	She knew that they both were causing trouble and Jeff and Jane were both punished.
Distractor	She quickly intervened and helped Jeff up off the floor, advising him to apologize to Jane.

4 EXPERIMENTAL SET-UP

We evaluate AAT in an OCL setting, focusing on the trade-off between episodic retention and structural generalization as learning progresses. This section describes the experimental protocol, baselines, and evaluation metrics used to assess learning dynamics and final performance. Unless otherwise stated, all models are trained under identical optimization and data stream conditions.

4.1 BASELINES AND EVALUATION

AAT's combined loss (Algorithm 1) introduces a structural stabilization term alongside the instance-level objective. At first exposure, this additional signal can be instantiated in two qualitatively different ways: using a structural abstraction of the current instance (x_{abs}), as in AAT, or using previously observed instances sampled from a replay buffer, as in experience replay (ER). Both approaches aim to reduce interference, but they do so through fundamentally different mechanisms. AAT exploits relational structure within the current example, whereas ER maintains temporal coverage by revisiting past examples. Their direct comparison therefore isolates the respective contributions of structural abstraction and historical replay to retention and generalization. We use standard fine-tuning (SFT), which optimizes only the instance-level objective, as a reference for measuring the effect of the additional stabilization signal. We also compare against elastic weight consolidation (EWC), which regularizes changes to parameters deemed important for prior examples (Kirkpatrick et al., 2017), and A-GEM, which constrains updates using gradients computed from stored examples (Chaudhry et al., 2018). All methods use the same number of local optimization steps, n, per batch to ensure comparable computation. Further implementation details are provided in Appendix A.4.

We adopt an interval-based evaluation protocol, assessing the model every k training steps. At evaluation step i, metrics are computed over the cumulative batches since the last evaluation, $\mathcal{B}_{cur} = \bigcup_{j=i-k+1}^{i} b_j$, and the full training history $\mathcal{H}_i = \bigcup_{j=1}^{i-k} b_j$. Correctness in RCB is defined by the model's predictive confidence: given a triple (e_1, r, e_2) serialized into a natural language prompt, an instance is correct if $\pi_i(e_2 \mid e_1, r) > \tau$. For NAB, an instance is correct if the conditional log-likelihood of the correct ending exceeds that of the distractor given the same narrative context. We define $\mathbb{I}(\pi_i(d)) = 1$ if the prediction is correct and 0 otherwise.

Online Accuracy ($\text{Acc}_{\text{online}}$). Measures model plasticity, acquiring recently encountered information:

$$\text{Acc}_{\text{online}}(i) = \frac{1}{|\mathcal{B}_{cur}|} \sum_{d \in \mathcal{B}_{cur}} \mathbb{I}(\pi_i(d)) \tag{1}$$

Cumulative Accuracy ($\text{Acc}_{\text{cumul}}$). Provides a global view of the model's knowledge base, capturing backward transfer as learning new relational structures may clarify previously forgotten edges:

$$\text{Acc}_{\text{cumul}}(i) = \frac{1}{|\bigcup_{j=1}^{i} b_j|} \sum_{d \in \bigcup_{j=1}^{i} b_j} \mathbb{I}(\pi_i(d)) \tag{2}$$

Forgetting (F). Measures stability by tracking loss of previously acquired knowledge. An instance is forgotten if correctly predicted at some prior step $t < i$ but incorrectly predicted at step i, normalized by the set of instances the model learned at least once:

$$F(i) = \frac{|\{d \in \mathcal{H}_i \mid \mathbb{I}(\pi_i(d)) = 0 \wedge \exists t < i : \mathbb{I}(\pi_t(d) = 1)\}|}{|\{d \in \mathcal{H}_i \mid \exists t < i : \mathbb{I}(\pi_t(d) = 1)\}|} \tag{3}$$

4.2 EXPERIMENTS & ABLATIONS

We use Qwen2.5-1.5B (Team, 2024) as the primary backbone model trained with a language modeling objective. Unless otherwise specified, AAT uses $\alpha = 0.5$ and $n = 5$ (implementation details in Section A.4). We first examine the impact of abstraction on training by tracking forgetting and online accuracy throughout training on RCB, reported separately for known and unknown edges. To empirically investigate the theoretical claim that abstraction reduces gradient variance, we perform a loss landscape analysis using linear interpolation in model weight space (Li et al., 2018), measuring variance and coefficient of variation along the primary training trajectory and a perpendicular axis.

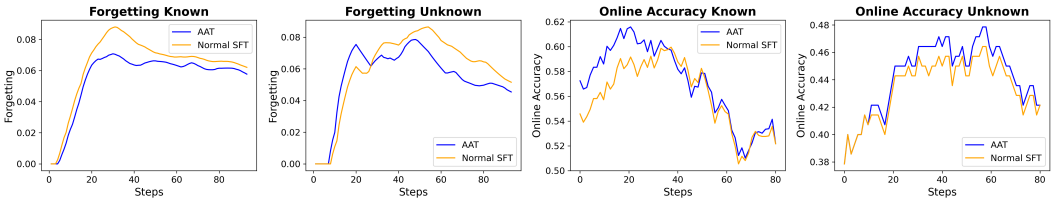


Figure 2: Online accuracy and forgetting over training steps on the RCB.

To analyze abstraction effectiveness, we ablate the structural abstraction levels: beyond entity masking, we evaluate category-based abstraction (entities replaced with semantic categories such as city, institution, or person) and random abstracts (entity-masked sentences combined without structural coherence) as a control. We also evaluate AAT on NAB to assess whether the cognitive benefits of abstraction generalize from explicit symbolic structure to implicit narrative-level relational motifs. For NAB, we use $n = 1$ and $\alpha = 0.15$; with 90% of narratives per proverb for training and 10% held out as unknown. Additional ablations on sensitivity to α, n, and data order in Section A.3.

Finally, we compare AAT against all baselines across RCB metrics: cumulative accuracy ($\text{Acc}_{\text{cumul}}$), step-averaged online accuracy ($\text{Acc}_{\text{online}}$), and final-step forgetting (F), measured separately for known and unknown edges. We additionally train SmolLM-1.7B (Allal et al., 2025) to assess whether AAT's inductive bias generalizes across model families with substantially different pretraining.

5 Results

We organize results around three questions. First, what does abstraction do to learning dynamics and loss geometry? Second, what properties of the abstraction drive its effect? Third, how does learning via abstraction compare to learning via instance replay, and what does each mechanism preserve?

5.1 Training Behavior and Loss Geometry

As shown in Figure 2, SFT and AAT exhibit similar high-level dynamics: forgetting rises following domain transitions and then gradually stabilizes, while online accuracy evolves in the complementary direction. Known-edge accuracy initially increases before declining later in training, whereas unknown-edge accuracy improves substantially as training progresses. Despite these shared trends, AAT exhibits lower sustained forgetting after the initial transition and generally higher online accuracy, particularly on unknown edges and during the early stages of known-edge learning. This pattern is consistent with the structural abstraction signal reducing interference once learning stabilizes.

This effect is also reflected in the geometry of the loss landscape. Compared with SFT, AAT reduces loss-surface variance by 17.00% and lowers the coefficient of variation from 0.1549 to 0.1375, consistent with optimization in smoother, lower-interference regions. Along the training trajectory, the abstract loss improves by an average of 0.243 nats (up to 0.427 nats), while the corresponding increase in instance loss remains small (0.038 nats on average). Although absolute loss is marginally higher in AAT (2.64% on average), its lower variability indicates a smoother optimization landscape.

We further examine sensitivity to data ordering across five random permutations of the RCB stream (Table 4). SFT exhibits the highest variability on unknown-edge cumulative accuracy (0.3571 ± 0.0325), while AAT achieves higher mean performance with lower variability (0.3781 ± 0.0213). AAT also reduces variability in overall cumulative accuracy (0.4891 ± 0.0091 vs. 0.4683 ± 0.0159 for SFT). Under one-sided paired t-tests with global Holm correction, AAT significantly improves cumulative accuracy relative to SFT. These results suggest that abstraction improves performance while making outcomes less sensitive to data ordering.

These results parallel the cognitive predictions motivating our experimental design. Interference accounts of human memory predict that schema-congruent encoding can reduce interference between related memories by organizing new experiences around shared structure (Masís-Obando et al., 2022). AAT exhibits an analogous pattern: lower sustained forgetting, smoother optimization, and reduced variability across training orders. Prior work on progressive abstraction shows that the structure

Table 2: Performance under RCB's alternative abstraction levels and training strategies on NAB.

(a) Performance of abstraction levels.

Method	Cum. Accuracy ↑		Forgetting ↓	
	Unknown	Known	Unknown	Known
Random	0.390	0.476	0.085	0.116
Cat. Mask	0.406	0.540	0.066	0.071
Ent. Mask	**0.409**	**0.545**	**0.064**	**0.068**

(b) Performance comparison on NAB across ten seeds.

Method	Cum. Accuracy ↑		Forgetting ↓	
	Unknown	Known	Unknown	Known
SFT	0.597 ± 0.006	0.862 ± 0.016	0.166 ± 0.005	$\mathbf{0.030 \pm 0.008}$
ER	0.592 ± 0.008	$\mathbf{0.886 \pm 0.019}$	0.190 ± 0.007	0.034 ± 0.008
AAT	$\mathbf{0.601 \pm 0.006}$	0.861 ± 0.017	$\mathbf{0.160 \pm 0.004}$	$\mathbf{0.030 \pm 0.009}$

learned from sequential experience can depend strongly on presentation order (Skorstad et al., 1988; Kuehne et al., 2000). Against this background, AAT's lower variability across random orderings suggests that explicitly supplying structural abstraction can make learning more robust to sequential variation. These correspondences do not imply that the underlying mechanisms are identical, but suggest that abstraction induces computational behavior aligned with schema-based accounts.

5.2 What Makes Abstraction Effective

Having established that abstraction stabilizes learning dynamics, we next ask what properties of the abstraction drive this effect. Table 2a compares three abstraction levels: entity masking, category-based abstraction, and random abstracts. Category-based abstractions replace entities with semantic category labels (e.g., "Person_3 is happily married to Person_4" for Alice and John in Figure 1), while random abstracts preserve the same number of relations as the corresponding structured abstraction but randomly recombine relations into a new relational graph (e.g., "Entity_3 is a professor in Entity_4" for Alice and John in Figure 1).

Random abstracts degrade performance relative to the structured masking methods, yielding lower cumulative accuracy and higher forgetting on both edge types. Because this condition adds an auxiliary relational signal while breaking the correspondence between the instance and its abstract structure, the benefit of AAT cannot be explained by additional data or auxiliary supervision alone. Instead, the improvement depends on the abstract representation preserving the relational structure of the instance with which it is jointly optimized. When that correspondence is destroyed, the additional abstract signal becomes detrimental rather than stabilizing.

The results further exhibit a clear gradation, random $<$ category $<$ mask. Category-based abstraction retains coarse semantic information about the entities, whereas entity masking removes surface identity more completely, exposing the shared relational structure more directly. This ordering is consistent with predictions from SMT: analogical generalization improves as surface-specific information is suppressed and relational structure becomes more salient. Together, these results show that AAT's benefit depends both on maintaining the structural correspondence between an instance and its abstraction and on the degree to which the abstraction suppresses surface information.

5.3 Generalization to Alternative Abstractions

To assess whether these findings generalize beyond explicitly structured relational graphs, Table 2b reports performance on NAB across ten seeds, where abstraction is implicit in shared proverb-level meaning across surface-dissimilar narratives. ER achieves the highest known-narrative accuracy, improving over SFT by 0.0241, and this gain remains significant after global Holm correction ($p_{\text{adj}} = 2.41 \times 10^{-7}$). However, ER performs worse than SFT on unknown narratives and exhibits substantially greater forgetting there (0.190 vs. 0.166). AAT shows the complementary pattern. It achieves the highest unknown-narrative accuracy while reducing unknown-narrative forgetting from 0.166 to 0.160. Both effects remain significant after global Holm correction ($p_{\text{adj}} = 1.23 \times 10^{-4}$ for accuracy; $p_{\text{adj}} = 0.0053$ for forgetting). Known-narrative performance does not differ significantly from SFT after correction. This profile suggests that AAT's primary benefit on NAB lies in preserving and generalizing shared relational structure rather than improving retention of observed narratives.

The NAB setting is especially informative because the relevant abstraction is not explicitly encoded in the input. Unlike RCB, there is no entity masking or predefined relational template; the shared structure is a moral or causal motif expressed across narratives that differ in domain, entities, and event structure. AAT's significant gains on unknown-narrative accuracy and forgetting therefore

Table 3: Step-averaged online accuracy, final-step forgetting, and cumulative accuracy on the RCB benchmark. Results are mean ± standard deviation across three seeds.

Model	Method	Online Accuracy ↑		Forgetting ↓		Cumulative Accuracy ↑		
		Unknown	Known	Unknown	Known	Unknown	Known	All
Qwen-1.5B	SFT	0.3777 ±0.032	0.5453 ±0.003	0.0873 ±0.050	0.0626 ±0.020	0.3482 ±0.039	0.5227 ±0.012	0.4674 ±0.020
	ER	0.3958 ±0.008	0.5584 ±0.014	0.0900 ±0.032	0.0629 ±0.019	0.3645 ±0.019	0.5371 ±0.016	0.4832 ±0.017
	A-GEM	0.3945 ±0.008	0.5536 ±0.013	0.0856 ±0.038	**0.0487 ±0.015**	0.3645 ±0.022	0.5400 ±0.015	0.4852 ±0.017
	EWC	0.4008 ±0.007	0.5617 ±0.012	**0.0785 ±0.033**	**0.0487 ±0.013**	0.3732 ±0.019	**0.5478 ±0.010**	**0.4932 ±0.013**
	AAT	**0.4034 ±0.030**	**0.5673 ±0.008**	0.0792 ±0.042	0.0563 ±0.025	**0.3745 ±0.026**	0.5474 ±0.007	0.4927 ±0.007
Qwen-3B	SFT	0.3827 ±0.041	0.5744 ±0.026	0.1010 ±0.018	0.0682 ±0.029	0.3457 ±0.043	0.5482 ±0.040	0.4844 ±0.041
	ER	0.3983 ±0.034	0.5757 ±0.028	0.1150 ±0.030	0.0664 ±0.036	0.3544 ±0.039	0.5504 ±0.046	0.4886 ±0.044
	A-GEM	0.3931 ±0.044	0.5739 ±0.026	0.0987 ±0.021	0.0751 ±0.035	0.3561 ±0.048	0.5439 ±0.042	0.4848 ±0.043
	EWC	0.3974 ±0.044	**0.5807 ±0.016**	0.1022 ±0.018	0.0567 ±0.013	0.3587 ±0.046	**0.5575 ±0.025**	**0.4949 ±0.030**
	AAT	**0.4044 ±0.030**	0.5700 ±0.031	**0.0845 ±0.021**	**0.0518 ±0.014**	**0.3717 ±0.033**	0.5498 ±0.039	0.4937 ±0.037
SmolLM-1.7B	SFT	0.5946 ±0.005	0.7470 ±0.007	0.0543 ±0.030	0.0229 ±0.014	0.5675 ±0.015	0.7397 ±0.007	0.6858 ±0.006
	ER	0.6030 ±0.002	0.7461 ±0.007	0.0583 ±0.036	0.0256 ±0.019	0.5735 ±0.015	0.7365 ±0.008	0.6854 ±0.009
	A-GEM	0.6024 ±0.002	0.7496 ±0.004	0.0621 ±0.038	0.0200 ±0.014	0.5706 ±0.017	0.7433 ±0.005	0.6891 ±0.007
	EWC	0.6050 ±0.003	**0.7511 ±0.007**	0.0609 ±0.035	0.0194 ±0.016	0.5732 ±0.014	0.7456 ±0.006	0.6915 ±0.006
	AAT	**0.6060 ±0.000**	0.7503 ±0.006	**0.0524 ±0.026**	**0.0172 ±0.014**	**0.5808 ±0.006**	**0.7462 ±0.006**	**0.6944 ±0.004**

suggest that its benefit is not specific to the explicit relational structure of RCB. More generally, jointly optimizing over instances and abstractions appears to bias learning toward shared structure even when that structure is defined at a higher semantic level.

5.4 Abstraction versus Continual Learning Baselines

Having established what abstraction contributes to learning, we next compare AAT with standard continual learning approaches. As described in Section 4.1, these methods stabilize learning through different mechanisms: ER replays stored instances, A-GEM constrains updates using gradients from past examples, EWC regularizes parameters important to previous observations, whereas AAT derives its stabilizing signal from the structural abstraction of the current instance. The comparison therefore asks whether these mechanisms produce different profiles of retention and generalization.

Performance on RCB is reported in Table 3. Across all three models, AAT improves cumulative accuracy over normal SFT on both known and unknown edges, with the largest gains consistently occurring on unknown relations. AAT also reduces final-step forgetting relative to SFT on both edge types across all three models. Thus, AAT improves retention while providing its strongest gains on structural generalization. For Qwen-1.5B and Qwen-3B, unknown-edge cumulative accuracy increases from 0.3482 to 0.3745 and from 0.3457 to 0.3717, respectively, while improvements on known edges are smaller. SmolLM shows the same qualitative pattern, with AAT increasing unknown-edge cumulative accuracy from 0.5675 to 0.5808 while maintaining a smaller gain on known edges. Thus, AAT's primary advantage lies in structural generalization rather than only retention of previously observed relations.

The continual learning baselines exhibit a different profile. ER, A-GEM, and EWC are often competitive on known-edge accuracy and forgetting, but their gains on unknown relations are less consistent across models. AAT achieves the highest unknown-edge cumulative accuracy for all three models while remaining competitive on known edges and reducing final forgetting relative to SFT throughout. This pattern is consistent with the different inductive biases of the methods:

conventional continual learning objectives primarily protect previously acquired information, whereas AAT promotes relational structure that can transfer across instances.

The overall cumulative-accuracy improvements are also supported by significance testing across random data orderings. Using one-sided paired t-tests against SFT, with Holm correction for multiple comparisons, all four methods yield significant improvements in overall cumulative accuracy for Qwen-1.5B ($p_{\text{adj}} < 0.05$). For SmolLM-1.7B, only EWC ($p_{\text{adj}} = 0.0101$) and AAT ($p_{\text{adj}} = 0.0230$) remain significant after correction. For Qwen-3B, none of the pairwise comparisons reach significance after correction ($p_{\text{adj}} > 0.05$). Together with AAT's consistent gains on unknown relations and lower final forgetting relative to SFT across all three models, these results suggest that abstraction can provide a complementary route to stability; rather than only protecting previously acquired instances or parameters, it promotes structure that transfers across examples.

6 Related Work

Benchmarks. NLP continual learning benchmarks focus on task-incremental settings with explicit boundaries (Aghajanyan et al., 2020; Wu et al., 2021; Yadav et al., 2023; Wang et al., 2023; Zhao et al., 2025; Zhang et al., 2023), extending to domain or language shifts (Ke et al., 2023; Castellucci et al., 2021; M'hamdi et al., 2023; Winata et al., 2023). Parallel work explores OCL with non-stationary streams (Li et al., 2025) or knowledge-graph updates (Jang et al., 2021; Wu et al., 2023), but these benchmarks typically represent each update as a single fact or sentence and do not distinguish episodic retention from structural generalization. RCB and NAB are specifically designed to probe this distinction across different abstraction granularities, filling a gap left by existing benchmarks.

Methods. Continual leaning methods generally mitigate forgetting through regularization (Kirkpatrick et al., 2017; Chaudhry et al., 2018) or ER (Ostapenko et al., 2022; Smith et al., 2024; Bellitto et al., 2024; Zhuo et al., 2023; Yan et al., 2022). Replay is often effective (Van de Ven et al., 2020) but requires maintaining a memory buffer, scaling poorly in strictly online settings. Architectural strategies such as adapters and parameter expansion reduce interference (Gururangan et al., 2022; M'hamdi et al., 2023; Chen et al., 2024) but typically require task boundaries and continuous parameter growth. AAT occupies a distinct position: it introduces abstraction as a loss-level inductive bias at the moment of observation, targeting the instance-specific gradients that drive interference.

Abstraction in Cognition. AAT draws on SMT, which models analogy as alignment of relational structure across domains while abstracting from surface features. Work on analogical transfer shows that relational encoding supports cross-context generalization beyond instance-level similarity (Gick & Holyoak, 1983; Clement et al., 1994). Neuroscience studies further suggest that related experiences are organized around shared components, supporting abstraction and stable memory (Masís-Obando et al., 2022; Mattar & Daw, 2018). AAT operationalizes these ideas as a training signal to address challenges of non-stationary language model training.

7 Conclusion

Drawing on cognitive accounts of schema-based learning, we introduced AAT to test whether structural abstraction can promote stable learning under non-stationarity. Across two model families and two benchmarks, AAT improves generalization to unobserved instances and exhibits lower sustained forgetting, while multi-seed RCB analysis shows reduced variability across data orderings. Ablations show that these gains depend on preserving the correspondence between an instance and its abstraction, with performance improving as surface-specific information is suppressed. Compared with conventional continual learning methods, AAT's largest and most consistent gains occur on unknown relations, suggesting a complementary route to stability through transferable structure.

These findings motivate several directions for future work. For machine learning, an important question is whether the same principles extend to larger models, less explicitly structured tasks, and naturalistic continual learning settings. For cognitive science, the results provide a computational framework for testing when structural abstraction is sufficient to support stable generalization and for evaluating other theories of schema-based learning in artificial systems. More broadly, we view this work as a step toward using cognitive theory not only as inspiration for model design, but as a source of testable computational hypotheses about learning and generalization.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, including grammar correction, word choice, and improving the clarity and flow of the manuscript. All AI-assisted revisions were subsequently reviewed and proofread by the authors.

We also used GitHub Copilot for code autocompletion during implementation. AI-suggested code was reviewed and tested by the authors before use.

For construction of the Narrative Abstraction Benchmark (NAB), we used an open-weight language model to generate synthetic narratives. The generation procedure and prompts used to create these data are disclosed in the paper and appendix.

We did not use generative AI tools to formulate or prove mathematical claims. The authors take responsibility for the final content of the paper, including all text, code, analyses, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the code and datasets used in this work as supplementary material. The appendix reports the experimental hyperparameters, random seeds, implementation details, and additional benchmark construction information needed to reproduce the reported results. Theoretical assumptions and the proof of the main proposition are also provided in the main paper and appendix. Upon publication, we plan to release the code and datasets publicly on GitHub.

REFERENCES

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*, 2020.

Hongjoon Ahn, Jinu Hyeon, Youngmin Oh, Bosun Hwang, and Taesup Moon. Prevalence of negative transfer in continual reinforcement learning: Analyses and a simple baseline. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 75036–75060, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ba9e3d60610f3525717665966d86e0cd-Paper-Conference.pdf.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Giovanni Bellitto, Federica Proietto Salanitri, Matteo Pennisi, Matteo Boschini, Lorenzo Bonicelli, Angelo Porrello, Simone Calderara, Simone Palazzo, and Concetto Spampinato. Saliency-driven experience replay for continual learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 103356–103383. Curran Associates, Inc., 2024. doi: 10.52202/079017-3284. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bb1e9f32181a8d6a834670d5b3e1c48d-Paper-Conference.pdf.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b704ea2c39778f07c617f6b7ce480e9e-Paper.pdf.

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. Learning to solve NLP tasks in an incremental number of languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 837–847, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.106. URL https://aclanthology.org/2021.acl-short.106/.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Cheng Chen, Junchen Zhu, Xu Luo, Heng T Shen, Jingkuan Song, and Lianli Gao. Coin: A benchmark of continual instruction tuning for multimodel large language models. *Advances in Neural Information Processing Systems*, 37:57817–57840, 2024.

Catherine A Clement, Ronald Mawby, and Denise E Giles. The effects of manifest relational similarity on analog retrieval. *Journal of Memory and Language*, 33(3):396–420, 1994.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024.

Vivek Farias and Adam Jozefiak. Self-normalized resets for plasticity in continual learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 20852–20886, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/359c0c4dc53b6bf64041b4c8334f1c11-Paper-Conference.pdf.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2): 155–170, 1983.

Dedre Gentner. Why we're so smart. *Language in mind: Advances in the study of language and thought*, 195235, 2003.

Dedre Gentner, Brian Bowdle, Phillip Wolff, Consuelo Boronat, et al. Metaphor is like analogy. *The analogical mind: Perspectives from cognitive science*, pp. 199–253, 2001.

Sayan Ghosh and Shashank Srivastava. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

Mary L Gick and Keith J Holyoak. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38, 1983.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015. URL https://arxiv.org/abs/1312.6211.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. DEMix layers: Disentangling domains for modular language modeling. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5557–5576, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.407. URL https://aclanthology.org/2022.naacl-main.407/.

Douglas R Hofstadter et al. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pp. 499–538, 2001.

RP Honeck. Proverb in mind, the cognitive science of proverbial wit and wisdom, mahwah, nj and london. *Lawrence Ebaum*, 1997.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*, 2021.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Sven Kuehne, Kenneth Forbus, Dedre Gentner, and Bryan Quinn. Seql: Category learning as progressive abstraction using structure mapping. In *Proceedings of the 22nd annual meeting of the cognitive science society*, pp. 770–775. Pasadena, CA, 2000.

Alex Lewandowski, Dale Schuurmans, and Marlos C. Machado. Plastic learning with deep fourier features. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 31112–31137, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4d42358702dff82e1436550a05ade260-Paper-Conference.pdf.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, and Fartash Faghri. TiC-LM: A web-scale benchmark for time-continual LLM pretraining. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32231–32273, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1551. URL https://aclanthology.org/2025.acl-long.1551/.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lyle23b.html.

Rolando Masís-Obando, Kenneth A Norman, and Christopher Baldassano. Schema representations in distinct brain networks support narrative memory during encoding and retrieval. *Elife*, 11:e70445, 2022.

Marcelo G Mattar and Nathaniel D Daw. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609–1617, 2018.

Meryem M'hamdi, Xiang Ren, and Jonathan May. Cross-lingual continual learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3908–3943, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.217. URL https://aclanthology.org/2023.acl-long.217/.

Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15699–15717. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mirzadeh22a.html.

Gregory L Murphy and Douglas L Medin. The role of theories in conceptual coherence. *Psychological review*, 92(3):289, 1985.

Cuong V Nguyen, Alessandro Achille, Michael Lam, Tal Hassner, Vijay Mahadevan, and Stefano Soatto. Toward understanding catastrophic forgetting in continual learning. *arXiv preprint arXiv:1908.01091*, 2019.

Oleksiy Ostapenko, Timothee Lesort, Pau Rodriguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Continual learning with foundation models: An empirical study of latent replay. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 60–91. PMLR, 22–24 Aug 2022. URL https://proceedings.mlr.press/v199/ostapenko22a.html.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Paper.pdf.

Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press, 2013.

Janice Skorstad, Dedre Gentner, and Douglas Medin. Abstraction processes during concept learning: A structural view. In *Proceedings of the Tenth Annual Meeting of the Cognitive Science Society*, volume 10, pp. 419–425, 1988. URL https://escholarship.org/uc/item/9mg8p8hp.

James Seale Smith, Lazar Valkov, Shaunak Halbe, Vyshnavi Gutta, Rogerio Feris, Zsolt Kira, and Leonid Karlinsky. Adaptive memory replay for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3605–3615, June 2024.

Timothy Tadros, Giri P Krishnan, Ramyaa Ramyaa, and Maxim Bazhenov. Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. *Nature communications*, 13(1): 7742, 2022.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):4069, 2020.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Jiuqi Wang, Rohan Chandra, and Shangtong Zhang. Experience replay addresses loss of plasticity in continual learning. *arXiv preprint arXiv:2503.20018*, 2025.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, et al. Trace: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*, 2023.

Genta Indra Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 768–777, 2023.

Tongtong Wu, Xuekai Li, Yuan-Fang Li, Gholamreza Haffari, Guilin Qi, Yujin Zhu, and Guoqiang Xu. Curriculum-meta learning for order-robust continual relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10363–10369, 2021.

Yuhao Wu, Tongjun Shi, Karthick Sharma, Chun Wei Seah, and Shuhao Zhang. Online continual knowledge learning for language models. *arXiv preprint arXiv:2311.09632*, 2023.

Prateek Yadav, Qing Sun, Hantian Ding, Xiaopeng Li, Dejiao Zhang, Ming Tan, Parminder Bhatia, Xiaofei Ma, Ramesh Nallapati, Murali Krishna Ramanathan, Mohit Bansal, and Bing Xiang. Exploring continual learning for code generation models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 782–792, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.68. URL https://aclanthology.org/2023.acl-short.68/.

Qingsen Yan, Dong Gong, Yuhang Liu, Anton van den Hengel, and Javen Qinfeng Shi. Learning bayesian sparse networks with full experience replay for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 109–118, June 2022.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. CITB: A benchmark for continual instruction tuning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9443–9455, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.633. URL https://aclanthology.org/2023.findings-emnlp.633/.

Hongbo Zhao, Fei Zhu, Haiyang Guo, Meng Wang, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453*, 2025.

Tao Zhuo, Zhiyong Cheng, Zan Gao, Hehe Fan, and Mohan Kankanhalli. Continual learning with strong experience replay. *arXiv preprint arXiv:2305.13622*, 2023.

A Appendix

The appendix contains limitations (Section A.1), the proof of Proposition 1 (Section A.2), additional ablations (Section A.3), implementation details (Section A.4), NAB construction procedures (Sections A.5 and A.6), and dataset statistics and samples (Section A.7).

A.1 Limitations

Our study has several limitations. First, the theoretical analysis relies on a simplified decomposition of relational and entity-specific representations, including the assumption that they occupy disjoint parameter subsets. This assumption is not expected to hold exactly in Transformers; rather, it provides an idealized setting for isolating the mechanism by which abstraction can suppress entity-specific contributions while preserving relational signal. The analysis should therefore be interpreted as a mechanistic account under simplifying assumptions, rather than a complete characterization of optimization dynamics in language models.

Second, our experiments focus on relatively small language models and controlled non-stationary settings. Although the observed patterns are consistent across two model families and two benchmarks with different forms of abstraction, it remains unclear whether they will extend to substantially larger models, longer training streams, or more naturalistic environments.

Third, RCB is deliberately structured to separate observed relations from relations recoverable through relational dependencies. This controlled design facilitates mechanistic analysis but does not capture the full complexity of real-world knowledge acquisition. NAB partially addresses this limitation by testing implicit, proverb-level abstraction across surface-dissimilar narratives, but it also remains a constructed benchmark. Additionally, because RCB is constructed from Wikidata, some relations may have appeared in model pretraining; the zero-shot evaluation quantifies pre-existing performance but cannot rule out latent pretrained knowledge.

Finally, the connection to cognitive science established in this work is behavioral and computational rather than mechanistic. As discussed in Section 5, similarities between AAT's learning behavior and findings from human abstraction do not imply that language models and humans implement abstraction through the same internal processes.

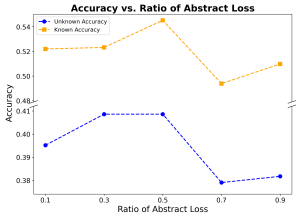


Figure 3: Effect of abstraction loss weight α on cumulative accuracy.

Table 4: Effect of data order on cumulative accuracy across five seeds.

Method	$\text{Acc}^{\text{unknown}}_{\text{cumul}}$	$\text{Acc}^{\text{known}}_{\text{cumul}}$	$\text{Acc}^{\text{all}}_{\text{cumul}}$
SFT	0.3571 $\pm$ 0.0325	0.5197 $\pm$ 0.0111	0.4683 $\pm$ 0.0159
ER	0.3702 $\pm$ 0.0165	0.5427 $\pm$ 0.0143	0.4887 $\pm$ 0.0148
AAT	0.3781 $\pm$ 0.0213	0.5404 $\pm$ 0.0122	0.4891 $\pm$ 0.0091

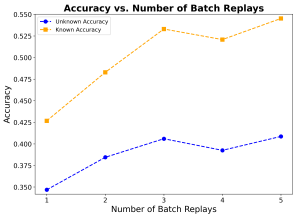


Figure 4: Effect of local replay count n on cumulative accuracy.

A.2 Proof of Proposition 1

Proof. By **A3**, the abstract loss contributes only a relational gradient: $\nabla_\theta \mathcal{L}_i^{\text{abs}} \propto W^\top \nabla \ell \cdot \nabla_\theta \phi_r^{(i)}$. The combined gradient follows by linearity:

$$\nabla_\theta \mathcal{L}_i^{\text{total}} \propto W^\top \nabla \ell \cdot \left[\nabla_\theta \phi_r^{(i)} + (1-\alpha) \nabla_\theta \phi_e^{(i)} \right].$$

By **A1**, the two components operate on disjoint parameter subsets, so:

$$\text{Var}\left(\nabla_\theta \mathcal{L}_i^{\text{total}}\right) \propto \sigma_r^2 + (1-\alpha)^2 \sigma_e^2.$$

Since $\alpha \in (0,1)$, we have $(1-\alpha)^2 < 1$, so the entity variance term is strictly reduced relative to the instance-only case $\sigma_r^2 + \sigma_e^2$. The signal-to-noise improvement follows because $(1-\alpha) < 1$ strictly decreases the entity-specific term in the denominator. □

A.3 Additional Ablation Studies

We conduct a series of ablation studies to analyze the sensitivity of AAT to its key hyperparameters and environmental factors. To study the abstraction loss weight, we train models with $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ while fixing the local replay count to $n = 5$. Conversely, to evaluate the effect of local replay, we fix $\alpha = 0.5$ and vary n from 1 to 5. In both cases we report cumulative accuracy at the final training step. To assess sensitivity to data order, we train each method on five random permutations of the data stream and report the mean and standard deviation of cumulative accuracy. In the online setting, performance variability across permutations reflects robustness to sequential interference and early sample bias.

A.3.1 Effect of AAT hyperparameters

The effect of abstract loss weight (α) in AAT is reported in Figure 3. For $\alpha < 0.5$, we observe substantial improvements on unknown edges, indicating enhanced inductive generalization, while gains on known edges remain limited. Performance peaks at $\alpha = 0.5$, where both known and unknown edges improve simultaneously, suggesting a favorable balance between instance-level supervision and abstraction-based regularization.

Beyond this point, performance degrades for both edge types. Although abstraction acts as a regularizer that encourages learning shared relational structure and reduces interference, assigning it excessive weight appears to suppress instance-specific learning. One plausible explanation is that heavily weighting the abstraction loss leads the model to overfit to repetitive abstract inputs, as multiple concrete instances share the same abstract representation. This result underscores the importance of balancing abstract and instance-level objectives rather than allowing abstraction to dominate the training signal.

Figure 4 shows the effect local replay count (n) on AAT. We observe a consistent improvement in performance on both known and unknown edges as the number of local replays increases. This trend indicates that additional local updates help stabilize learning in the strictly online setting, where each data point is observed only once. The gains from increasing n are particularly pronounced for known edges, suggesting that local replay enables more effective consolidation of newly introduced entities and their relations. While there is a slight performance dip at $n = 4$, performance recovers

at $n = 5$, indicating that the overall trend is robust. These results demonstrate that moderate local replay complements abstraction by improving instance-level learning without requiring long-term memory buffers.

A.3.2 Effect of Data Order (Random Seeds)

We evaluate sensitivity to data ordering across five random permutations of the RCB stream (Table 4). SFT exhibits the largest variability on unknown-edge cumulative accuracy (0.3571 ± 0.0325), whereas both ER and AAT reduce this variance. AAT achieves the highest mean unknown-edge accuracy (0.3781 ± 0.0213) and the lowest variability in overall cumulative accuracy (0.4891 ± 0.0091), compared with 0.4683 ± 0.0159 for SFT.

Using one-sided paired t-tests with global Holm correction across the method–metric comparisons, both ER and AAT significantly improve known-edge and overall cumulative accuracy relative to SFT ($p_{\text{adj}} < 0.05$). AAT also improves unknown-edge accuracy by 0.0209, although this comparison is significant at the 0.10 level after correction ($p_{\text{adj}} = 0.0823$). ER's unknown-edge improvement is likewise not significant.

Overall, these results suggest that both replay and abstraction improve robustness to data ordering, while AAT yields the highest mean performance on unknown edges and the most stable overall accuracy without requiring storage of past examples.

A.4 Implementation Details

Training setup. We use a learning rate of 5×10^{-5} across all training methods. Unless otherwise specified, experiments use a per-device batch size of 4 and are run on three NVIDIA RTX A6000 GPUs. Our implementation is based on the Hugging Face ecosystem, using `transformers 4.56.1`, `torch 2.8.0`, and `accelerate 1.7.0`. For multi-seed experiments, we draw seeds in order from the fixed list $[42, 16, 25, 123, 54, 11, 33, 66, 77, 99]$, using the first k seeds for experiments with k runs. The seeds induce different data orderings while keeping all other training configurations fixed.

AAT hyperparameters. The default AAT configuration for RCB uses $n = 5$ local updates and $\alpha = 0.5$. NAB instead uses $n = 1$ and $\alpha = 0.15$. We use fewer local updates for NAB because the two benchmarks differ substantially in the information that must be acquired from each example. RCB instances contain multiple explicit relations and entity-specific facts, including relatively rare entities, making repeated local updates useful for acquiring the information contained in a single instance. NAB examples are natural-language narratives whose evaluation depends primarily on retaining the higher-level relational motif rather than memorizing each entity or surface detail. Additional local replay therefore provides substantially less benefit in this setting, and we use $n = 1$ to avoid unnecessary repeated optimization on individual narratives. The effect of n and α on AAT is examined further in Section A.3.

Continual-learning baselines. For experience replay (ER), we maintain the replay buffer using reservoir sampling, allowing each previously observed example an equal probability of remaining in the fixed-capacity buffer as the stream progresses. The buffer is updated at every training step, and replay examples are sampled with a replay batch size equal to the current training batch size. We use a warm-up period during which replay is disabled until the buffer reaches its target capacity, preventing early updates from repeatedly sampling from a very small set of examples.

We evaluate ER configurations with different buffer capacities and local replay counts and select the configuration with the highest cumulative validation accuracy. The best-performing configuration uses a buffer of 100 examples. We additionally examine the effect of local replay for ER in Table 5; increasing the number of local updates from $n = 1$ to $n = 5$ improves cumulative accuracy in the configuration shown. This comparison also ensures that the ER baseline is not disadvantaged by receiving fewer local optimization steps than AAT.

For EWC, we use a regularization coefficient of $\lambda = 0.1$ and a decay factor of 0.9. For A-GEM, we use a memory buffer of 128 examples. We use the same local replay count for EWC and A-GEM as for AAT and ER so that all methods receive a comparable number of optimization steps. Baseline-specific hyperparameters are selected using cumulative accuracy on the validation set.

RCB evaluation and threshold calibration. As described in the main text, RCB predictions are conceptually defined in terms of model confidence. In implementation, we operate directly in logit space and apply a calibrated confidence threshold τ to the model outputs. For each model backbone, we collect logits from a held-out calibration subset using the pretrained model before continual training and evaluate accuracy across candidate threshold values. We select τ at the elbow of the resulting accuracy-threshold curve, corresponding to the point beyond which increasing the threshold yields diminishing improvement. This calibration is performed once per model backbone. The resulting threshold is then fixed and used for SFT, AAT, ER, A-GEM, and EWC, ensuring that differences between training methods cannot arise from method-specific threshold tuning.

Compute. On the largest benchmark, NAB, a single training run with evaluation every 10 steps requires approximately 1 hour and 20 minutes on our hardware. Each run uses three NVIDIA RTX A6000 GPUs in parallel to achieve the target effective batch size. Runs are executed sequentially, and the full experimental suite required approximately 11 days of wall-clock training time.

Table 5: RCB performance for ER with a 100-example buffer using one versus five local updates (n), evaluated with seed 42. Online accuracy and forgetting are averaged across training steps; cumulative accuracy is measured over all observed examples.

Model	Method	$\text{Acc}^{\text{unknown}}_{\text{online}}$	$\text{Acc}^{\text{known}}_{\text{online}}$	$\text{F}^{\text{unknown}}$	F^{known}	$\text{Acc}^{\text{unknown}}_{\text{cumul}}$	$\text{Acc}^{\text{known}}_{\text{cumul}}$	$\text{Acc}^{\text{all}}_{\text{cumul}}$
Qwen-1.5B	Buffer-100-lp1	0.3772	0.5283	0.0950	0.0817	0.3450	0.5051	0.4551
	Buffer-100-lp5	0.4023	0.5656	0.0626	0.0635	0.3835	0.5515	0.4990

A.5 Narrative Expansion

To construct the NAB used in Section 5.3, we start from a collection of narratives paired with their corresponding proverbs from ePiC dataset (Ghosh & Srivastava, 2022) and expand the dataset by generating 15 additional narratives for each proverb. The original dataset contains 250 proverbs, with 10 narratives associated with each proverb. This expansion increases the number of narrative instances per proverb while preserving the original set of proverbs. For narrative expansion, we prompted Qwen2.5-7B-Instruct (Team, 2024). Prompt design is informed by principles from SMT (Gentner, 1983). For each proverb, an existing narrative is selected as a seed. Prompts discourage lexical overlap and object reuse, require a domain-level shift, and emphasize alignment at the level of abstract relational structure rather than surface form. In addition to generating narrative text, the model is prompted to identify aligned spans between the proverb and the narrative corresponding to shared relational elements. These span annotations are retained for analysis but are not used during training or evaluation. The full expansion prompt is shown in Figure 6.

A.6 Distractor Generation

In NAB we convert each narrative instance into a pairwise continuation comparison task. Each narrative is split into a shared narrative part and two alternative continuations: (i) a *correct ending* that is consistent with the proverb's abstract meaning, and (ii) a *distractor ending* that is locally coherent but violates that abstract relation. For each narrative, we use sentence-level segmentation to break down the narrative into a shared narrative part and the final sentence of the narrative, and consider the final sentence as the correct ending. For distractor generation, we prompted Qwen2.5-7B-Instruct (Team, 2024) with the proverb, the shared narrative part, and the correct ending, instructing the model to produce a single-sentence continuation that flows naturally from the narrative context but violates or ignores the proverb's abstract meaning. Distractors are constrained to be plausible story continuations and are matched in length to the correct ending, ensuring that superficial cues such as fluency or stylistic coherence are insufficient for solving the task. The model additionally produces a brief explanation describing why the distractor does not align with the proverb, which is retained for analysis but not shown to the model during training or evaluation. The full distractor generation prompt is shown in Figure 5. Each resulting instance defines a pairwise continuation comparison task in which the model is evaluated by comparing the log-likelihood assigned to each candidate ending given the same narrative part. Because both endings are locally coherent continuations of the same context, a higher likelihood therefore requires sensitivity to the implicit proverb-level abstraction rather than reliance on surface-level narrative consistency.

A.7 Data Statistics and Samples

We introduced the Relational Cycle Benchmark (RCB) and the Narrative Abstraction Benchmark (NAB) to study the role of abstraction in non-stationary online learning. Table 6 reports RCB statistics by domain, including the number of instances, the number and fraction for which a derivable relation is identified, and the resulting counts of known and unknown edges. When a derivable relation exists, it is held out as the unknown edge and the remaining relations are treated as known; instances without a derivable relation contribute only to known-edge evaluation. Representative samples from RCB and NAB are provided in Table 9 and Table 8.

We additionally report zero-shot performance for different sizes of Qwen2.5 and SmolLM-1.7B on both benchmarks in Table 7. On RCB, the 1.5B and 3B models perform similarly, while the 7B model shows a marked improvement, suggesting that larger models have substantially stronger access to the relational and factual knowledge required by the benchmark. On NAB, performance changes comparatively little across model sizes, with no clear scaling trend over the evaluated range. Comparing these zero-shot results with Table 3 shows that training improves performance on known edges, but yields much smaller gains on unknown relations. This suggests that exposure to the constituent entities and observed relations does not by itself ensure inference of the held-out relational dependency.

Table 6: Dataset statistics for RCB. Instances with an identifiable logical dependency contribute one derivable relation as an unknown edge; instances without such a dependency contribute only known edges.

Domain	**Instances**	**Derivable**	**% Deriv.**	**Known Edges**	**Unknown Edges**
Genealogy	500	200	40.0%	1000	200
Profession	180	180	100.0%	900	180
Arts	115	50	43.5%	345	50
Science	100	25	25.0%	200	25
History	100	100	100.0%	200	100
Music	100	25	25.0%	200	25
Geopolitics	75	25	33.3%	225	25
Sports	75	75	100.0%	225	75
Total	1245	680	54.6%	3295	680

Table 7: Zero-shot performance across model sizes on the RCB and the NAB.

Model	RCB			NAB		
	Unknown	Known	Total	Unknown	Known	Total
Qwen-1.5B	0.3469	0.4695	0.4311	0.5420	0.5103	0.5134
Qwen-3B	0.3020	0.4537	0.4062	0.5222	0.5266	0.5262
Qwen-7B	0.4357	0.6062	0.5528	0.5157	0.5198	0.5194
SmolLM-1.7B	0.4450	0.5934	0.5469	0.4711	0.4943	0.4920

Table 8: Random samples from the NAB: Proverbs, Narratives and Continuation Task.

Proverb	Narrative Context	Correct Ending	Distractor Ending
Jam tomorrow...	The man believed his friend when he kept borrowing money from him, telling him he would pay him tomorrow. Every day he told him he would pay him soon.	He kept giving him money, thinking of the future, but never saw the loans repaid.	He finally got fed up with the constant requests and decided to save all his money for emergencies.
Keep your powder dry	Margo always knew to start any trip with a spare tire, emergency cash, and a car repair kit. She stayed calm and prepared for any catastrophe.	She was yet to need any of this, but she always figured she'd be best prepared if catastrophe ever struck.	With all the supplies in her car, Margo felt ready for anything and drove off, never expecting to encounter issues.
Laughter is the best medicine	My grandfather was ill and in a somber mood. I read him a funny speech about a plagiarism scandal, and his eyes lit up with fascination.	The somber mood in the room dissipated like dew as we all wondered loudly, and mirthfully, how that was possible.	The somber mood in the room dissipated as my grandfather's eyes lit up with curiosity about the scandal.
Procrastination is thief of time	My son put off an oil change that would have taken 20 minutes. Six months later, he was sitting at the side of the road with a seized-up engine.	Ultimately, we figured the delay resulted in over 29 hours of combined waiting, extra travel, and car searching.	Fortunately, my son managed to get the engine fixed quickly by borrowing a friend's car, saving him from delays.
It takes two to tango	Jeff and Jane were getting into a fight. Jane pointed the finger, saying it was all Jeff's fault. Their mother knew better though.	She knew that they both were causing trouble and Jeff and Jane were both punished.	She quickly intervened and helped Jeff up off the floor, advising him to apologize to Jane.

Table 9: Random samples from the RCB: Text Context and Knowledge Graph Edges

Text Context	Known Edges (Sample)	Unknown Edge (Target)
Ljubljana is in Slovenia. Academy of Theatre, Radio, Film and Television is in Ljubljana. Polde Bibič works for the University of Ljubljana and was educated at the Academy.	(Polde Bibič, works for, Univ. of Ljubljana), (Polde Bibič, educated at, Acad. of Theatre...), (Ljubljana, country, Slovenia)	(Univ. of Ljubljana, based in country, Slovenia)
Closing time is a dance film and has Mette Ingvartsen as a cast member.	(Closing time, cast member, Mette Ingvartsen), (Closing time, genre, dance)	(Mette Ingvartsen, field of work, dance)

```
You are an expert dataset creator for testing Analogical Reasoning.
Your task is to generate a "Distractor Ending" for a story.

# GOAL
You will be given a Proverb, a Story Context, and the Correct Ending.
You must write a single sentence that:
1.  Flows naturally from the Story Context.
2.  Violates or Ignores the meaning of the Proverb.

# CRITERIA FOR THE DISTRACTOR
- Plausible Storytelling: It should look like a real ending. Do not write gibberish.
- Thematic Mismatch: If the proverb says "haste makes waste" (bad outcome),
your distractor should perhaps show haste leading to success.
- Length: Keep it roughly the same length as the Correct Ending.

# OUTPUT FORMAT
Strict JSON only:
{
    "distractor_ending": "[Your new sentence here]",
    "explanation": "[Briefly explain why this ending does not fit the proverb]"
}
```

Figure 5: Distractor generation prompt used for NAB.

```
You are an expert Cognitive Scientist and NLP Researcher in Analogical
Reasoning. Your task is to expand the ePiC dataset (Proverbs in Context)
by applying Gentner's "Structure-Mapping Theory".

# TASK OVERVIEW
Generate a NEW narrative that is a structural analog to the proverb.
Move from "Instance Learning" to "Abstraction" by focusing on
higher-order relations rather than surface-level similarities.

# EXECUTION STEPS
1. Relational Extraction: Identify the core relational predicates.
2. Pragmatic Goal: Identify the "Purpose" (warn, comfort, criticize).
3. Domain Leap: Select a target domain different from the seed.
4. Systematic Alignment: Draft a narrative where the causal pattern
   allows the relations to map to the proverb.
5. Evaluation: Review for "Object Overlap."
6. Span Discovery: Identify structural counterparts.

# CONSTRAINTS
- No Surface Overlap: No shared objects/themes with the proverb.
- Systematicity: Narrative must explain the "why" (causal chains).
- Span Alignment: Align up to 5 meaningful structural components.

# OUTPUT FORMAT (Strict JSON)
{
    "analysis": {
        "abstract_schema": "...",
        "domain_leap": "..."
    },
    "fields": {
        "quote": "{{quote_placeholder}}",
        "narrative": "...",
        "span_quote_1": "", "span_narrative_1": "",
        "span_quote_2": "", "span_narrative_2": "",
        "span_quote_3": "", "span_narrative_3": "",
        "span_quote_4": "", "span_narrative_4": "",
        "span_quote_5": "", "span_narrative_5": ""
    },
    "judge_scores": {
        "relational_similarity": "1-10",
        "surface_dissimilarity": "1-10",
        "rationale": "..."
    }
}
```

Figure 6: Narrative expansion prompt used for NAB.