Same-Head Analogical Completion: A Shortcut-Controlled Diagnostic Suite for Source-Conditioned Relation Completion

Anonymous ACL submission

Abstract

Analogy-style relation evaluations often conflate source-relation use with target-head plausibility, entity similarity, prompt format, and candidate construction. We introduce *same-head analogical completion*, a shortcut-controlled diagnostic suite that makes these explanations separately measurable: given a source pair (A, B) and target anchor C, a model ranks candidate tails D_i that are all attested for the same C under different relations. On ANALOGYKB, a Wikidata-derived analogy benchmark, the suite supports three claims. First, same-head fields expose source use: target-only scoring stays near chance, while source-aware rankers are higher and collapse under source shuffling. Second, candidate construction is itself an evaluation axis: tail-similarity matching and hard-negative selector stress preserve the source-use signal but change model-family rankings. Third, a supervised target-space-stabilized predictive bottleneck becomes useful under held-out target heads, but stronger RelBERT, RoBERTa, and GPT-4.1-mini diagnostics bound the practical range and held-out-relation transfer changes the ordering. The contribution is therefore not a universal ranker, but a reproducible protocol for separating what relation models use under controlled same-head completion.

1 Introduction

Human analogy often concerns relations rather than isolated entities. Given *France:French*, a target anchor such as *Brazil* suggests *Portuguese*; given *Marie Curie:Nobel Prize in Physics*, another anchor may call for the award, office, place, or family relation that best matches the source pair. The output is not simply the most likely next token. It is a relation-conditioned completion: infer the relation expressed by (A, B), then select a target pair (C, D) that realizes that relation.

We study a diagnostic suite centered on the controlled candidate field. The suite asks what model families use when target-head plausibility, source-tail similarity, source-conditioned relation matching, selector-specific candidate hardness, and specialized relation-pair representations are experimentally separated. Same-head completion turns these explanations into benchmark axes. Predictive bottlenecks are one probe in the suite. A Joint-Embedding Predictive Architecture (JEPA)-style model (Assran et al., 2023) predicts a latent target relation state and compares it to candidate target-pair embeddings; the diagnostic question is when that bottleneck helps once shortcut controls and strong direct rankers are already in place.

The central challenge is shortcut control, a recurring issue in benchmark design and model diagnostics (Gururangan et al., 2018; McCoy et al., 2019; Geirhos et al., 2020). If candidates differ in target-head plausibility, a source-agnostic model (one that ignores (A, B)) can win. If candidate labels are exposed, the task becomes relation classification. If the predictor sees every candidate jointly with the source pair, it may behave like a direct joint multilayer perceptron (MLP) ranker rather than a predictive bottleneck. We therefore study a controlled candidate-ranking task:

$$\arg\max_i s_i, \quad s_i \approx \mathrm{sim}(r(A, B), r(C, D_i)) \,. \quad (1)$$

In our same-head analogical completion setting, every D_i is an attested tail for the same C, but under different relations. Thus C-only plausibility is insufficient; the model must use (A, B) to determine which relation around C is intended.

We build the task from ANALOGYKB, a benchmark of analogy relation pairs (Yuan et al., 2024), focusing on Wikidata because it offers enough multi-relation target heads for 8-way same-head ranking. The benchmark's main claim is methodological: same-head candidate fields make source-conditioned completion, target-side shortcuts, selector choice, and generalization regime separately

observable. The main generalization setting is held-out target heads with a seen relation inventory. The expanded 669-relation held-out diagnostic adds a complementary axis: source use can persist even when unseen-relation model ordering changes.

We use a small set of terms consistently. A *source-aware* model receives (A, B), whereas a *source-agnostic* baseline scores candidates from the target side only. *Source shuffling* replaces (A, B) during evaluation to test whether a model's score really depends on the source pair. The *candidate field* is the set of candidate tails attached to the same C. A *hard-negative selector* is the frozen embedding model used to choose difficult non-gold tails during dataset construction; it is separate from the model being evaluated. *Tail-similarity matching* means that non-gold tails are selected to be close to the source tail B, making simple entity-similarity heuristics less decisive. *Target-space stabilization* means pretraining or regularizing the target-pair embedding space before bottleneck scoring; the strongest version uses a supervised target-contrastive loss, which pulls gold target pairs with the same relation label closer together during training.

Empirically, the study is organized around three claims that together support the diagnostic protocol. First, same-head fields make source use measurable: target-only scoring is near chance, source-aware rankers are higher, and source shuffling drops them back into the chance range. Second, candidate construction is part of the benchmark rather than a preprocessing detail: tail-similarity matching exposes a strong entity-similarity shortcut, and changing only the hard-negative selector preserves source use while changing architecture rankings. Third, the bottleneck result is a bounded mechanism claim: target-space stabilization enables source-conditioned projection under held-out target heads, while stronger relation encoders, source-only ablations, and held-out-relation tests define where that mechanism does and does not explain performance.

Contributions. We contribute: (i) same-head analogical completion, a shortcut-controlled diagnostic for source-conditioned relation completion; (ii) a set of controlled diagnostics that separate source use, entity-similarity pressure, selector dependence, held-out-head transfer, and held-out-relation transfer; and (iii) a bounded mechanism study of when a JEPA-style predictive bottleneck helps once matched direct rankers and practical relation encoders are in the same candidate fields. The release includes exact split settings, data loaders, experiment runners, bootstrap and diagnostic scripts, prompt templates, and machine-readable JSON/CSV result files; full prediction dumps are treated as artifact-archive material rather than prose supplement material.

Scope. The scope is deliberately controlled: we study source-conditioned ranking over attested candidate tails, not open-ended relation discovery, label-free relation learning, broad unseen-relation transfer, broad analogy solving, or general relation matching. This controlled scope is what makes the benchmark diagnostic. It lets us test whether a model uses the source relation under a candidate construction where target-head plausibility alone is weak, and it turns selector dependence and held-out-relation behavior into interpretable benchmark outputs rather than incidental artifacts.

2 Related Work

Analogy is often studied as structure mapping rather than surface similarity (Gentner, 1983; Gentner and Markman, 1997; Turney, 2006; Jurgens et al., 2012), with natural language processing (NLP) benchmarks covering lexical, multimodal, and long-context analogies (Mikolov et al., 2013; Gladkova et al., 2016; Zhang et al., 2023; Ye et al., 2024). ANALOGYKB supplies the relation pairs (Yuan et al., 2024); our contribution is the same-head completion control, following diagnostic benchmark work that exposes artifacts rather than relying on aggregate accuracy (Gururangan et al., 2018; McCoy et al., 2019; Geirhos et al., 2020). Knowledge graph (KG) embedding and text-conditioned knowledge graph completion (KGC) methods rank or verbalize known-relation triples (Bordes et al., 2013; Yang et al., 2015; Trouillon et al., 2016; Sun et al., 2019; Yao et al., 2019; Chen et al., 2022; Yao et al., 2025); our task withholds the target relation identifier. Textual relation encoders are therefore strong practical baselines (Soares et al., 2019; Rossiello et al., 2019; Mtumbuka and Schockaert, 2024; Ushio et al., 2025). JEPA models predict in representation space (Assran et al., 2023; Terver et al., 2026); we identify when that bottleneck helps same-head completion under controlled candidate fields.

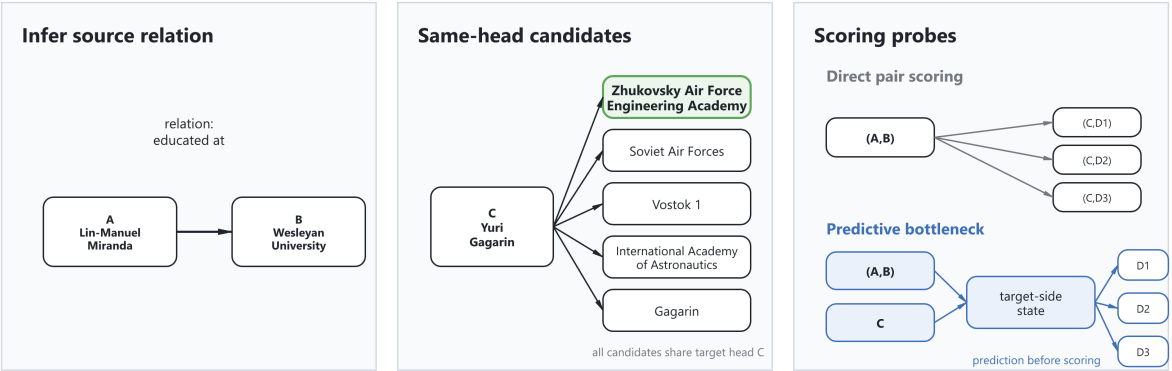

Figure 1: Controlled same-head analogical completion and one bottleneck probe. The task first requires inferring the relation expressed by the source pair, then applying it around a target head whose candidate tails are all attested for that same head. The JEPA-style probe predicts a target-side relation state before candidate scoring, whereas direct pair-scoring baselines compare each candidate pair to the source pair.

3 Task

Each example contains a source pair (A, B), a target head C, a gold tail D^+, and $K - 1$ negative tails D^-, where K is the candidate-set size. The model ranks candidate tails:

$$s(A, B, C, D^+) > s(A, B, C, D^-). \qquad (2)$$

We evaluate top-1 accuracy (Acc@1), mean reciprocal rank (MRR), recall at three (R@3), and source-shuffled Acc@1, which is Acc@1 after the source pair has been replaced.

Same-head control. For every test instance, all candidates are attested tails for the same target head C, but under different relations. This differs from random or nearest-neighbor negative sampling, which can amplify target-side priors. Same-head control makes the target anchor ambiguous by design. Figure 1 summarizes why this control is central: the target head stays fixed, so the model must use the source pair to decide which relation around that target head is intended.

Pair split. Train and test contain different pairs from overlapping relation inventory and target-head distribution. This measures in-distribution relation completion.

Held-out-head split. Target heads C in test are removed from training before constructing examples. This is the key generalization diagnostic: the model must apply a learned source relation to a new target anchor and candidate field.

Source shuffle. At evaluation time, we replace the source pair with another source pair in the batch. A source-conditioned model should drop toward the chance range.

Diagnostic interpretation. We do not treat high accuracy alone as evidence of analogical relation use. A positive diagnostic result must be above the source-agnostic baseline, degrade when the source pair is shuffled, and remain interpretable under stricter candidate construction. Architecture comparisons are therefore always read together with the candidate field on which they are measured.

4 Models

Source-agnostic. Scores D_i from C, D_i only. This tests whether the candidate set leaks target-side shortcuts.

Frozen embedding baselines. We include non-parametric frozen baselines with no task training: tail cosine $\cos(B, D_i)$, offset cosine $\cos(B - A, D_i - C)$, target-head cosine $\cos(C, D_i)$, and source-head cosine $\cos(A, D_i)$. We report these with frozen BERT (Devlin et al., 2019) and, for the hardest setting, with an off-the-shelf Sentence-BERT-style MiniLM encoder (Reimers and Gurevych, 2019). We also include a frozen RelBERT pair-similarity baseline (Ushio et al., 2025), which directly embeds (A, B) and (C, D_i) as relation pairs and ranks by cosine similarity. These baselines test whether the benchmark can be solved by entity or relation embedding proximity rather than learned relation transfer. Standard transductive knowledge graph embedding (KGE) link pre-

Table 1: Model families. We use descriptive names in the text and reserve implementation names for the reproducibility appendix.

Family	What it tests	Scoring form	Role in the study
Target-only ranker	Can the candidate field be solved without the source pair?	$s(C, D_i)$	Shortcut check
Frozen similarity	Do off-the-shelf entity or relation embeddings already solve the task?	$\cos(B, D_i)$, offset cosine, RelBERT pair cosine	Non-parametric baselines
Siamese pair encoder	Is direct relation-pair matching enough without prediction?	$\cos(r(A, B), r(C, D_i))$	Strong learned baseline
Joint MLP ranker	How strong is a model that scores every candidate jointly?	$f(A, B, C, D_i)$	Non-bottleneck upper baseline
Predictive bottleneck	Can JEPA infer a source relation and predict a target-side relation state before seeing candidate scores?	$\cos(P(r(A, B), C), r(C, D_i))$	Main architecture

dictors (Bordes et al., 2013; Yang et al., 2015; Trouillon et al., 2016; Sun et al., 2019) require known r and learned entity embeddings, so held-out-head source-conditioned completion would either hide needed entities or give oracle relation access; we instead report text-inductive baselines plus an oracle prototype diagnostic.

Seq2seq likelihood baseline. To compare against a local instruction-tuned generator, we prompt FLAN-T5-base (Chung et al., 2024) with the source pair, target head, and candidate list, then rank candidates by length-normalized conditional likelihood. We also run a retrieval-augmented variant that prepends two nearest training demonstrations selected by frozen source-pair similarity. This tests whether a small instruction-tuned seq2seq model can solve the controlled candidate ranking directly from text likelihood, with or without simple in-context retrieval. Separate full three-seed API diagnostics with GPT-4.1-mini use two output formats: a single-choice prompt that reports Acc@1 only, and a complete-ranking prompt that reports Acc@1, MRR, and R@3. We treat these as prompted API diagnostics rather than calibrated LLM comparisons because the output format changes behavior.

Textual cross-encoder baseline. We also fine-tune a modest RoBERTa-base (Liu et al., 2019) cross-encoder on raw entity-string pairs, scoring each candidate from "A [REL] B" and "C [REL] D_i". This is not capacity-matched to the frozen-feature MLPs; it bounds stronger trainable textual joint scoring.

Siamese pair encoder. Encodes $r(A, B)$ and $r(C, D_i)$ independently with shared-form relation encoders and scores by similarity. This is a Siamese relation-matching baseline related to prior analogy representation work (Rossiello et al., 2019).

Joint MLP ranker. Scores each candidate from joint features of A, B, C, D_i. This is the strongest non-bottleneck baseline; it can condition on each candidate jointly with the source pair.

Predictive bottleneck. The predictive model separates source relation inference from target candidate scoring:

$$z_{src} = E_s(A, B), \quad \hat{z} = P(z_{src}, C),$$
$$z_i = E_t(C, D_i), \quad s_i = \cos(\hat{z}, z_i).$$

The predictor produces a target-side relation state before candidate similarity scoring. The default form includes C, but we treat explicit predictor-side anchor conditioning as a testable design choice rather than a required part of the claim: a source-only predictor ablation removes C from P while leaving each target-pair embedding $E_t(C, D_i)$ and the same-head candidate field unchanged.

Target-space stabilization. Naive JEPA learns target-pair representations and the predictor simultaneously and performs poorly. We therefore pretrain E_s, E_t with pair matching, then freeze E_t, or freeze both E_s and E_t, before predictive training.

Relation-aware auxiliary pressure. During training only, we add an auxiliary relation classification loss on source and gold target pair embeddings. Relation labels are not used at evaluation. This encourages the target space to remain relation-discriminative while preserving candidate ranking as the task.

Target-contrastive relation space. The final repair adds a supervised contrastive loss over gold target-pair embeddings, following contrastive representation learning and supervised contrastive objectives (van den Oord et al., 2018; Chen et al., 2020; Khosla et al., 2020). For a minibatch of gold target representations $z_j = E_t(C_j, D_j^+)$ with relation labels y_j, positives for j are other examples with the same exact ANALOGYKB relation identifier. We do not merge aliases, inverses, or near-duplicate compositional relations when forming positives; this keeps the supervision reproducible but makes the label-granularity limitation explicit. Let $\alpha_{jk} = \exp(\cos(z_j, z_k)/\tau)$. We minimize

$$\mathcal{L}_{tc}(j) = -\frac{1}{|P(j)|} \sum_{p \in P(j)} \log \frac{\alpha_{jp}}{\sum_{k \neq j} \alpha_{jk}}.$$

This does not use relation labels for test-time scoring and does not replace candidate ranking with relation classification. It only shapes the target relation space into which the predictive bottleneck must map.

5 Experimental Setup

We use ANALOGYKB Wikidata same-relation pairs (Yuan et al., 2024). We preserve ANALOGYKB's relation identifiers as labels; if an entry contains a relation-name list, we join the listed names into one label, but we do not manually merge aliases, inverses, or compositional near-duplicates. Under the current cap of 300 relations and 500 pairs per relation, the loader obtains 101,330 pairs and 750 eligible target heads with at least eight attested tails across multiple relations.

Candidate fields and splits. The main candidate count is 8. Same-head fields are constructed only from tails attested for the same C under different relations. The tail-matched variants further choose non-gold tails whose frozen-selector similarity to the source tail B is close to or higher than the gold tail, making simple B-to-D_i proximity less decisive. Selector-stress experiments keep the model family, feature backbone, and split protocol fixed while changing only the frozen embedding model used to construct the hard-negative field. In the tail-matched held-out-head setting, train/test sizes range from 4,180–4,240 and 3,240–3,520 examples, with 290–294 test relations after coverage filtering.

Model selection and reporting. To avoid selecting target-contrastive hyperparameters on the test split, we use a deterministic train/dev/test held-out-head protocol and select the target-contrastive weight and temperature by three-seed mean dev Acc@1 before reporting the held-out test comparison. A package audit verifies zero exact train/dev/test fact overlap and zero target-head overlap for these splits. We also retain a five-seed seed-group confirmation audit for the earlier $(0.2, 0.2)$ setting, where seeds 3–4 are checked after the initial seeds 0–2 sweep. Unless noted otherwise, reported learned-model means use three seeds and matched model comparisons use the same prediction instances and candidate fields; appendix intervals use paired bootstrap, with relation- and target-head-cluster bootstraps for the main held-out-head comparison. Text features are frozen `bert-base-uncased` embeddings of entity strings only; we do not use descriptions, aliases, or document context. All main learned-model runs use CUDA-enabled GPU acceleration. Qualitative examples are explicitly seed-0 diagnostics.

6 Results

We report the main diagnostics in the order needed to isolate competing explanations. The first two tables test whether same-head fields remove target-only plausibility and expose the remaining entity-similarity shortcut. The selector-stress table then changes only the hard-negative field and measures whether source use and model ordering move together. The held-out-head table is the central mechanism result: it places the stabilized bottleneck, matched learned rankers, and stronger practical anchors in the same candidate field. The appendix contains the extended ablations and artifact map.

6.1 Same-Head Control Exposes Source Use and Shortcuts

Table 2 gives the first diagnostic result. Same-head completion is not solved by the target field alone: the source-agnostic model is near chance, while source-aware learned models are much higher. Frozen tail cosine is also strong, showing that many analogical completions are partly aligned with semantic proximity between the source tail B and a candidate tail D_i. That result motivates the stricter tail-matched diagnostic: tail cosine ignores A, C, and the source relation, and its rank quality is below joint MLP scoring.

Table 2: Wikidata same-head pair split with 8 candidates per query and three seeds; train and test draw pairs from the same loaded relations. Chance Acc@1 is 0.125. Source-aware learned models are far above source-agnostic; the strong frozen tail-similarity baseline motivates the stricter tail-matched diagnostic.

Model	Acc@1	MRR	R@3
Joint MLP ranker	0.3517	0.5632	0.7049
Frozen tail cosine	0.3475	0.5278	0.6038
Frozen offset cosine	0.3048	0.4967	0.5721
Siamese pair encoder	0.3172	0.5439	0.7027
Naive predictive bottleneck	0.2839	0.5099	0.6357
Source-agnostic	0.1364	0.3482	0.3777

Table 3: Three-seed Wikidata pair split with 8 same-head candidates additionally matched against source-tail similarity. Frozen tail cosine remains strong on Acc@1, but learned source-conditioned models have much stronger rank quality.

Model	Acc@1	MRR	R@3
Frozen tail cosine	0.3379	0.5075	0.5615
Relation-shaped predictive bottleneck	0.3117	0.5362	0.6933
Joint MLP ranker	0.2997	0.5334	0.6946
Siamese pair encoder	0.2885	0.5113	0.6337
Source-agnostic	0.1343	0.3470	0.3858

6.2 Tail Matching and Selector Choice Are Benchmark Axes

Because frozen tail cosine is strong, we add a stricter diagnostic in which same-head negative tails are chosen to have source-tail similarity close to or higher than the gold tail. The selector stress suite uses off-the-shelf sentence-embedding models–BERT, MiniLM, GTE-small (a General Text Embeddings model), and E5-small-v2 (a contrastively trained text-embedding model)–only to construct candidate fields, not as learned rankers. Table 3 shows that tail matching changes the comparison: the relation-shaped predictive bottleneck and the joint MLP become statistically tied under paired bootstrap, with bottleneck-minus-joint MLP Acc@1 $+0.0137$, 95% confidence interval (CI) $[-0.0115, 0.0390]$, MRR $+0.0034$, CI $[-0.0132, 0.0201]$, and R@3 -0.0066, CI $[-0.0308, 0.0176]$. Table 4 formalizes selector stress as two separate diagnostics. The source-use diagnostic is stable: source-aware models collapse under source shuffling across selectors. The architecture-ranking diagnostic is selector-dependent: the JEPA-over-Joint-MLP Acc@1 delta is positive with BERT and MiniLM hard-negative selectors, but reverses with GTE-small and E5-small-v2. Because the learned feature encoder is held fixed while only the candidate field changes, these sign flips are a methodological finding: hard-negative construction is part of the evaluation axis, not a hidden preprocessing detail.

Figure 2 clarifies the shared conclusion across selectors. GTE/E5-selected negatives create a field where direct candidate-wise conditioning and pair matching are more useful than a single predicted target state, plausibly because the selector produces a different hard-negative distribution than the BERT embedding feature space used by the learned rankers. Selector stress therefore becomes one of the suite's outputs: it separates source-use robustness from architecture-ranking robustness and specifies the candidate-field conditions under which the bottleneck result appears.

6.3 Held-Out Heads Separate Model Families

Held-out target heads show why source-use controls and model-quality diagnostics must be separated. In the pre-contrastive held-out-head diagnostic, the relation-shaped bottleneck remains source-aware: source shuffling drops it into the chance range. The joint MLP and frozen tail cosine are nevertheless stronger on raw ranking quality (Appendix Table 8). Thus passing source-shuffle control is necessary but not sufficient; the suite distinguishes whether a model uses the source pair from how well it transfers that signal to a new target-head field.

The central held-out-head diagnostic in Table 5 separates practical strength from diagnostic behavior. Same-head candidates are not solved by target-only scoring, and entity similarity remains a real pressure. The stabilized bottleneck is above the matched Joint MLP and Siamese pair encoder, while source shuffling collapses source-aware learned models. RelBERT, RoBERTa, and GPT-4.1-mini calibrate the stronger practical range under the same candidate field, and GPT-4.1-mini illustrates prompt-format sensitivity: single-choice and complete-ranking prompts differ by about five Acc@1 points (Appendix Table 11).

The source-only bottleneck clarifies the mechanism: the useful operation is source-conditioned projection into a stabilized target-pair space, not explicit C conditioning inside P. A second cross-cutting stress test is held-out relations. In the expanded 669-relation diagnostic, the predictive bottleneck is far above target-only scoring (Acc@1 $+0.1731$) but below the joint MLP (-0.0293) and Siamese encoder (-0.0160; Appendix Table 17).

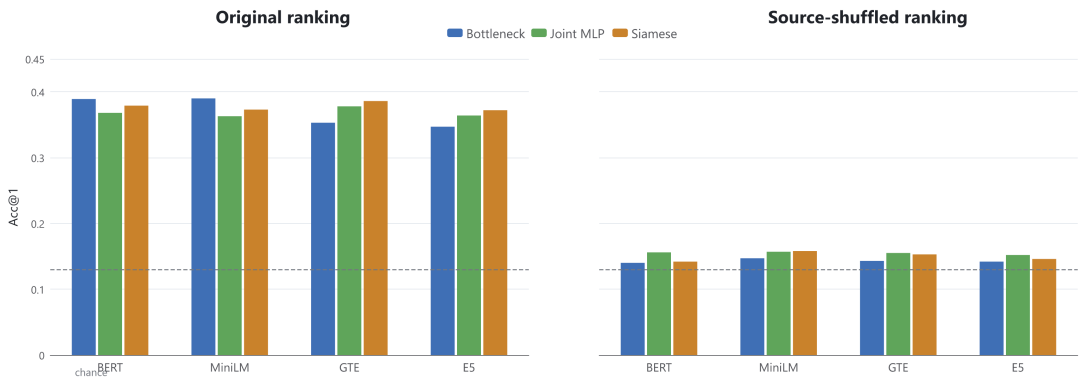

Figure 2: Hard-negative selector stress on tail-matched held-out heads. The bottleneck gain over direct rankers is positive with BERT and MiniLM candidate construction but reverses with GTE-small and E5-small-v2. The source-shuffled panel shows the stable diagnostic: all source-aware models collapse toward chance across selectors, so source use is robust while model ordering tracks candidate construction.

Table 4: Selector stress as a benchmark axis on tail-matched held-out heads. Δ Acc@1 is predictive bottleneck minus Joint MLP. Across the suite, source-shuffled Acc@1 for source-aware models is roughly 0.142–0.158, so source use is robust; architecture ranking tracks the hard-negative selector.

Selector	Δ Acc@1	Source-use diagnostic	Architecture-ranking diagnostic
BERT	$+0.0213$	Stable source-use signal: source shuffle near chance	Bottleneck > Joint MLP under this candidate field
MiniLM	$+0.0266$	Stable source-use signal: source shuffle near chance	Bottleneck > Joint MLP under this candidate field
GTE-small	-0.0243	Stable source-use signal: source shuffle near chance	Joint MLP > bottleneck under this candidate field
E5-small-v2	-0.0168	Stable source-use signal: source shuffle near chance	Joint MLP > bottleneck under this candidate field

Table 5: Core held-out-head diagnostic on 8-way tail-matched same-head fields. This compact main-text view gives the practical range and the matched bottleneck comparison; Appendix Table 11 gives the full notes and auxiliary rows.

Model	Acc@1	MRR	R@3	Shuffle Acc.
GPT-4.1-mini single choice	0.4659	–	–	–
Frozen RelBERT pair cosine	0.4474	0.6233	0.7340	–
Fine-tuned RoBERTa cross-encoder	0.4253	0.6110	0.7302	–
Target-contrastive bottleneck	0.3888	0.5912	0.7297	0.1399
Source-only bottleneck	0.3967	0.5960	0.7340	0.1430
Joint MLP ranker	0.3677	0.5726	0.7052	0.1559
Siamese pair encoder	0.3791	0.5821	0.7223	0.1423
Frozen tail cosine	0.3175	0.4919	0.5457	–
Source-agnostic	0.2137	0.4187	0.4744	0.2137

Same-head control therefore exposes source use while also showing that unseen-relation transfer is a distinct generalization regime.

RelBERT-space diagnostics ask how the bottleneck behaves inside a stronger relation-pair representation. Stronger target spaces improve bottleneck rank quality over frozen RelBERT in MRR and R@3, and candidate-wise direct adaptation is stronger still (Appendix Table 16). RelBERT therefore serves as a controlled diagnostic: source-conditioned bottleneck scores add non-redundant signal, and direct candidate-wise adaptation quantifies the remaining bottleneck cost.

6.4 Bottleneck Probe: When Stabilization Helps

The bottleneck probe identifies the representation condition under which latent prediction becomes useful. On the pair split, target-pair pretraining, encoder freezing, and relation-aware auxiliary pressure improve the predictive bottleneck but leave a gap to joint MLP scoring (Appendix Table 9). On held-out heads, the same pre-contrastive bottleneck is substantially below the joint MLP: Acc@1 -0.0738, MRR -0.0564, and R@3 -0.0566, with paired CIs fully below zero.

The effective repair is representational rather than a scoring-layer tweak. Field centering improves the pre-contrastive bottleneck, and mean/variance adaptation adds a small MRR gain,

but the decisive step is supervised target-contrastive shaping of the target-pair space. A dev-selected search chooses weight 0.4 and temperature 0.2, with nearby nonzero settings also strong and the no-contrastive row much weaker (Appendix Figure 3). The resulting bottleneck reaches Acc@1 0.3888, MRR 0.5912, and R@3 0.7297 on the tail-matched held-out-head split, compared with the joint MLP's 0.3677, 0.5726, and 0.7052. The gain is positive under example-, relation-, and target-head-level resampling (Appendix Table 12).

These diagnostics specify the mechanism. The same contrastive target-space loss applied to a Siamese pair encoder is close, and cluster intervals for that comparison are weaker. The predicted-state margin improves after stabilization: mean margin moves from -0.2801 to -0.0884, and median margin from -0.2374 to -0.0695 (Appendix Table 13). The median remains negative, so the target-pair space is better aligned but not cleanly separated. The source-only predictor is the clearest interpretation check. Removing C from P, while keeping C inside each target-pair embedding and candidate field, reaches $0.3967/0.5960/0.7340$, slightly above the C-conditioned form, and still collapses under source shuffling. Thus the mechanism is source-conditioned projection into a stabilized target-pair space; explicit predictor-side anchor conditioning is not the driver. At 16 candidates, the bottleneck still exceeds the joint MLP by Acc@1 $+0.0232$, MRR $+0.0211$, and R@3 $+0.0277$, showing that the diagnostic signal persists as the same-head field grows.

7 Discussion

The main contribution is an evaluation decomposition. Same-head control makes target-only shortcuts, entity similarity, source conditioning, selector choice, and target-space bottleneck behavior separately measurable. Source shuffling tests use of (A, B), tail matching tests source-tail similarity, selector stress tests candidate construction, and held-out-relation transfer tests extension beyond the seen relation inventory. The selector result is therefore a benchmark finding in its own right: source use is stable under source shuffling across the tested selectors, while model ordering changes with the hard-negative field.

This decomposition matters because a single analogy accuracy number can hide different failure modes. A model may use the source pair but fail under harder candidate fields, or it may rank well by exploiting relation-pair priors without robust source-conditioned completion. Same-head completion turns these possibilities into separate measurements rather than post-hoc explanations.

The resulting recommendation is procedural. Relation-completion benchmarks should report the candidate-field construction, include target-only and source-shuffled controls, and compare bottleneck and direct rankers on identical candidate fields before interpreting an accuracy gain as relation transfer. In our setting, these controls change the conclusion: the bottleneck is a useful mechanism under stabilized held-out-head same-head fields, but it is not the best practical ranker and it does not dominate across selector or relation generalization regimes.

The bottleneck probe supplies a mechanism case study inside that suite. Target-space stabilization changes predictive JEPA from a weak probe into a source-conditioned projection mechanism; the source-only ablation shows that the useful operation is projection into a stabilized target-pair space, not predictor-side target-anchor input. RelBERT, RoBERTa, and GPT-4.1-mini calibrate the practical range, while RelBERT-space adapters, margins, cluster bootstraps, ConceptNet, $K = 16$, frozen-backbone checks, ambiguity audits, and prototype/memory baselines specify where the mechanism helps and where direct rankers or stronger representations change the answer.

8 Conclusion

Controlled same-head analogical completion decomposes relation-completion accuracy into source use, candidate hardness, selector dependence, external-baseline strength, and held-out-relation behavior. The central contribution is a reproducible diagnostic suite for separating what relation models use under same-head completion. Within that suite, target-space stabilization yields a clear bottleneck mechanism: source-conditioned projection helps under specific candidate fields, while selector stress, RelBERT/RoBERTa/GPT diagnostics, and held-out relations define the conditions under which that mechanism explains performance. The result is not a universal ranker, but a benchmark suite that makes shortcut explanations testable.

Limitations

The benchmark ranks attested Wikidata candidates, not open-ended discoveries. We add relation- and target-head-cluster bootstraps for the main matched comparison, but not two-way or mixed-effects uncertainty for every diagnostic. The ambiguity audit is automatic or fixed-rubric; the fixed-rubric CSV is a bounded data audit, not completed external annotation evidence. The best bottleneck uses supervised relation labels during training, though never for test-time scoring. Its positive result is specific to tail-matched held-out heads under the tested selector fields; stronger RelBERT/RoBERTa/GPT-4.1-mini diagnostics, GTE/E5 selector stress, and held-out-relation transfer define the current boundary. Richer entity text, label-efficient stabilization, broader hard negatives, completed human adjudication, relation-superfamily sensitivity, two-way cluster uncertainty, and broader multi-domain validation remain open.

Ethics Statement

This work uses an existing public benchmark derived from Wikidata-style factual relation pairs. We do not collect human-subject data, annotate private text, or infer sensitive attributes. The main risk is misuse: ranking attested relation candidates should not be interpreted as discovering unknown facts or validating Wikidata statements, and downstream use should require provenance checks and human validation.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26.

Chen Chen, Yufei Wang, Bing Li, and Kwok-Yan Lam. 2022. Knowledge is flat: A Seq2Seq generative framework for various knowledge graph completion. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4005–4017, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

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

Dedre Gentner and Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of*

the First Joint Conference on Lexical and Computational Semantics, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Frank Martin Mtumbuka and Steven Schockaert. 2024. Entity or relation embeddings? an analysis of encoding strategies for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6003–6022, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Gaetano Rossiello, Alfio Gliozzo, Robert Farrell, Nicolas Fauceglia, and Michael Glass. 2019. Learning relational representations by analogy using hierarchical Siamese networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3235–3245, Minneapolis, Minnesota. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Basile Terver, Randall Balestriero, Megi Dervishi, David Fan, Quentin Garrido, Tushar Nagarajan, Koustuv Sinha, Wancong Zhang, Mike Rabbat, Yann LeCun, and Amir Bar. 2026. A lightweight library for energy-based joint-embedding predictive architectures. *arXiv preprint arXiv:2602.03604*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2025. RelBERT: Embedding relations with language models. *Artificial Intelligence*, 347(C):104359.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2025. Exploring large language models for knowledge graph completion. In *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. AnaloBench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13060–13082, Miami, Florida, USA. Association for Computational Linguistics.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1249–1265, Bangkok, Thailand. Association for Computational Linguistics.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. 2023. Multimodal analogical reasoning over knowledge graphs. In *International Conference on Learning Representations*.

A Supplementary Tables and Figures

Reader map. Table 6 links the main diagnostic claims to the appendix evidence and package artifacts.

Table 6: Appendix reader map. The table links each diagnostic claim to its main-text summary and detailed appendix evidence.

Claim	Main-paper summary	Detailed evidence
Same-head completion exposes source use and shortcuts	Tables 2–3; Sections 6.1–6.2	Figure 8; source-shuffle Figure 6; diagnostic matrix Table 11
Selector choice is an evaluation axis	Figure 2; Table 4; Section 6.2	Hard-negative selector stress and anti-similarity sanity checks in Section D
Stabilization identifies the bottleneck mechanism	Sections 6.3–6.4	Tables 9, 11, 12, and 13; Figures 3–5
External baselines bound practical performance	Section 6.3	Detailed matrix Table 11; RelBERT adapter Table 16; disagreement Figure 7
Held-out relations are a separate generalization axis	Sections 6.3 and 7	Held-out-relation Table 17 and limitations
Resource, size, and backbone checks support scope	Discussion	ConceptNet Table 10; backbone Table 18; candidate-size Table 20

Figure 3: Target-contrastive hyperparameter grid. Hyperparameters are selected by dev Acc@1 before test reporting; the no-contrastive row is much weaker, while several nonzero settings are close. This supports a stabilization effect rather than a single test-tuned point.

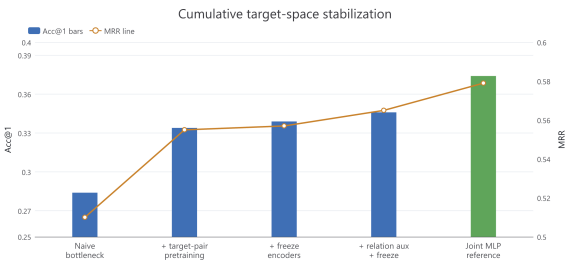

Figure 4: Cumulative effect of target-space stabilization on the three-seed Wikidata pair split with 8 candidates. Bars are Acc@1 and the line is MRR. Pretraining, encoder freezing, and relation-aware auxiliary pressure each lift the predictive bottleneck, but a gap to the non-bottleneck Joint MLP reference remains. The supervised target-contrastive held-out-head result is reported separately in Section 6.3. Visualizes Table 9.

B Reproducibility Notes

The implementation artifacts are the ANALOGYKB loader, completion and baseline runners, paired-bootstrap and prediction-analysis scripts, and per-relation summaries. The main dev-selected target-contrastive runs use frozen `bert-base-uncased` features, CUDA, seeds 0–2, 300 Wikidata relations, at most 500 pairs per relation, 8-way same-head tail-similarity-matched candidate ranking, a train/dev/test held-out-head split, 400 pair-pretraining steps, 200 predictive training steps, source/target freezing after pretraining, relation auxiliary weight 0.1, target-contrastive weight 0.4, and temperature 0.2. The earlier $(0.2, 0.2)$ setting is retained for five-seed confirmation, candidate-size, label-exposure, and several architecture diagnostics. External baselines include frozen RelBERT pair cosine, frozen MiniLM/BERT entity similarity, and local FLAN-T5-base length-normalized likelihood. Local release-environment checks used an NVIDIA GeForce RTX 3060 Ti GPU with 8GB memory and driver 560.94; the paper does not make wall-clock runtime claims because runtime depends on GPU availability, cached feature files, and API rate limits. The GPT-4.1-mini diagnostic summaries in the package were produced on May 12–13, 2026 local time, use the exact model string `gpt-4.1-mini`, and record the prompts and expected prediction-dump filenames. Full API prediction JSONL dumps are artifact-archive material rather than supplement material. ANALOGYKB is a public resource; the ACL Anthology entry points to `https://github.com/siyuyuan/analogykb`. The included loaders and runners specify the expected local placement for the Wikidata and ConceptNet

Table 7: Appendix diagnostic summary. The main text reports and interprets the numerical tables directly; this table only collects the same diagnostics for navigation.

Diagnostic	Main result	Interpretation
Same-head source use	Source-agnostic scoring is near chance on the pair split (0.136 Acc@1; Table 2) and remains weak on held-out heads (0.214; Table 5); source-aware models score higher and drop to roughly 0.140–0.156 under source shuffling.	Target-head plausibility alone is insufficient.
Tail-matched candidates	Frozen tail cosine is strong (0.348 pair-split Acc@1; 0.338 under tail matching), but learned source-conditioned rankers retain stronger MRR/R@3 in the stricter field (Table 3).	Entity similarity is a real shortcut pressure.
Selector stress	Changing only the hard-negative selector flips the bottleneck–Joint MLP Acc@1 delta: BERT $+0.0213$, MiniLM $+0.0266$, GTE -0.0243, E5 -0.0168; source-shuffled Acc@1 stays near chance (Table 4; Figure 2).	Source use is robust, but model ordering is not.
Held-out target heads	The stabilized bottleneck reaches 0.389/0.591/0.730 Acc@1/MRR/R@3, above matched Joint MLP 0.368/0.573/0.705 and Siamese 0.379/0.582/0.722, but below RelBERT/RoBERTa/GPT practical anchors (Table 5).	The bottleneck helps in a bounded setting.
Held-out relations and ablations	In the expanded held-out-relation diagnostic the bottleneck is above target-only scoring by $+0.1731$ Acc@1 but below the Joint MLP by -0.0293; source-only, margin, cluster-bootstrap, and $K = 16$ diagnostics mark the boundary.	Unseen-relation transfer is a different regime.

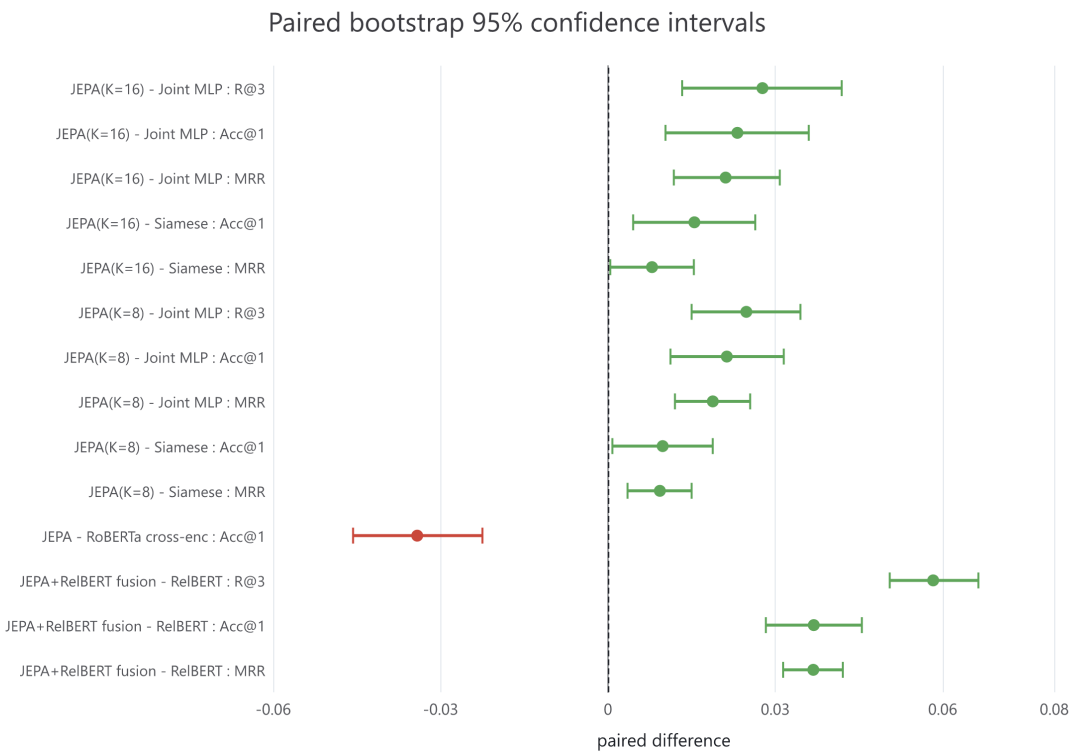

Figure 5: Paired-bootstrap 95% confidence intervals from the tail-similarity-matched held-out-head setting. Green intervals show that the stabilized bottleneck (or the bottleneck+RelBERT fixed-weight fusion) is paired-significantly above the comparator on the same prediction instances; the red interval shows the bounded loss against the RoBERTa cross-encoder. All intervals come from 10k paired bootstrap resamples and correspond to the numbers reported in Section 6.3.

same-relation files. A domain-selection audit over ConceptNet, OpenPI-C, ProPara, Re-DocRED, and MAVEN-ERE documents why only ConceptNet is a smaller same-task sanity check and why the others require different task constructions.

How to read the package. The supplement separates executable code, data loaders, and immutable machine-readable result files. The runner `experiments/run_analogykb_completion_smoke.py` creates the learned-model results and prediction JSONL files. The frozen baselines are produced by `experiments/run_analogykb_frozen_baselines.py`; the local seq2seq likelihood baseline is produced by `experiments/`

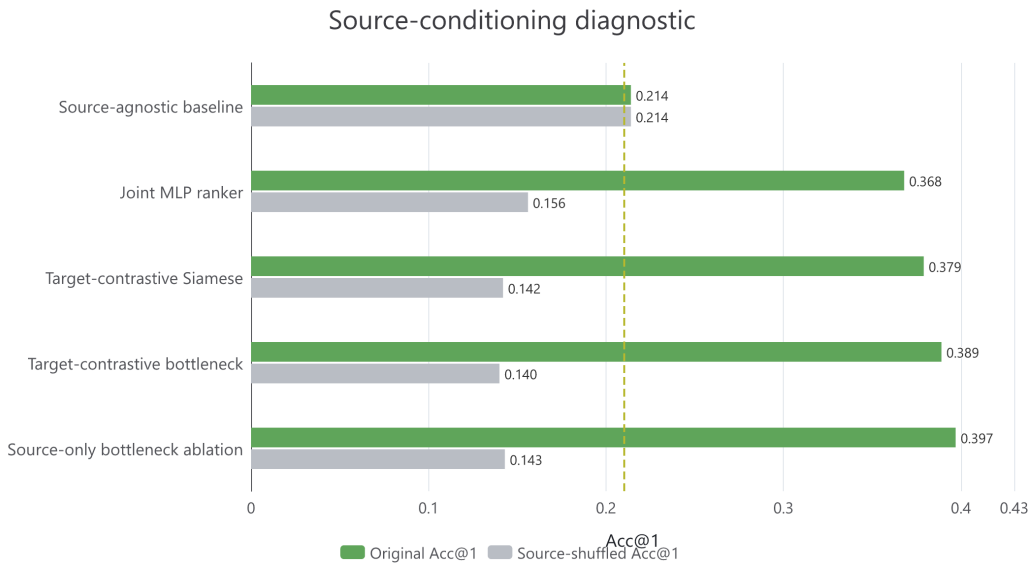

Figure 6: Source-conditioning diagnostic. Source-aware models lose roughly 21–25 Acc@1 points under source shuffling, while the source-agnostic baseline is invariant.

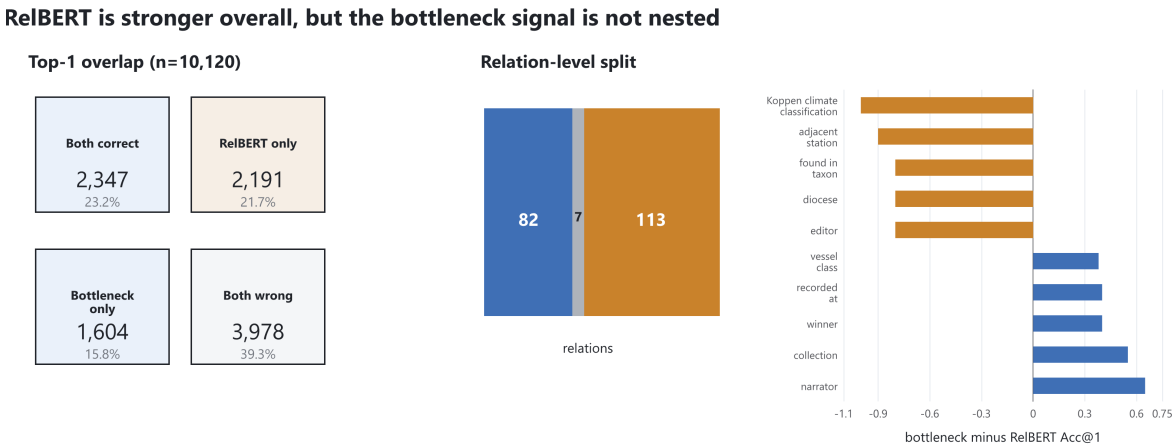

Figure 7: RelBERT disagreement diagnostic. Frozen RelBERT is stronger overall, but wins are not nested: the bottleneck has 1,604 unique top-1 wins and is higher on 82 relations.

Table 8: Held-out target heads with 8 same-head candidates and three seeds. Target heads in test are unseen during training. Source shuffling collapses source-aware learned models, but the joint MLP generalizes much better than the pre-contrastive bottleneck. Frozen tail cosine is a strong non-parametric baseline.

Model	Acc@1	MRR	R@3	Shuffle Acc.
Joint MLP ranker	0.4012	0.5987	0.7376	0.1478
Frozen tail cosine	0.3614	0.5429	0.6194	–
Relation-shaped predictive bottleneck	0.3270	0.5420	0.6809	0.1540
Source-agnostic	0.1893	0.4009	0.4615	0.1893

Table 9: Stepwise target-space stabilization for the predictive bottleneck on the three-seed, 8-way Wikidata pair split. Each row adds a more stable target relation space, but these repairs alone do not beat joint MLP scoring.

Model	Acc@1	MRR	R@3
Joint MLP ranker	0.3735	0.5788	0.7219
+ relation auxiliary + freeze both encoders	0.3458	0.5647	0.7190
+ freeze source and target encoders	0.3387	0.5573	0.7030
+ target-pair pretraining	0.3336	0.5551	0.7094
Naive predictive bottleneck	0.2839	0.5099	0.6357

`run_analogykb_llm_baseline.py`. Paired significance tests use `experiments/paired_bootstrap_analogykb.py`. The reported tables are not hand-entered from a separate spreadsheet: each reported bootstrap interval and diagnostic row is traceable to the JSON/CSV files under `results/`.

Data construction in words. For each gold analogy-style source relation, we select a target head C that has multiple attested outgoing relation pairs in ANALOGYKB Wikidata. The gold tail D^+ is the tail that realizes the source relation around C. The negative tails are also attested tails

Table 10: ConceptNet cross-resource sanity check. ConceptNet has fewer eligible same-head anchors and uses 4-way ranking, so it is not broad multi-dataset validation. It nevertheless reproduces the held-out-head direction: bottleneck minus Joint MLP is Acc@1 $+0.0363$, CI $[0.0062, 0.0664]$, MRR $+0.0324$, CI $[0.0130, 0.0520]$, and R@3 $+0.0342$, CI $[0.0116, 0.0568]$, while the Siamese comparison is tied.

ConceptNet 4-way held-out heads	Acc@1	MRR	R@3
Predictive bottleneck	0.4686	0.6751	0.8724
Siamese pair encoder	0.4589	0.6698	0.8729
Joint MLP ranker	0.4326	0.6430	0.8388
Target-only ranker	0.2785	0.5369	0.7492

of the same C, but under different relations. In the tail-similarity-matched setting, negatives are further chosen so that frozen BERT similarity between B and D^- is close to or higher than the similarity between B and D^+. This makes a simple "choose the tail most similar to B" heuristic less decisive and forces the learned models to use the source relation more carefully.

Splits and what they test. The pair split asks whether models can solve examples drawn from the same broad relation/head distribution. The held-out-head split is stricter: target heads C in test are removed from training, so models must apply a source relation to a new target anchor and a new candidate field. The held-out-relation split removes relation labels before example construction and is included as a stress test for unseen-relation transfer. Because many held-out relations do not retain same-head candidate fields, we report it as a diagnostic rather than as the main claim.

Metrics and uncertainty. Acc@1 is exact top-choice accuracy over the candidate set. MRR rewards ranking the gold tail near the top even when it is not first. R@3 measures whether the gold tail is in the top three. Source-shuffled Acc@1 replaces (A, B) with an unrelated source pair at evaluation time; a source-conditioned model should drop under this intervention. Bootstrap confidence intervals are paired over the same prediction instances, so a positive interval means that model A consistently outranks model B on the same examples rather than merely having a higher unpaired mean.

C Architecture and Capacity Details

All learned models consume frozen text features. The source and target relation encoders are two-layer MLPs over concatenated head, tail, and difference features. The predictive bottleneck uses $E_s(A, B)$, concatenates it with the target head feature C, and maps the result through a two-layer predictor to a target-side relation state. Candidate scores are cosine similarities between the predicted state and each $E_t(C, D_i)$. The Siamese pair encoder uses the same relation encoder form for source and target pairs but does not condition the source relation on C. The Joint MLP ranker uses a two-layer MLP over concatenated (A, B, C, D_i) features for each candidate.

With frozen `bert-base-uncased` features, hidden size 128, and 300 relation labels for auxiliary heads, parameter counts are:

Model-family distinction. The Joint MLP ranker is deliberately non-bottlenecked: it scores each (A, B, C, D_i) candidate directly. The Siamese pair encoder compares an independent source-pair representation with an independent target-pair representation. The predictive bottleneck is stricter than both. It must infer the source relation, condition that relation on C, and produce a single predicted target-side state before candidate scoring. This is why the architecture can fail even when a direct ranker succeeds, and why target-space stabilization is the central experimental variable.

Why not call the direct ranker a transformer cross-encoder? All learned models in the main experiments use frozen entity-string features. The direct ranker is a joint MLP over these frozen features, not a fine-tuned transformer cross-encoder over raw text. This naming matters for the claim: the experiments compare predictive bottlenecks against capacity-matched learned rankers and frozen external relation encoders. The RoBERTa diagnostic in Table 11 is therefore reported separately as a stronger trainable textual ranker rather than a capacity-matched architecture ablation. A partial Wikipedia-title description cache, without gold Wikidata entity identifiers (QIDs) or aliases, lifts RoBERTa to 0.4522/0.6283/0.7458 Acc@1/MRR/R@3; fuller entity linking remains open. A same-budget cached DeBERTa-base (He et al., 2021) check was less stable and lower on average (0.2966 Acc@1, 0.4947 MRR), so we keep RoBERTa as the stronger local cross-encoder diagnostic.

RelBERT target-space diagnostic. RelBERT is included as a strong external relation encoder and remains stronger than the main JEPA architecture.

Table 11: Detailed diagnostic matrix for tail-similarity-matched held-out-head results with 8 candidates, 4.18k–4.24k train examples, and 3.24k–3.52k test examples per seed. The JEPA, Joint MLP, and Siamese rows use the three-seed train/dev/test protocol after selecting target-contrastive hyperparameters by dev Acc@1; frozen RelBERT and frozen tail cosine are five-seed means. GPT-4.1-mini[¶] rows are full three-seed API diagnostics over 10,120 examples: single-choice reports Acc@1 only, complete-ranking reports Acc@1/MRR/R@3. RoBERTa[†] is a three-seed, 200-step textual-ranker diagnostic; E5[§] and FLAN-T5 + retrieved demos[‡] are three-seed frozen/in-context diagnostics. The source-only row[♣] removes C only from the predictor input and is reported as a predictor-input ablation using the selected target-contrastive setting.

Model	Acc@1	MRR	R@3	Source-shuffled Acc@1
GPT-4.1-mini API single choice[¶]	0.4659	–	–	–
GPT-4.1-mini API full ranking[¶]	0.4147	0.5790	0.6494	–
Frozen RelBERT pair cosine	0.4474	0.6233	0.7340	–
Fine-tuned RoBERTa cross-encoder[†]	0.4253	0.6110	0.7302	–
Target-contrastive predictive bottleneck	0.3888	0.5912	0.7297	0.1399
Source-only predictive bottleneck[♣]	0.3967	0.5960	0.7340	0.1430
Joint MLP ranker	0.3677	0.5726	0.7052	0.1559
Target-contrastive Siamese pair encoder	0.3791	0.5821	0.7223	0.1423
Frozen E5-small-v2 tail cosine[§]	0.3443	0.5214	0.5918	–
Frozen MiniLM tail cosine	0.3308	0.5143	0.5923	–
Frozen tail cosine	0.3175	0.4919	0.5457	–
Source-agnostic	0.2137	0.4187	0.4744	0.2137
FLAN-T5 likelihood	0.1495	0.3647	0.4126	–
FLAN-T5 + retrieved demos[‡]	0.1392	0.3518	0.3914	–

Table 12: Cluster-bootstrap 95% confidence intervals for the tail-similarity-matched held-out-head setting, using 10k resamples over relation clusters (202 clusters) or target-head clusters (432 clusters). Values are model A minus model B over the same 10,120 paired prediction instances. The matched Joint MLP comparison remains positive under both cluster units; the Siamese comparison is weaker; RelBERT remains stronger on Acc@1 and MRR.

Comparison	Bootstrap unit	ΔAcc@1	ΔMRR	ΔR@3
JEPA – Joint MLP	Relation	$+0.0213\,[+0.0003, +0.0420]$	$+0.0188\,[+0.0032, +0.0337]$	$+0.0248\,[+0.0050, +0.0444]$
JEPA – Joint MLP	Target head	$+0.0213\,[+0.0048, +0.0379]$	$+0.0188\,[+0.0066, +0.0309]$	$+0.0248\,[+0.0086, +0.0414]$
JEPA – Siamese	Relation	$+0.0098\,[-0.0053, +0.0247]$	$+0.0093\,[-0.0008, +0.0193]$	$+0.0080\,[-0.0053, +0.0220]$
JEPA – Siamese	Target head	$+0.0098\,[-0.0033, +0.0228]$	$+0.0093\,[+0.0007, +0.0180]$	$+0.0080\,[-0.0050, +0.0213]$
JEPA – RelBERT	Relation	$-0.0597\,[-0.0875, -0.0310]$	$-0.0335\,[-0.0553, -0.0116]$	$-0.0052\,[-0.0325, +0.0208]$
JEPA – RelBERT	Target head	$-0.0597\,[-0.0799, -0.0395]$	$-0.0335\,[-0.0485, -0.0182]$	$-0.0052\,[-0.0256, +0.0151]$

Table 13: Predicted-state margin diagnostic from saved prediction dumps. Margin is the gold candidate score minus the highest non-gold score in the same-head field; Q25 and Q75 are the 25th and 75th percentiles. For predictive JEPA, scores are cosine similarities between the predicted target-side state and candidate target-pair embeddings. The diagnostic is not a paired significance test, but it shows that target-contrastive stabilization moves the predicted state closer to the gold target-pair embedding relative to the hardest negative.

Setting	n	Acc@1	Mean margin	Median margin	Q25	Q75	Gold / hard-neg mean
Pre-contrastive relation-auxiliary	10,120	0.3150	-0.2801	-0.2374	-0.6592	0.0805	0.1527 / 0.4328
Target-contrastive, dev-selected	10,120	0.3887	-0.0884	-0.0695	-0.2856	0.1171	0.3088 / 0.3972

To check whether this is merely a missing target-space choice, we ran a five-seed diagnostic that freezes RelBERT as the target-pair space $E_t(C, D)$ and trains a small predictor into that space on the same tail-matched held-out-head split. With the same BERT entity-string features as the main JEPA runs, the predictor reaches $0.3858/0.5860/0.7213$ Acc@1/MRR/R@3. Conditioning instead on source RelBERT plus the target-head embedding gives $0.4303/0.6246/0.7673$: below frozen RelBERT pair cosine on Acc@1 (-0.0171, CI $[-0.0249, -0.0093]$), tied on MRR within uncertainty, and higher on R@3 ($+0.0334$, CI $[0.0268, 0.0401]$). A source-RelBERT-only projection is stronger still at $0.4633/0.6469/0.7795$, above frozen RelBERT on all three metrics (Acc@1 $+0.0158$, CI $[0.0084, 0.0230]$). For the three learned RelBERT representation-space predictors, source shuffling reduces Acc@1 to 0.127–0.136, confirming that the diagnostic still depends on source-relation information; a relation-level audit shows the source projection above frozen RelBERT on 113/210 relations but weaker in high-overlap buckets. We also tested small trainable residual adapters on top of frozen RelBERT pair embeddings. The adapted predictive bottleneck is nearly tied with frozen RelBERT on Acc@1 but improves rank quality, while a candidate-wise direct adapter is stronger still (Table 16). A three-weight frozen-logit distillation sweep peaks at $0.4640/0.6499/0.7919$ for the adapted bottleneck; embedding-level distillation is

Table 14: Negative architecture probes before target-contrastive stabilization. The middle column states why each probe exists. Shallow readout, anchoring, distillation, and closed-label bottlenecks do not close the held-out-head gap; the positive target-contrastive result is reported separately in Table 11.

Probe	Question tested	Acc@1	MRR
Relation-shaped bottleneck	Baseline before target-contrastive shaping	0.3438	0.5502
Candidate-field centering	Does local target-field normalization fix transfer?	0.3348	0.5388
Field normalization	Do target-field mean/variance features help?	0.3382	0.5418
Query attention over field	Does a deeper candidate-field readout help?	0.3383	0.5422
Anchor-conditioned prediction	Does pooled target-anchor context help?	0.3364	0.5460
Target-anchor counterfactual	Does enforcing anchor sensitivity help?	0.3432	0.5497
Longer training	Is the gap just undertraining?	0.3446	0.5499
Teacher distillation	Can the joint MLP transfer its ranking signal?	0.3389	0.5478
Relation-logit bottleneck	Is closed-label relation classification enough?	0.3091	0.5263
Joint MLP ranker	Non-bottleneck reference	0.4119	0.6047

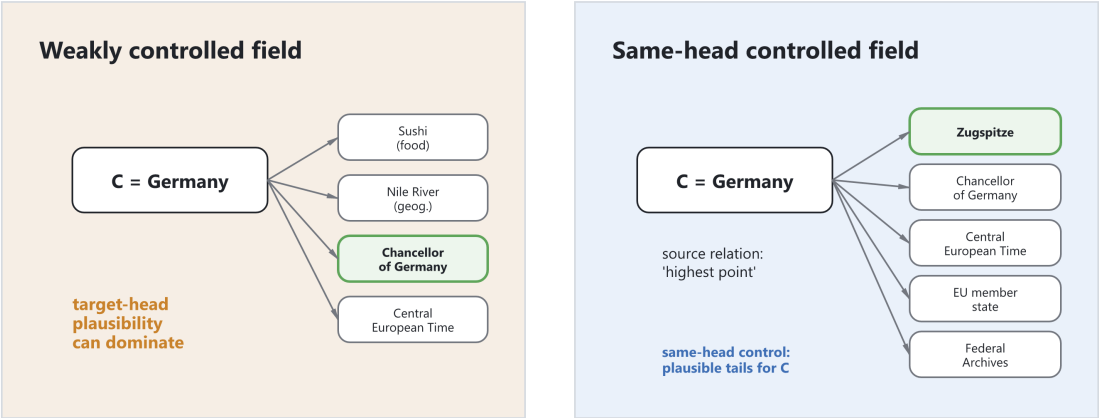

Figure 8: Why same-head candidates matter. Random or weakly controlled candidate fields can reward target-only plausibility or source-tail similarity. Same-head construction makes every candidate a plausible tail for the same target head, so success depends more directly on selecting the candidate relation that matches the source pair.

Table 15: Capacity comparison for the main learned models before encoder freezing after pretraining. The predictive bottleneck and Joint MLP ranker have similar parameter budgets.

Model	Parameters	Trainable
Target-only ranker	295k	295k
Siamese pair encoder	821k	821k
Joint MLP ranker	1.00M	1.00M
Predictive bottleneck	1.03M	1.03M

weaker, and both remain below the direct adapter's $0.4808/0.6617/0.7996$. This supports the target-space-stabilization thesis while sharpening the limitation: RelBERT-space JEPA is useful as a bottleneck diagnostic, but the strongest RelBERT-space ranker here is not bottlenecked. Thus RelBERT-logit distillation reduces but does not remove the bottleneck cost under this setting. Updating the RelBERT encoder inside the predictive bottleneck does not by itself solve the problem. A three-seed encoder-updated RelBERT-JEPA variant that predicts one target-side state from the encoded source pair reaches only $0.3054/0.5309/0.6798$, while source shuffling drops it to $0.1554/0.3680/0.4138$. Thus it uses the source relation, but remains below frozen RelBERT and far below direct fine-tuned RelBERT cosine.

Table 16: Five-seed RelBERT representation-space adapter diagnostic on tail-matched held-out heads. The adapted predictive bottleneck adds trainable residual adapters but still predicts one target-side state before candidate scoring. It improves over frozen RelBERT in MRR $+0.0128$, CI $[0.0074, 0.0184]$, and R@3 $+0.0455$, CI $[0.0383, 0.0529]$, but not Acc@1. The direct adapter is stronger than the adapted bottleneck on all metrics, with Acc@1 $+0.0360$, CI $[0.0281, 0.0439]$.

Model	Acc@1	MRR	R@3
Frozen RelBERT pair cosine	0.4474	0.6233	0.7340
Conditioned RelBERT bottleneck	0.4351	0.6258	0.7611
Adapted RelBERT bottleneck	0.4445	0.6359	0.7794
Source RelBERT projector	0.4678	0.6493	0.7818
Direct RelBERT adapter ranker	0.4808	0.6617	0.7996

D Additional Diagnostics

This appendix section collects stress tests and alternative-explanation checks. Each diagnostic asks what aspect of the same-head suite is being measured: candidate-set artifacts, relation-label exposure, direct-ranker capacity, selector construc-

tion, or target-space stabilization. The results make the benchmark outputs auditable without expanding the main claim.

Held-out relations. To test transfer beyond seen relation labels, we also split by relation before constructing same-head tail-similarity-matched examples. A coverage audit shows that the original 300-relation split yields only 320–520 test examples per seed, so Table 17 uses all 669 loaded relations with a 0.33 held-out split, yielding 1,320–1,460 test examples per seed. This split separates source-use detection from broad unseen-relation generalization:

Table 17: Three-seed expanded held-out-relation diagnostic with all 669 loaded Wikidata relations, a 0.33 relation-heldout split, and 8 same-head tail-matched candidates. The predictive bottleneck is above target-only scoring (Acc@1 $+0.1731$, CI $[0.1571, 0.1890]$) but below the direct MLP ranker (-0.0293, CI $[-0.0460, -0.0129]$) and Siamese pair encoder (-0.0160, CI $[-0.0274, -0.0043]$).

Model	Acc@1	MRR	R@3	Shuffle Acc.
Direct MLP ranker	0.3567	0.5589	0.6850	0.1560
Siamese pair encoder	0.3435	0.5493	0.6762	0.1480
Predictive bottleneck	0.3282	0.5385	0.6696	0.1444
Target-only ranker	0.1559	0.3703	0.4208	0.1559

Thus unseen-relation transfer is a separate benchmark axis. The central positive claim remains on held-out target heads under controlled same-head candidate fields, while this diagnostic records how the same models behave when the relation inventory changes. Seed-0 per-relation summaries in the package show heterogeneous behavior rather than a clean relation-type taxonomy: JEPA gains over direct scoring on relations such as native language, mother, and work location, but loses sharply on relations such as candidacy in election and continent. We treat this as per-relation audit evidence for future selector and relation-family analysis.

Hard-negative selection backbone. The tail-matched benchmark uses frozen BERT similarity to select same-head distractors whose tails are close to the source tail. We therefore rebuilt the held-out-head candidate fields with a Sentence-BERT-style `all-MiniLM-L6-v2` selector while keeping the learned models' frozen BERT entity features unchanged. Across three seeds, predictive JEPA remains above the Joint MLP ranker on Acc@1 $+0.0266$, CI $[0.0164, 0.0368]$, and MRR $+0.0174$, CI $[0.0107, 0.0241]$, and above the Siamese pair encoder on Acc@1 $+0.0168$, CI $[0.0078, 0.0259]$, and MRR $+0.0096$, CI $[0.0039, 0.0154]$. Three-seed GTE-small and E5-small-v2 selector stress tests reverse the architecture ranking: JEPA falls below the Joint MLP by Acc@1 -0.0243 and -0.0168, respectively. Thus selector stress records where source use is stable and where model ordering depends on candidate construction. As an anti-similarity sanity check, we also choose same-head distractors from the low end of source-tail similarity. This easier variant raises all source-aware models: predictive JEPA reaches $0.5075/0.6751/0.8017$, Joint MLP $0.5017/0.6672/0.7833$, and Siamese $0.4925/0.6644/0.7903$. Because this setting removes the main source-tail shortcut stress, we keep the tail-similarity-matched construction as the primary benchmark.

D.1 Cross-Resource, Backbone, and Domain Checks

ConceptNet cross-resource sanity check. To test whether the target-contrastive repair is specific to Wikidata, we also reran the same-head completion protocol on the ANALOGYKB ConceptNet split. ConceptNet has fewer usable same-head multi-relation anchors in our filters, so these are smaller 4-way diagnostics rather than replacements for the main tail-matched held-out-head Wikidata setting: our coverage audit finds 26 loaded ConceptNet relations, 6,726 capped pairs, 233 eligible same-head anchors, and 25 usable relations for 4-way ranking, compared with 300 Wikidata relations, 101,330 capped pairs, 750 eligible anchors, and 219 usable relations for 8-way ranking. In the 4-way pair split, across three seeds, target-contrastive JEPA reaches $0.5729/0.7379/0.8867$ Acc@1/MRR/R@3, compared with $0.4298/0.6475/0.8489$ for the Joint MLP. A 10k paired bootstrap gives JEPA-minus-Joint MLP Acc@1 $+0.1409$, CI $[0.1045, 0.1764]$, and MRR $+0.0891$, CI $[0.0661, 0.1119]$; against the Siamese pair encoder, JEPA is a statistical tie on Acc@1 and MRR. The 4-way held-out-head ConceptNet variant in Table 10 gives JEPA 0.4686 Acc@1 and 0.6751 MRR versus Joint MLP $0.4326/0.6430$, with JEPA-minus-Joint MLP Acc@1 $+0.0363$, CI $[0.0062, 0.0664]$. These checks support the mechanism outside Wikidata, while keeping Wikidata as the main benchmark because ConceptNet yields a smaller candidate field.

Frozen-feature backbone check. For the main benchmark we also reran the dev-selected tail-

matched held-out-head protocol with the candidate field held fixed by the original BERT selector but the learned-model features changed to frozen RoBERTa-base or all-MiniLM-L6-v2 entity-string embeddings. Table 18 reports paired deltas for the bottleneck rows against the matched Joint MLP. This is a narrow backbone diagnostic, not a claim that these are optimal RoBERTa or MiniLM systems.

A broader domain-selection audit reaches the same conclusion: Wikidata has much larger same-head coverage than ConceptNet under the controlled candidate field, while OpenPI-C and ProPara are procedural state-tracking resources and Re-DocRED is a document-level multi-label relation extraction resource. These resources are therefore reported as domain-selection evidence and future task-design directions rather than folded into a mismatched validation claim.

We also audited the ANALOGYKB analogous-relation files as a possible broader evaluation. They are not a drop-in replacement for the same-head same-relation protocol. The Wikidata analogous split primarily uses a `relation1`/`relation2` and `tuple_r1`/`tuple_r2` schema, which would change the task into explicit cross-relation analogy transfer. The ConceptNet analogous split has only five coarse relation-family entries. We therefore leave analogous-split evaluation to future work rather than mixing a different task into the present benchmark.

We additionally audited MAVEN-ERE event graphs as a possible event-relation validation domain. The data are promising for future context-entailed relation prediction: in a 1,000-document train sample, mined two-hop rules such as BEFORE,BEFORE→BEFORE retain high validation confidence. However, the obvious bridge task is not suitable as a second main benchmark here. For five high-confidence rules, we found 327,045 positives with rule support but no same-source rule-ambiguous negatives, so a symbolic path-rule lookup would solve the construction. We therefore report MAVEN-ERE as a task-design audit for future event-relation benchmarks.

D.2 Alternative Explanations and Ambiguity

Prototype and memory baselines. We also test the simpler hypothesis that a few relation prototypes or nearest training-pair memories are enough. As an additional frozen sentence-embedding check, E5-small-v2 and GTE-small tail-cosine heuristics reach only 0.3443 and 0.2962 Acc@1, respectively, under the same three-seed tail-matched held-out-head setting; JEPA remains higher than E5 by Acc@1 $+0.0467$, CI $[0.0351, 0.0584]$. For each training relation, we average frozen pair features $[h, t, t-h, h \odot t]$ into a centroid. At test time, the inferred-prototype baseline first maps the source pair to its nearest training-relation centroid and then ranks target candidates by similarity to that centroid. On the three-seed tail-matched held-out-head setting, this reaches only Acc@1 0.2977, MRR 0.4780, and R@3 0.5284. Predictive JEPA is higher by Acc@1 $+0.0931$, CI $[0.0816, 0.1046]$, MRR $+0.1126$, CI $[0.1043, 0.1210]$, and R@3 $+0.1982$, CI $[0.1862, 0.2101]$. An oracle variant that is given the test relation label before ranking is much stronger, with Acc@1 0.4231, MRR 0.5911, and R@3 0.6753; it exceeds JEPA on Acc@1 but is tied on MRR and below on R@3. This separates two issues: relation-label centroids are useful if the source relation is already known, but source-conditioned completion still requires inferring the target-side relation state from the source pair. A nonparametric kNN memory baseline that retrieves the 16 nearest training pairs to the source pair is weaker still, with Acc@1 0.2750, MRR 0.4580, and R@3 0.5066. Target-contrastive JEPA exceeds this memory baseline by Acc@1 $+0.1157$, CI $[0.1040, 0.1273]$, and MRR $+0.1326$, CI $[0.1242, 0.1408]$. Thus the main result is not explained by nearest-neighbor retrieval over frozen pair features.

Ambiguity and label granularity. Examples are built from one held-out source relation and one gold target pair, but same-head candidate fields can contain semantically adjacent tails. The benchmark therefore tests recovery of the labeled source relation rather than human adjudication of every plausible analogy. A seed-0 automatic audit found another candidate with the source relation in 1/3,520 examples and multi-labeled candidates in 203/3,520 examples: 39 gold and 164 negative candidates. In a fixed-rubric audit of a deterministic 61-row stratified sample, all 30 sampled multi-labeled gold rows were official gold answers with extra labels; among 31 sampled non-gold rows, 1 was a duplicate-gold case, 5 were label-granularity cases, 7 were adjacent-relation cases, and 18 were clean negatives. Thus the strict metric can under-credit a few granular overlaps, but the sampled audit does not suggest that many non-gold candidates

Table 18: Paired 10k bootstrap intervals for frozen-feature backbone checks. The candidate fields, hard-negative selector, split, and target-contrastive setting match the main BERT-base diagnostic; only the frozen entity-string feature backbone changes.

Features	Comparison	Acc@1 Δ	MRR Δ	R@3 Δ
RoBERTa-base	Bottleneck – Joint MLP	$+0.0143\,[+.0052, +.0232]$	$+0.0123\,[+.0069, +.0180]$	$+0.0178\,[+.0097, +.0259]$
RoBERTa-base	Source-only – Joint MLP	$+0.0210\,[+.0122, +.0299]$	$+0.0159\,[+.0105, +.0214]$	$+0.0154\,[+.0074, +.0236]$
MiniLM-L6	Bottleneck – Joint MLP	$+0.0117\,[+.0027, +.0208]$	$+0.0066\,[+.0008, +.0124]$	$+0.0019\,[-.0066, +.0105]$
MiniLM-L6	Source-only – Joint MLP	$+0.0147\,[+.0057, +.0238]$	$+0.0090\,[+.0032, +.0149]$	$+0.0064\,[-.0021, +.0147]$

are hidden duplicate answers. Because we do not consolidate aliases, inverses, or related Wikidata relations globally, we emphasize paired comparisons, source shuffle, and RelBERT/JEPA disagreement rather than treating every error as a clean reasoning failure. An all-669-relation lexical audit found no exact or normalized duplicate labels and only three high-overlap label pairs, so automatic label collapse would be sparse and still semantically unsafe without adjudication.

Table 19: Representative fixed-rubric ambiguity audit rows.

Rubric	Audited contrast	Interpretation
duplicate gold	Quebec–Laval vs. Quebec City for "contains administrative entity"	Strict single-gold scoring under-credits this row.
label granularity	Super Mario Bros.–Takashi Tezuka vs. Shigeru Miyamoto	Game-design and creative-contributor labels overlap.
adjacent relation	The Da Vinci Code–Paul Bettany vs. Dan Brown	Work provenance is adjacent to, but not cast membership.
clean negative	Super Mario Bros.–Super Nintendo vs. Shigeru Miyamoto	Creative contributor, not platform.

The released CSV records the fixed-rubric 61-row audit, and the paired validation JSON records the row counts and rubric categories. We treat this as a bounded data audit rather than completed external annotation evidence.

Candidate-set size. The main benchmark uses 8 candidates; Table 20 shows that the bottleneck gain over joint MLP and Siamese also holds at 16. Larger fields are coverage-limited: at 24 candidates the seed-0 test set drops to 1,040 examples, and at 32 no train/test examples satisfy the current filters.

Table 20: Candidate-size diagnostic on tail-matched held-out heads. At K=16, the bottleneck remains above Joint MLP by Acc@1 $+0.0232$, CI $[0.0103, 0.0360]$, and MRR $+0.0211$, CI $[0.0118, 0.0308]$.

Model	K=8 Acc@1	K=16 Acc@1	K=16 MRR
Predictive bottleneck	0.3815	0.3049	0.4860
Siamese pair encoder	0.3657	0.2883	0.4774
Joint MLP ranker	0.3688	0.2818	0.4652
Target-only ranker	0.2137	0.1073	0.2631

Target-contrastive sensitivity. Table 21 reports the dev-selected search. The no-contrastive row is much weaker, while the selected $(0.4, 0.2)$ setting and the earlier $(0.2, 0.2)$ setting are both strong. Removing the auxiliary relation-classification head but keeping target-contrastive loss remains close to the full model ($0.3870/0.5877/0.7233$ vs. $0.3905/0.5903/0.7265$), so the contrastive target-space loss is the primary stabilizer.

Table 21: Dev-selected target-contrastive search. Rows are sorted by mean dev Acc@1; only the selected row is used for the main held-out test comparison.

W	T	N	Dev	Acc.	MRR	R@3
0.40	0.2	3	0.3838	0.3888	0.5912	0.7297
0.20	0.2	3	0.3781	0.3821	0.5853	0.7263
0.20	0.1	3	0.3768	0.3804	0.5826	0.7180
0.40	0.1	3	0.3684	0.3804	0.5843	0.7254
0.10	0.1	3	0.3678	0.3728	0.5779	0.7202
0.40	0.4	3	0.3674	0.3792	0.5830	0.7226
0.10	0.2	3	0.3641	0.3748	0.5805	0.7241
0.20	0.4	3	0.3594	0.3702	0.5770	0.7187
0.10	0.4	3	0.3381	0.3581	0.5673	0.7036
0.00	0.2	3	0.3073	0.3227	0.5354	0.6658

The earlier $(0.2, 0.2)$ seed-group audit remains a robustness check: on seeds 3–4 it is above Siamese (Acc@1 $+0.0162$, MRR $+0.0127$) but only slightly above Joint MLP (Acc@1 $+0.0033$, CI $[-0.0097, 0.0164]$).

Does the predictor use the target head? Table 22 removes C from the predictor input while leaving target-pair encoding $E_t(C, D_i)$ unchanged. The source-only predictor is slightly above the full predictor, and source shuffling still collapses performance. Thus the current gain is source-conditioned prediction into a stabilized target-pair space, not evidence that appending a frozen C vector is the right anchor conditioning mechanism.

Table 22: Predictor-conditioning ablation. Removing C only from the predictor input does not hurt, while source shuffling remains near chance.

Model	Acc@1	MRR	R@3	Shuffle Acc.
Full predictor	0.3888	0.5912	0.7297	0.1399
Source-only predictor	0.3967	0.5960	0.7340	0.1430

Local target-head ambiguity and rank errors. Performance falls as target-head relation diversity grows: JEPA reaches $0.4273/0.6128$ Acc@1/MRR on heads with at most eight observed relations, $0.3844/0.5890$ on 9–15, and $0.3607/0.5747$ on 16–31. Joint MLP follows the same trend but is lower in every bucket. JEPA also has 3,454 rank-2/3 near-misses and 1,404 cases where it places the gold in the top three while Joint MLP ranks it below three, so many errors are local candidate-discrimination failures.

Label exposure and direct-ranker capacity. A three-seed check shows label dependence: 0%, 10%, and 25% exposed labels trail the joint MLP, 50% ties it, and 20% noisy labels drops the bottleneck to $0.3426/0.5538$ Acc@1/MRR. A hidden-256 joint MLP reaches $0.3730/0.5767$, but the full-label bottleneck remains above it with Acc@1 $+0.0176$ and MRR $+0.0137$. We keep exact ANALOGYKB identifiers and treat label granularity as a data limitation.