RoleLex: Recovering Cross-Speaker Referential Structure from Frozen Language-Model States

Anonymous authors
Paper under double-blind review

Abstract

When two people talk about the same thing, the words they use for it drift: in one MapTask dialogue a route giver's "the old mill" is answered with "the mill wheel," and both speakers then settle on "it." A system that tracks such references has to store each mention as it arrives, before it knows what will be asked later. We study how to build that stored key from a frozen dialogue decoder and find that no single layer holds the evidence. On MapTask dialogues, decoder depths read out separately pick different antecedents for 47–68% of queries, and the best depth changes from one backbone to the next. RoleLex keeps the depths apart instead of choosing between them: four depths are projected and normalized independently and each occupies its own block of a 256-dimensional key, so one dot product between keys equals the average of four depth-wise cosines. On 13,077 human-linked mentions from 128 MapTask dialogues, evaluated leave-one-map-out, this fixed key beats the best development-selected single depth for Qwen3.5-9B, Gemma 4 E4B-it, and OLMo 3 7B, matches or beats the best selected pair of depths, and costs the same parameters and arithmetic as a final-layer key. Controls show that what transfers across backbones is keeping the depth scores separate; a dense map over the same states has four times the parameters, does worse, and uses fewer than half of its output directions. The gains are largest when no antecedent shares the query's wording, carry over to deciding whether any antecedent exists, and reappear on OneCommon.

1 Introduction

Conversation builds shared worlds by returning to the same entities. In one HCRC MapTask dialogue, the route giver asks the follower to find "the old mill"; the follower, whose copy of the map labels that spot differently, answers "the mill wheel," and within a few turns both of them are calling it "it." Reference of this kind is collaborative. Descriptions are proposed, accepted, repaired, and reused as common ground develops (Clark & Wilkes-Gibbs, 1986), and a system that wants to know what the follower's "it" denotes has to link it to the giver's earlier mention, across a change of speaker and a change of wording, using only what has been said so far.

A frozen dialogue decoder already represents every mention it has processed, so the natural design is to read out the state at each mention, project it to a small key, cache the key, and later retrieve by cosine. The question is which state. Native cosine over contextual states is an unreliable readout (Ethayarajh, 2019; Zhou et al., 2021; 2022), and a learned projection of the final layer is the conventional fix. But the final layer is not where the referential evidence is best. In our experiments the strongest single-depth key sits midway through the decoder for Qwen and Gemma, the best depth differs across backbones, and, read out separately, the depths disagree about the antecedent on 47–68% of queries. Choosing a better layer recovers some of this evidence. It cannot recover all of it, because the layers do not agree.

The interface is what makes this hard. A retrieval key must be computed once, cached before future queries are known, and compared with a single dot product. Within that

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

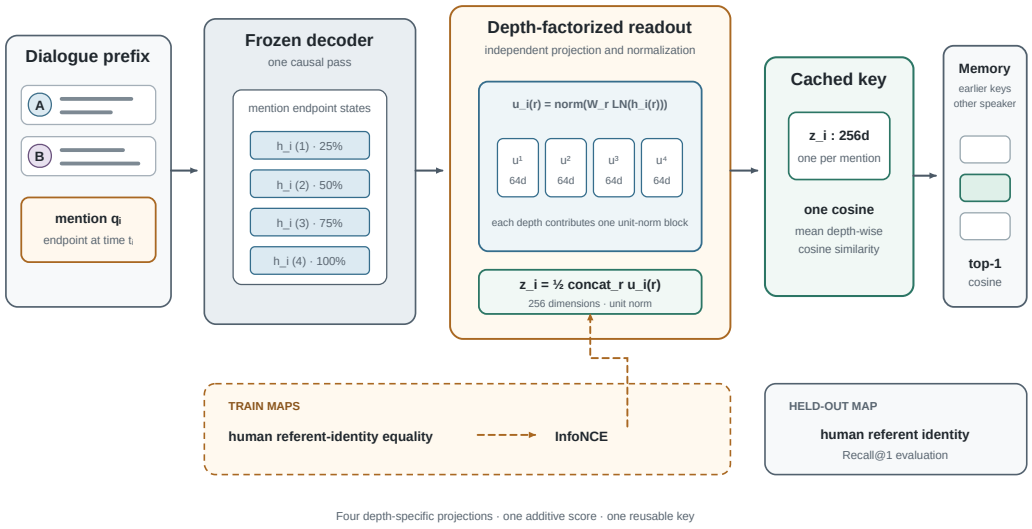


Figure 1: RoleLex in one pass. A frozen decoder reads the causal dialogue prefix once; at the mention endpoint, states from four depths are each layer-normalized, projected to 64 dimensions, and unit-normalized, then concatenated into one 256-dimensional key that is cached. Retrieval compares the query key against all cached keys from the other speaker with a single cosine, which by construction equals the mean of four depth-wise cosines. Only the four projections are trained, from human referent-identity equality on training maps; test maps are never seen.

interface, the two obvious ways to use several depths each give something up. Mixing the states before projection, as scalar mixtures do (Peters et al., 2018), forces four depth-specific relations into one coordinate system. Keeping a separate vector per depth abandons the single key.

RoleLex takes a third route: a direct sum. Each depth gets its own small linear map, its output is normalized on its own, and the normalized blocks are concatenated into one 256-dimensional key (Figure 1). Because the blocks occupy disjoint coordinates, the dot product between two keys is exactly the average of four depth-wise cosines. The interface is untouched; the key is still a unit vector searched by maximum inner product, and four $d \times 64$ maps cost the same as one $d \times 256$ map over the final layer. What changes is that no depth can overwrite another. The name records the two things that change between a query and its antecedent: the speaker's role and the lexical form.

Testing this idea requires retrieval cases with human-verified referents and a strictly causal candidate set. We build MapTask-RefMem by aligning 13,077 GMMT reference expressions from 128 dialogues (Li et al., 2026a) with the original HCRC landmark annotation. Each mention is a query; its memory is every earlier mention by the other speaker; 9,088 queries have a correct antecedent in memory and 3,989 do not. Evaluation is leave-one-map-out over sixteen maps, so a held-out map shares no landmarks with training. Only the projections are trained, with an InfoNCE objective (van den Oord et al., 2018) over human referent identities, and the decoder stays frozen. The companion Seeing Is Not Sharing study uses the same expressions to analyze interpretation matching and common-ground bias (Li et al., 2026b); we ask a different question of the same data, namely whether a causal frozen state can retrieve the other speaker's mention.

Our contributions are as follows.

- A benchmark for causal cross-speaker reference retrieval. MapTask-RefMem pairs human referent identities with a strictly online candidate set and a held-out-map protocol, and separates retrieval from abstention (Section 2).
- A depth-factorized retrieval key. RoleLex stores several decoder depths in one unit vector whose dot product is an additive similarity over depth, at the parameter and arithmetic cost of a final-layer key (Section 3).
- Evidence that the factorization itself is what helps. A fixed four-depth key beats the best development-selected single depth on all three backbones and matches the best selected pair (Section 4.2). Controls that mix depths before projection or let a dense map see all four states do worse; the dense map collapses to fewer than half of its output directions and falls further behind as training maps are removed (Section 4.3).
- Where and how far the gain reaches. The advantage concentrates on queries whose wording differs from every correct antecedent and on distant antecedents (Section 4.4), extends to deciding whether an antecedent exists at all (Section 4.5), and reappears on OneCommon with the architecture transferred unchanged (Section 4.7).

Different layers of a frozen decoder remember a mention differently. A single retrieval vector has room for all of them.

2 Cross-Speaker Reference Retrieval

2.1 The online task

For mention occurrence i, let x_i be the causal dialogue prefix ending at the annotated mention span, s_i its speaker, t_i its completion time, and r_i its human referent identity. Its eligible memory is

$$\mathcal{C}_i = \{j : t_j < t_i,\ s_j \neq s_i\}, \tag{1}$$

every earlier mention by the other speaker. We call i a retrieval case if some $j \in \mathcal{C}_i$ has $r_j = r_i$ and an abstention case otherwise. Given a similarity g, retrieval returns

$$\hat{j}_i = \arg\max_{j \in \mathcal{C}_i} g(i, j), \tag{2}$$

and succeeds when $r_{\hat{j}_i} = r_i$. For joint retrieval or abstention, the system returns no antecedent when $\max_{j \in \mathcal{C}_i} g(i, j) < \theta$, with θ chosen on each fold's development map. The task isolates referential access: mention spans and speaker roles are given, and what is tested is whether the preceding cross-speaker history can be searched for the right antecedent.

2.2 MapTask-RefMem and held-out evaluation

We align every released GMMT reference occurrence (Li et al., 2026a) with the original HCRC MapTask reference annotation, so that each mention carries a human landmark identity. The result is 13,077 queries from 128 dialogues and 16 maps: 9,088 retrieval cases and 3,989 abstention cases. Among the retrieval cases, 7,031 have a preferred antecedent worded differently from the query (the different-surface slice), and 5,919 have no correct eligible antecedent that repeats the query's surface at all (the stricter no-same-surface slice); 765 belong to the official HCRC lexical-variant subset.

Each fold holds out one complete map for testing. A deterministic SHA256 rule picks a distinct development map, and the remaining fourteen maps are used for training. Human landmark identity defines the training relation: two occurrences are positives if their identities match. Everything stays causal, since both the frozen state and the eligible memory contain only dialogue available when the mention ends, and human labels from the held-out map are touched only for scoring. Appendix A documents the alignment and folds.

Our primary metric is micro Recall@1 over retrieval cases; we also report map-macro Recall@1 and the number of held-out maps on which a paired improvement is positive. Over

all 13,077 queries we report joint accuracy, abstention F1, and retrieval-versus-abstention AUROC.

3 RoleLex

3.1 A depth-factorized key

A frozen causal decoder supplies mention-endpoint states at R fixed depths,

$$h_i^{(r)} = f_{\mathrm{LM}}^{(r)}(x_i) \in \mathbb{R}^d, \qquad r \in \mathcal{D}, \quad |\mathcal{D}| = R, \tag{3}$$

all from the same forward pass. Our default sets $R = 4$ at roughly 25%, 50%, 75%, and 100% of the decoder. Each state is layer-normalized without affine parameters, projected by its own bias-free map, and unit-normalized:

$$v_i^{(r)} = W_r \mathrm{LN}_r(h_i^{(r)}), \qquad u_i^{(r)} = \frac{v_i^{(r)}}{\|v_i^{(r)}\|_2} \in \mathbb{R}^{256/R}. \tag{4}$$

The key is the scaled concatenation of these blocks,

$$z_i = \frac{1}{\sqrt{R}} \left[u_i^{(1)}; \ldots; u_i^{(R)} \right] \in \mathbb{R}^{256}, \tag{5}$$

which has unit norm by construction. Because the blocks occupy disjoint coordinates, a dot product between two keys is an average of depth-wise cosines,

$$z_i^\top z_j = \frac{1}{R} \sum_{r \in \mathcal{D}} u_i^{(r)\top} u_j^{(r)}, \tag{6}$$

and inference is ordinary maximum-inner-product search, $\hat{j}_i = \arg\max_{j \in \mathcal{C}_i} z_i^\top z_j$.

The restriction here is structural rather than a matter of capacity. A dense map over the concatenated states would contain every cross-depth bilinear term; RoleLex reserves one coordinate block per depth and normalizes that depth's evidence before it is added to the others (Appendix C writes out the corresponding kernel). A depth with a confident, wrong opinion about one pair can therefore contribute at most one quarter of the score, and a depth whose evidence would be drowned out under a shared norm keeps its full weight.

3.2 What the factorization controls, and what it does not

The construction extends to fixed nonnegative weights β_r with $\sum_r \beta_r = 1$:

$$z_i(\beta) = \left[\sqrt{\beta_1} u_i^{(1)}; \ldots; \sqrt{\beta_R} u_i^{(R)} \right], \qquad z_i(\beta)^\top z_j(\beta) = \sum_r \beta_r u_i^{(r)\top} u_j^{(r)}. \tag{7}$$

We use uniform $\beta_r = 1/R$ as the primary key because it is fixed across folds, backbones, and datasets, and we also fit softmax-parameterized weights to ask whether the data prefer a different fixed kernel.

A parameter-matched alternative concatenates the unnormalized projections $v_i^{(r)}$ and normalizes once, globally. Writing $\alpha_i^{(r)} = \|v_i^{(r)}\|_2 / \sqrt{\sum_s \|v_i^{(s)}\|_2^2}$, its cosine is

$$\bar{z}_i^\top \bar{z}_j = \sum_{r \in \mathcal{D}} \alpha_i^{(r)} \alpha_j^{(r)} u_i^{(r)\top} u_j^{(r)}. \tag{8}$$

Global normalization thus turns block norms into sample-dependent depth weights, whereas Equation 7 keeps the coefficients fixed and inspectable. The two versions turn out to be close on Qwen and apart on Gemma (Section 4.3), which already hints that the multidepth gain has a backbone-dependent geometric component.

Four $d \times 64$ maps match one $d \times 256$ final-depth map in parameters and arithmetic, so the following controls share a budget with RoleLex unless stated otherwise. They replace block normalization with global normalization; share one $d \times 64$ map across the four depths (a quarter of the parameters); apply an unrestricted $4d \times 256$ map to the same four normalized states (four times the parameters); or collapse the states before projection with a learned scalar mixture or a plain mean. Fixed-budget variants also spread the 256 coordinates over eight or sixteen uniformly spaced depths.

3.3 Referent-supervised metric learning

For each training query with a positive, z_i^+ is the latest eligible other-speaker occurrence with the same human referent identity, and eight eligible occurrences with different identities serve as negatives $\{z_{ik}^-\}_{k=1}^8$, mined as the hardest by centered frozen-state cosine at a fixed middle layer. The four projections are trained jointly to minimize

$$\mathcal{L}_i = -\log \frac{\exp(z_i^\top z_i^+/\tau)}{\exp(z_i^\top z_i^+/\tau) + \sum_{k=1}^8 \exp(z_i^\top z_{ik}^-/\tau)}, \qquad \tau = 0.07, \tag{9}$$

with the decoder frozen throughout. We run 12 epochs of AdamW (Loshchilov & Hutter, 2019) at learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, and seed 3. The four projections hold 1,048,576 parameters for Qwen3.5-9B and OLMo 3 7B ($d = 4096$) and 655,360 for Gemma 4 E4B-it ($d = 2560$).

4 Experiments

4.1 Setup and comparisons

We use the frozen Qwen/Qwen3.5-9B (Qwen Team, 2026),[1] google/gemma-4-E4B-it (Gemma Team, 2026),[2] and allenai/Olmo-3-7B-Instruct (Team OLMo, 2025),[3] with end-point states of 4,096, 2,560, and 4,096 dimensions. The four depths are layers 8/16/24/32 for Qwen and OLMo and 10/21/31/42 for Gemma, in each case about a quarter, half, three quarters, and all of the decoder. The released Gemma implementation returns layer 42 after its final RMS normalization; Appendix Table 10 checks that this does not matter.

The central comparison is the fixed-final-depth key: a 256-dimensional projection of the last layer trained with the same loss, negatives, optimizer, and parameter budget. We also train every one- and two-depth subset and let each fold pick the best on its development map, which gives the single- and two-depth baselines an advantage RoleLex does not have. The key width, schedule, temperature, and four-depth default were fixed before the four-seed confirmation runs and then held constant across backbones and datasets. Broader comparisons include dynamic BM25 (Robertson & Zaragoza, 2009), a fully fine-tuned BGE-base bi-encoder over causal mention contexts (Xiao et al., 2024), Qwen3-Embedding-0.6B (Zhang et al., 2025), frozen embedding and coreference systems, prompting, and nonlinear projection heads. Appendix B gives the complete protocol.

4.2 Several depths beat the best single depth

Table 1 holds the main result. Every column uses a 256-dimensional key and the same number of projection parameters; the one- and two-depth columns choose, on each fold's development map, among the four single depths or the six pairs. Reading across the Qwen rows, the fixed four-depth key reaches 90.66 overall, 1.39 points above the best selected single depth and 0.40 above the best selected pair. With Gemma the margins grow to 7.68 and 2.17 points. With OLMo the four-depth key gains 0.66 over the best single depth and trails the best pair by 0.09, inside seed noise. The different-surface rows repeat the ordering. Against the conventional final-layer key the gains are 4.47, 10.66, and 5.85 points, and they are positive on every one of the 16 held-out maps for every backbone (two-sided exact sign test, $p = 3.05 \times 10^{-5}$).

For scale, dynamic BM25 reaches 73.62 overall, so every RoleLex backbone is well clear of lexical matching. The strongest comparison is a BGE-base encoder fully fine-tuned on the same folds and labels, which reaches 90.51; Qwen RoleLex reaches 90.59 in the matched seed-3 run while touching none of the decoder's weights. Section 4.6 returns to what separates these two systems. Appendix Table 17 lists the embedding, prompting, coreference, and nonlinear comparisons.

[1] https://huggingface.co/Qwen/Qwen3.5-9B
[2] https://huggingface.co/google/gemma-4-E4B-it
[3] https://huggingface.co/allenai/Olmo-3-7B-Instruct

Table 1: MapTask-RefMem Recall@1 (percent) with a fixed 256-dimensional key and matched projection budget. Four-seed means ± standard deviations. The one- and two-depth columns are selected per fold on development data; the final-depth and RoleLex columns are fixed. "Different" restricts to queries whose preferred antecedent is worded differently.

Backbone	Slice	Fixed final	Best one depth	Best two depths	RoleLex (four)
Qwen3.5-9B	Overall	86.19 ± 0.19	89.27 ± 0.23	90.26 ± 0.20	$\mathbf{90.66 \pm 0.11}$
	Different	83.48 ± 0.28	86.80 ± 0.34	88.00 ± 0.30	$\mathbf{88.43 \pm 0.13}$
Gemma 4 E4B-it	Overall	74.72 ± 0.14	77.70 ± 0.25	83.21 ± 0.25	$\mathbf{85.38 \pm 0.17}$
	Different	71.22 ± 0.12	73.51 ± 0.23	79.41 ± 0.41	$\mathbf{81.99 \pm 0.23}$
OLMo 3 7B	Overall	81.86 ± 0.26	87.05 ± 0.46	$\mathbf{87.80 \pm 0.08}$	87.71 ± 0.13
	Different	78.27 ± 0.33	83.85 ± 0.57	$\mathbf{84.80 \pm 0.11}$	84.63 ± 0.18

Table 2: Four-seed mean MapTask-RefMem Recall@1 for alternative ways of turning the same four decoder states into one 256-dimensional key. Shared uses one quarter of RoleLex's parameters; dense uses four times as many.

Backbone	Scalar mix	Shared blocks	RoleLex	Global norm	Dense
Qwen3.5-9B	90.55	90.81	90.66	90.62	88.85
Gemma 4 E4B-it	82.23	84.55	85.38	84.20	82.72
OLMo 3 7B	86.12	86.60	87.71	87.87	85.71

4.3 The gain comes from keeping depth scores apart

A four-depth key could help for two reasons: because it keeps the depths' contributions separate, or because of how it parameterizes their combination. Table 2 separates the two. Scalar mixing, which collapses the four states into one before projection, matches RoleLex on Qwen but trails by 3.15 points on Gemma and 1.59 on OLMo. A dense map sees the same four normalized states with four times the parameters and trails on all three backbones. A longer, development-tuned optimization sweep narrows the dense gap to 0.53 points on Qwen and 0.65 on Gemma without closing it.

Within the separable family the picture is less tidy. Sharing one projection across depths edges ahead on Qwen while separate projections win on Gemma and OLMo. Block and global normalization tie on Qwen; block normalization is better on Gemma and global normalization on OLMo. What survives on all three backbones is the additive multidepth score. There is no universal rule for how to share projections or where to normalize.

Why does the more expressive dense map lose? Training dynamics point to an answer. Under the matched schedule, RoleLex uses 203–222 of its 256 output directions effectively, the dense map only 96–123. A lower learning rate restores the dense map's rank but not its held-out score, and on Gemma the dense map reaches a lower training loss while scoring worse on held-out maps, so the deficit is not under-optimization. The gap also widens when training maps are removed: with two maps instead of fourteen, RoleLex leads by 2.04 points on Qwen and 4.14 on Gemma. Meanwhile the four blocks, decoded independently, pick different antecedents for more than half of all queries. Our reading is that the factorization is an inductive bias that lets a data-limited learner hold on to evidence the depths do not share; it does more than delete a harmful term after the fact. Appendix Tables 11, 12, and 13 give the full numbers.

Depth resolution is a separate knob. Qwen changes little between four and eight blocks, Gemma keeps improving with more and narrower blocks, and OLMo slightly prefers two selected depths. We keep four quarter-depths as the model-independent default. On Gemma, a four-seed confirmation reaches 87.08 with eight depths and 88.07 with sixteen, against 85.38 with four, and choosing among 4/8/16 on each fold's development map alone reaches 87.93 overall and 84.97 on different-surface queries (Appendix Table 9).

Table 3: Retrieval-or-abstention on all 13,077 queries (four-seed means, percent). Joint accuracy counts a correct antecedent or a correct abstention. The threshold on maximum cosine is chosen per fold on development data. AUROC is over queries with nonempty eligible memory.

	Joint accuracy		Abstention F1		AUROC	
Backbone	Final	RoleLex	Final	RoleLex	Final	RoleLex
Qwen3.5-9B	76.97	80.94	73.35	75.52	88.11	90.38
Gemma 4 E4B-it	65.68	74.43	65.12	69.46	81.25	85.94
OLMo 3 7B	71.64	76.55	67.45	70.67	83.65	87.37

4.4 The gain lands where wording stops helping

If the extra depths mostly re-encode the surface form, the gain should shrink on queries whose antecedents are worded differently. It does the opposite. The advantage over the final-layer key is larger for pronouns than for lexical noun phrases and grows with antecedent distance; when the preferred antecedent lies 21–50 mentions back, it reaches 7.95 points for Qwen, 24.10 for Gemma, and 11.79 for OLMo. On the 5,919-query no-same-surface slice, where no correct memory mention repeats the query's words, RoleLex lifts Qwen from 81.64 to 86.59, Gemma from 69.10 to 79.35, and OLMo from 75.57 to 82.14. Appendix Tables 20 and 24 give the full stratification and examples, including cases where the added depths overturn a correct final-layer match.

Single-depth probes show where the extra evidence lives. For Qwen the best learned key is at layer 16, midway through the decoder, and for Gemma the middle layers beat the final layer, even though native cosine is strongest at the final layer of both (Appendix Table 16). The depths disagree about the antecedent on 47–68% of decisions. They are not interchangeable versions of one key; they are different views of the same relation.

4.5 Knowing when memory holds no antecedent

The maximum key similarity doubles as an abstention score. For each fold we choose one threshold on the development map and apply it unchanged to the held-out map. Table 3 evaluates all 13,077 queries. RoleLex improves joint retrieval-or-abstention accuracy, abstention F1, and the threshold-free separation between retrievable and non-retrievable queries on every backbone. The AUROC uses the 3,449 abstention cases whose eligible memory is nonempty, so every negative has at least one distractor.

4.6 An adapted encoder, and what a cached key costs

A dedicated retrieval encoder is the demanding comparison, since all of its weights can adapt to the task. Fine-tuning every parameter of BGE-base on the same folds reaches 90.51 overall and 88.29 on different-surface queries; Qwen RoleLex reaches 90.59 and 88.38 in the matched seed-3 run with the decoder frozen. The substrate matters, though. A frozen Qwen3-Embedding-0.6B representation reaches 65.98, and fitting a 256-dimensional metric head on top raises it only to 66.26. A contrastive head on any pretrained embedding does not reproduce the result; the contextual decoder states do. Appendix Tables 19 and 25 add mention-boundary corruption and entity-level consolidation as further checks.

The cost picture depends on whether the decoder is already running. RoleLex compresses a Qwen endpoint from 4,096 dimensions (8,192 FP16 bytes) to 256 (512 bytes); its four projections take 1.05M multiply–accumulates, the same as the final-depth key, and searching 10,000 keys takes 0.033 ms on an A100. When a dialogue decoder is producing the states anyway, the marginal work is projection and search. Standalone, the numbers favor the encoder: on eight transcripts with 644 mentions, one offline Qwen pass takes 1.216 s while BGE-base encodes the 644 causal contexts in 0.863 s, and recomputing each Qwen prefix independently costs 132.36 ms per query. RoleLex is the right tool when the decoder's

Table 4: OneCommon-RefMem Recall@1 (four-seed means). "Competitive different" keeps queries with at least one incorrect candidate and a differently worded preferred antecedent.

	Overall		Competitive different	
Backbone	Final	RoleLex	Final	RoleLex
Qwen3.5-9B	86.29	88.93	78.35	82.80
Gemma 4 E4B-it	82.33	83.83	72.13	74.65
OLMo 3 7B	84.87	88.60	76.01	81.92

states are already there or one transcript pass can be amortized over many queries. Table 7 separates the regimes.

4.7 The construction transfers to OneCommon

OneCommon is a second collaborative-reference domain, with partially observable visual context and evolving referent interpretations (Udagawa & Aizawa, 2019; 2020). We move the architecture, relative depths, key width, loss, and optimizer over unchanged and fit only the projections on OneCommon identities. Across four seeds RoleLex raises Qwen from 86.29 to 88.93, Gemma from 82.33 to 83.83, and OLMo from 84.87 to 88.60 (Table 4); on competitive different-surface cases the gains are 4.45, 2.51, and 5.91 points, and every backbone improves in every seed. The aggregation controls repeat the MapTask pattern in outline: several separable keys are close, and nothing that collapses depth before projection dominates (Appendix Table 22). Projections trained on one dataset transfer weakly to the other (Appendix Table 23), so what carries over is the construction, not one learned metric.

5 Discussion and Limitations

The experiments support a simple picture of how a frozen decoder remembers a mention. There is no single referential key that later layers merely refine. Different depths hold partly different relations, and evidence that has faded from a final-layer metric can still be read from an earlier one. RoleLex keeps these readings in coordinate blocks whose similarities add rather than interfere. The picture also says which details should be expected to travel. Using several depths helped on every model and dataset we tried; how to share projections, how many blocks to use, and where to normalize did not settle on one answer. The dense control sharpens the contrast: cross-depth terms add expressivity, and with limited supervision that expressivity is a cost.

Several limits bound these claims. Mention spans and speaker roles are given, so the benchmark measures referential access rather than mention detection. Both corpora are English task-oriented dialogue with a small set of physical referents, and the training relation comes from human identity labels; a label-efficiency check with generated identities keeps most of the gain (Appendix A), but we have not tested noisier supervision. The backbones span 4–9 billion parameters and three families; we do not know how the depth pattern behaves at other scales. The projections are linear by design, and nonlinear heads in our capacity controls did not help, but a richer family might. Finally, RoleLex is most natural as the retrieval layer of a dialogue model whose states already exist. When every context must be encoded from scratch, a separately adapted encoder is faster and reaches comparable accuracy.

6 Related Work

Reference and entity tracking in dialogue. Neural coreference resolvers score mention–antecedent pairs over contextual spans (Lee et al., 2017); online variants make the causal constraint explicit (Xu & Choi, 2022), and Maverick is a strong encoder-based resolver that accepts given mentions and speaker metadata (Martinelli et al., 2024). Entity-tracking evaluations probe discourse-state updates (Kim & Schuster, 2023), and entity-linking probes have

examined what multi-party dialogue models retain (Aina et al., 2019). Grounded dialogue adds partial observability and negotiated common ground: reference-centric models maintain explicit entity state (Fried et al., 2021), while OneCommon (Udagawa & Aizawa, 2019; 2020) and GMMT (Li et al., 2026a) annotate how referent interpretations evolve. Seeing Is Not Sharing studies interpretation and common-ground bias over the GMMT expressions (Li et al., 2026b); MapTask-RefMem uses the same expressions to study causal retrieval.

Cached keys for retrieval. Bi-encoders make candidate representations reusable. Siamese sentence encoders (Reimers & Gurevych, 2019), dense passage retrieval (Karpukhin et al., 2020), and their descendants (Zhao et al., 2024), including the BGE family (Xiao et al., 2024) and Qwen3 Embedding (Zhang et al., 2025), all obtain candidate-independent keys from a dedicated text encoder. RoleLex obtains them from several depths of a decoder that is already running.

Probes, layers, and multi-layer features. Low-rank linear maps can expose organized structure in frozen representations (Hewitt & Manning, 2019; Hernandez & Andreas, 2021). The closest precedent is the pronoun probe of Sorodoc et al. (2020), which projects states through one affine map and separates same-chain from different-chain pairs by cosine; RoleLex extends projection-plus-cosine from one representation to a depth-factorized key. That different layers serve different tasks is well established (Tenney et al., 2019; Jawahar et al., 2019; Rogers et al., 2020; de Vries et al., 2020), as is the changing geometry of contextual spaces across depth (Ethayarajh, 2019). ELMo's scalar mixtures (Peters et al., 2018) and BERT's concatenated upper layers (Devlin et al., 2019) are the classic ways to use several layers at once, and Xie & Lukasiewicz (2025) build passage-retrieval representations from multiple encoder layers with average, scalar-mixture, or self-contrastive pooling. Those constructions reduce depth to shared coordinates on the document side while the query stays a final-layer vector. Here one key must serve as both query and memory, and Section 4.3 tests the shared-coordinate alternatives directly.

Binding in controlled settings. Feng & Steinhardt (2024) identify binding-ID vectors for entity–attribute associations and confirm their role by intervention; Dai et al. (2024) isolate a low-rank ordering subspace, and later work finds grid-like entity–relation organization in controlled multi-sentence prompts (Dai et al., 2026). These are interventional results in synthetic settings. We examine a downstream consequence in naturally occurring dialogue: whether the referential relation is linearly retrievable from a causal state, and at which depths.

7 Conclusion

A frozen decoder does not keep one memory of a mention; it keeps several, at different depths, and they disagree often enough to matter. RoleLex stores those memories side by side in a single unit vector whose dot product is an average of depth-wise cosines. The construction costs no more than a final-layer key, beats the best layer one could have selected on three decoder families, and carries over to a second dialogue domain without retuning. Depth, in a frozen representation, is worth preserving as well as searching.

AI use statement

In this work, we used generative AI tools to implement the experimental code and to run the experiments described in the paper. We have not used generative AI tools for research ideation, for designing or adjusting the method, or for designing the experimental methodology; these were carried out entirely by the authors. Additionally, we used generative AI tools to search for and summarize related literature, to draft and polish parts of the text, and to design drafts of the figures. We have reviewed all AI-assisted work: all AI-written code was manually reviewed by the authors before use, every reported number was traced to its result file, and all AI-suggested references were checked against the original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

Appendix A describes the GMMT–HCRC alignment, the population counts, and the deterministic fold construction. Appendix B gives extraction layers, optimizer settings, seeds, negative mining, threshold selection, latency measurement, and the OneCommon protocol. The supplement contains the data products, training and evaluation code, configurations, per-fold predictions, and the commands that regenerate every table.

References

Laura Aina, Carina Silberer, Ionut-Teodor Sorodoc, Matthijs Westera, and Gemma Boleda. What do entity-centric models learn? insights from entity linking in multi-party dialogue. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 3772–3783, 2019. doi: 10.18653/v1/N19-1378. URL https://aclanthology.org/N19-1378/.

Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. Cognition, 22(1):1–39, 1986. doi: 10.1016/0010-0277(86)90010-7.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. doi: 10.18653/v1/2024.emnlp-main.967.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Cell-based representation of relational binding in language models. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2194. URL https://aclanthology.org/2026.acl-long.2194/.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In Findings of the Association for Computational Linguistics: EMNLP 2020, 2020. doi: 10.18653/v1/2020.findings-emnlp.389.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019. doi: 10.18653/v1/D19-1006.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In International Conference on Learning Representations, 2024.

Daniel Fried, Justin T. Chiu, and Dan Klein. Reference-centric models for grounded collaborative dialogue. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021. doi: 10.18653/v1/2021.emnlp-main.163.

Gemma Team. Gemma 4 technical report. arXiv, 2026.

Evan Hernandez and Jacob Andreas. The low-dimensional linear geometry of contextualized word representations. In Proceedings of the 25th Conference on Computational Natural Language Learning, 2021. doi: 10.18653/v1/2021.conll-1.7.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3651–3657, 2019. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356/.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.213.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017. doi: 10.18653/v1/D17-1018.

Nan Li, Albert Gatt, and Massimo Poesio. Grounded misunderstandings in asymmetric dialogue: A perspectivist annotation scheme for maptask. In Proceedings of the 2026 Language Resources and Evaluation Conference, 2026a. doi: 10.63317/59anbt78wyj7.

Nan Li, Albert Gatt, and Massimo Poesio. Seeing is not sharing: Some vision-language models overestimate common ground in asymmetric dialogue. SIGDIAL, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. Maverick: Efficient and accurate coreference resolution defying recent trends. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 13380–13394, 2024. doi: 10.18653/v1/2024.acl-long.722. URL https://aclanthology.org/2024.acl-long.722/.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 2227–2237, 2018. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202/.

Qwen Team. Qwen3.5-9B model card. Hugging Face model release, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Nils Reimers and Iryna Gurevych. 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, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 2020. doi: 10.1162/tacl_a_00349.

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. Probing for referential information in language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.384.

Team OLMo. OLMo 3. arXiv, 2025. URL https://arxiv.org/abs/2512.13961.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4593–4601, 2019. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Takuma Udagawa and Akiko Aizawa. A natural language corpus of common grounding under continuous and partially-observable context. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019. doi: 10.1609/aaai.v33i01.33017120.

Takuma Udagawa and Akiko Aizawa. An annotated corpus of reference resolution for interpreting common grounding. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020. doi: 10.1609/AAAI.V34I05.6442.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian yun Nie. C-pack: Packed resources for general chinese embeddings. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024. doi: 10.1145/3626772.3657878.

Zhongbin Xie and Thomas Lukasiewicz. Investigating multi-layer representations for dense passage retrieval. In Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 24522–24536, 2025. doi: 10.18653/v1/2025.findings-emnlp.1333. URL https://aclanthology.org/2025.findings-emnlp.1333/.

Liyan Xu and Jinho D. Choi. Online coreference resolution for dialogue processing: Improving mention-linking on real-time conversations. In Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pp. 350–361, 2022. doi: 10.18653/v1/2022.starsem-1.30. URL https://aclanthology.org/2022.starsem-1.30/.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. arXiv, 2025. URL https://arxiv.org/abs/2506.05176.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. ACM Transactions on Information Systems, 42 (4), 2024. doi: 10.1145/3637870.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. Frequency-based distortions in contextualized word embeddings. arXiv preprint arXiv:2104.08465, 2021.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. Problems with cosine as a measure of embedding similarity for high frequency words. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.45.

A Benchmark Construction and Data Audit

A.1 Alignment, population, and folds

Each released GMMT occurrence is aligned to the original HCRC human reference annotation. Candidate construction applies Equation 1 using the annotated span endpoint, timestamp, and speaker. Human landmark equality determines retrieval eligibility and final scoring. A secondary set of 1,267 timestamp-cutoff pairs audits temporal validity.

Table 5: MapTask-RefMem evaluation population.

Quantity	Count
Dialogues	128
Maps	16
All mention queries	13,077
Retrieval queries	9,088
Abstention queries	3,989
Different-surface retrieval queries	7,031
Official lexical-variant queries	765
Timestamp-cutoff pairs	1,267

For each test map, the stored deterministic SHA256 rule selects a distinct development map; the other fourteen maps form the training partition. All centering, PCA, hard-negative mining, preprocessing statistics, and learned parameters use this partition. Development selection uses the human identities of the development map. Human identities from the test map are accessed only for final scoring.

A.2 Alternative lower-cost supervision

The main experiments use HCRC human referent identities throughout training and development. As a label-efficiency check, we replace those relations with equality between GMMT's GPT-5-generated current-speaker landmark IDs, while retaining the same held-out human evaluation. The generated pair relation has 99.76% precision and 75.19% recall against human identity equality. Despite its missing links, it preserves most of the RoleLex gain.

Table 6: RoleLex with human and model-generated training relations (four-seed mean Recall@1 $\pm$ seed standard deviation, percent). Both conditions use held-out human evaluation.

Backbone	Human overall	Generated overall	Human different	Generated different
Qwen3.5-9B	$\mathbf{90.66 \pm 0.11}$	88.16 ± 0.21	$\mathbf{88.43 \pm 0.13}$	85.32 ± 0.26
Gemma 4 E4B-it	$\mathbf{85.38 \pm 0.17}$	82.57 ± 0.41	$\mathbf{81.99 \pm 0.23}$	78.53 ± 0.45
OLMo 3 7B	$\mathbf{87.71 \pm 0.13}$	84.70 ± 0.09	$\mathbf{84.63 \pm 0.18}$	81.00 ± 0.13

B Complete Experimental Details

B.1 Representation extraction and optimization

The frozen Qwen/Qwen3.5-9B and allenai/Olmo-3-7B-Instruct states have dimension 4,096; the frozen google/gemma-4-E4B-it state has dimension 2,560. Extraction uses each checkpoint's native tokenizer and the hidden state at the final subtoken of the annotated mention in its causal dialogue prefix. Qwen layer 32 and Gemma layer 42 are the states returned after their models' final normalization; in particular, Gemma's reported final block is post-final-RMSNorm. The final-normalization control hooks the input to that normalization and otherwise repeats the full four-seed protocol (Table 10). The raw contextual baseline centers states using training-map statistics and compares them by cosine. Contextual-projection ablations apply non-affine layer normalization, a bias-free $r \times d$ projection, and ℓ_2 normalization for $r \in \{128, 256\}$.

RoleLex sends four contextual states through independent bias-free $64 \times d$ maps. Each projected block is independently normalized and scaled by $1/2$ before concatenation. The resulting 256-dimensional vector is unit normalized by construction. The four projections contain exactly as many parameters and MACs as one $256 \times d$ final-depth projection.

The primary depth set extracts layers 8/16/24/32 for Qwen and OLMo, and 10/21/31/42 for Gemma from the same token endpoints. Primary RoleLex uses all four layers. Fixed-budget granularity controls use eight uniformly spaced layers with width 32 or sixteen uniformly spaced layers with width 16. For the fixed-budget count ablation, every singleton and pair is trained separately; each fold chooses the singleton or pair with maximum retrieval Recall@1 on its development map, with ties broken lexicographically. Output width is divided evenly across active depths, so all counts retain 256 total dimensions. A fixed negative-mining layer—Qwen/OLMo layer 24 or Gemma layer 21—is shared across the input-depth ablation.

Each positive query uses the latest eligible same-identity occurrence. Eight differently labeled hard negatives are selected by centered raw-state cosine. Optimization uses twelve AdamW epochs, learning rate 0.001, weight decay 0.0001, batch size 256, temperature 0.07, and seed 3. The complete stability study uses seeds 3, 7, 11, and 19 for the fixed-final-depth, four-depth, scalar-mixture, global-normalization, dense-concatenation, and development-selected one/two-depth keys over all sixteen folds for Qwen and Gemma. OLMo uses the same four seeds for its primary, aggregation, and development-selected one/two-depth comparisons. Every decoder parameter remains frozen.

Aggregation controls retain the same 256-dimensional output and optimization. The learned-weight block key parameterizes four fixed coefficients with a softmax. The scalar-mixture control combines normalized states with learned softmax weights and applies one $d \times 256$ projection. Raw-mean and normalized-mean controls average the four states before the same projection. The dense control applies one unrestricted $4d \times 256$ projection. Negative mining is tested with the fixed middle depth, the final depth, and uniform eligible negatives.

For retrieval-or-abstention evaluation, confidence is the maximum cosine to an eligible memory key. Each fold selects the threshold that maximizes joint accuracy on its development map and applies it unchanged to the test map. Queries with empty eligible memory abstain directly. Threshold-free AUROC and average precision are computed separately; the reported AUROC excludes empty memories so that every abstention example contains at least one distractor.

The two nonlinear capacity comparisons use the same held-out maps, human referent identities, latest-positive rule, eight hard negatives, 12 epochs, seed 3, AdamW learning rate 0.001, and weight decay 0.0001 as the contextual projection. The two-layer projection applies non-affine layer normalization, a bias-free 4096-to-w linear layer, GELU, and a bias-free w-to-128 linear layer for $w \in \{128, 256, 512, 1024\}$. Its InfoNCE batches contain 256 queries; parameter counts range from 540,672 to 4,325,376.

The pairwise coreference MLP independently normalizes each raw query and candidate state. It concatenates the query, candidate, and their absolute difference before a 12,288-to-w affine layer, GELU, and a w-to-1 affine layer for $w \in \{64, 256, 1024\}$. The MLP scores the same positive-plus-eight-negative candidate sets with cross-entropy and uses batches of 1,024 queries. Width is selected per fold by development Recall@1, with ties preferring fewer parameters.

Other controls are an untrained random rank-128 projection, train-map PCA at rank 128, supervised contextual ranks 32/64/128/256, random versus hard negatives, endpoint shifts, and the contextual projection objective over frozen BGE expression and causal-context embeddings.

The trainable encoder baseline fully updates all 109,482,240 parameters of BAAI/bge-base-en-v1.5. Each mention is encoded independently from the last 4,000 characters of its causal transcript followed by the target expression, left-truncated to 512 wordpieces. It uses the same folds, human identities, latest positive, eight frozen-Qwen-mined hard negatives, and temperature 0.07 as RoleLex. AdamW runs for three epochs with learning rate 2×10^{-5}, weight decay 0.01, batch size 16, and seed 3. The epoch with highest development Recall@1

is evaluated on the held-out test map; ties prefer the earlier epoch. Epochs 1, 2, and 3 are selected in 1, 8, and 7 folds, respectively.

The Qwen3-Embedding-0.6B baseline uses the same 512-token causal contexts, official last-token pooling, and a task-specific query instruction. We report its native 256-dimensional Matryoshka representation and a shared 1024-to-256 linear head trained with the same human-identity tuples and hard negatives as RoleLex. The head starts from the native 256-dimensional slice; each fold selects among that initialization and 12 trained epochs on its development map. Four seeds update 262,144 parameters while the 595.78M-parameter embedding model remains frozen.

For prompting, the frozen instruction checkpoint receives the causal dialogue prefix and current expression and generates an earlier expression used by the other speaker; deterministic normalized-text matching maps the output to the latest eligible mention. Maverick uses the official sapienzanlp/maverick-mes-ontonotes checkpoint, the annotated mention spans and speaker sequence, and an online window capped at 500 subword tokens. The most recent eligible mention in the query's predicted cluster is returned.

B.2 Measures, thresholds, and latency

For retrieval set R, micro Recall@1 is

$$\mathrm{R@1}_{\mathrm{micro}} = \frac{1}{|R|} \sum_{i \in R} \mathbf{1}[r_{\hat{j}_i} = r_i]. \tag{10}$$

Map-macro Recall@1 is the unweighted mean of the corresponding measure over the sixteen held-out maps. Latency is synchronized model execution on one NVIDIA A100-SXM4-80GB, measured from tokenized model inputs. The fixed timing subset comprises the first eight dialogues, 644 annotated mentions, and 434 retrieval queries. Independent queries and prompting use batch size 8. Offline indexing is the median of three complete transcript passes; prompt decoding is total generation time minus a separately timed prefill pass.

Table 7: Memory and latency measurements. Projection and search are post-encoding marginal costs.

Quantity or operation	Value
Raw Qwen key	4,096 dimensions, 8,192 FP16 bytes
RoleLex key	256 dimensions, 512 FP16 bytes
Form one Qwen RoleLex key from captured states	0.488 ms
Form 2,048 Qwen RoleLex keys	1.647 ms
Search 10,000 RoleLex keys	0.033 ms
Qwen four depth projections	1,048,576 multiply-accumulates
Qwen decoder only, 8 transcripts	1.216 s
Qwen + final/four-depth projection	1.223 / 1.222 s
Gemma decoder only, 8 transcripts	0.953 s
Gemma + final/four-depth projection	0.956 / 0.961 s
Four-depth marginal, Qwen/Gemma	0.0095 / 0.0113 ms per mention
BGE-base, 644 causal mention contexts	0.863 s (1.340 ms/mention)
Independent prefix state, batch 8	132.357 ms/query

The BGE measurement uses its 512-token limit and batches of eight; the 644 contexts average 428.39 wordpieces. Fine-tuning changes the weights but not the measured execution graph. On the same subset, independently processed decoder prefixes average 1,003 input tokens.

B.3 OneCommon protocol

The released OneCommon annotations contain 41,492 markables in 5,191 dialogues. We retain 24,596 exact-span, clean, unambiguous singleton mentions. Agent-specific referent IDs are mapped to the underlying physical-dot ID, and eligible memory follows Equation 1.

The release-compatible seed-1 80/10/10 dialogue split contains 4,152/519/520 train/development/test dialogues and 7,057/909/917 retrievable queries. Train–test

and development–test scenario overlap are both zero. There are 4,644 training queries with a positive and a negative. Test queries have 2.39 candidates on average and at most 10. Among them, 866 have a differently worded preferred antecedent, 597 have at least one incorrect candidate, and 120 belong to the competitive pronoun-like slice. The development split is unused.

The frozen backbones, four fixed relative depths, four 64-dimensional blocks, normalization, eight negatives, temperature, optimizer, batch size, and epoch count match MapTask. The fixed-final-depth 256-dimensional key is the matched comparison. OneCommon seeds are 3, 11, 29, and 47. Its training split uses released human referent IDs, with the architecture transferred directly.

B.4 Entity-memory statistics

MapTask provides 9,088 entity-memory queries. Of these, 6,556 have at least two target mentions in eligible memory and 6,041 have at least two target surface forms. The eligible memory contains 34.46 mentions and 7.93 entities per query on average. OneCommon provides 917 queries. Among them, 168 have multiple target mentions and 153 have multiple target surface forms. Mention-max over entity groups is decision-equivalent to the original mention-retrieval rule.

C Additional Results

C.1 Fixed-budget architecture controls

Table 8: Seed-3 depth-count ablation on MapTask-RefMem (Recall@1, percent). All rows use a 256-dimensional key and the same number of projection parameters. One- and two-depth subsets are selected on each fold's development map; the final-depth and four-depth rows are fixed.

Backbone	Input depths	Overall	Different	Same
Qwen3.5-9B	Final only	86.27	83.63	95.28
	One, dev-selected	89.52	87.21	97.42
	Two, dev-selected	90.54	88.44	97.72
	Four, fixed (RoleLex)	90.59	88.38	98.15
Gemma 4 E4B-it	Final only	74.63	71.07	86.78
	One, dev-selected	77.89	73.53	92.81
	Two, dev-selected	83.55	79.93	95.92
	Four, fixed (RoleLex)	85.18	81.77	96.84
OLMo 3 7B	Final only	81.64	77.95	94.21
	One, dev-selected	87.00	83.79	98.01
	Two, dev-selected	87.71	84.71	97.96
	Four, fixed (RoleLex)	87.51	84.41	98.10

Across the 64 Qwen seed–fold selections, the best singleton is layer 16/24/32 in 45/18/1 cases; the two most frequent pairs are 16–32 (38) and 16–24 (10). Gemma selects layer 10/21/31/42 in 22/10/29/3 cases, with 10–31 (41) and 10–42 (11) the most frequent pairs. OLMo selects layer 16 in 57 of 64 seed–folds; its most frequent pairs are 16–32 (39) and 16–24 (14).

Across 64 Gemma seed–fold decisions, development data selects sixteen depths 49 times and eight depths 15 times; it never selects four. This selection uses no held-out-map labels.

The final normalization does not create the multi-depth gain. With pre-norm states the RoleLex advantage is unchanged for Qwen and slightly larger for Gemma.

Table 9: Depth granularity at a fixed 256-dimensional key and matched projection budget (Recall@1, percent). Qwen rows are seed-3 diagnostics; Gemma rows are four-seed means ± standard deviations. "Selected" chooses among 4/8/16 only on each fold's development map.

Backbone	Depths	Width/depth	Parameters	Overall	Different
Qwen3.5-9B	4	64	1.05M	90.59	88.38
	8	32	1.05M	90.80	88.47
	16	16	1.05M	90.27	87.80
Gemma 4 E4B-it	4	64	0.66M	85.38 ± 0.17	81.99 ± 0.23
	8	32	0.66M	87.08 ± 0.13	83.92 ± 0.20
	16	16	0.66M	$\mathbf{88.07 \pm 0.07}$	$\mathbf{85.12 \pm 0.10}$
	Selected	$256/R$	0.66M	87.93 ± 0.10	84.97 ± 0.12

Table 10: Final-normalization control (four-seed mean Recall@1, percent). "Pre" hooks the input to the decoder's learned final normalization; RoleLex changes only its last input block.

Backbone	Post final	Pre final	Post RoleLex	Pre RoleLex
Qwen3.5-9B	86.19	86.52	90.66	90.67
Gemma 4 E4B-it	74.72	74.31	85.38	85.47

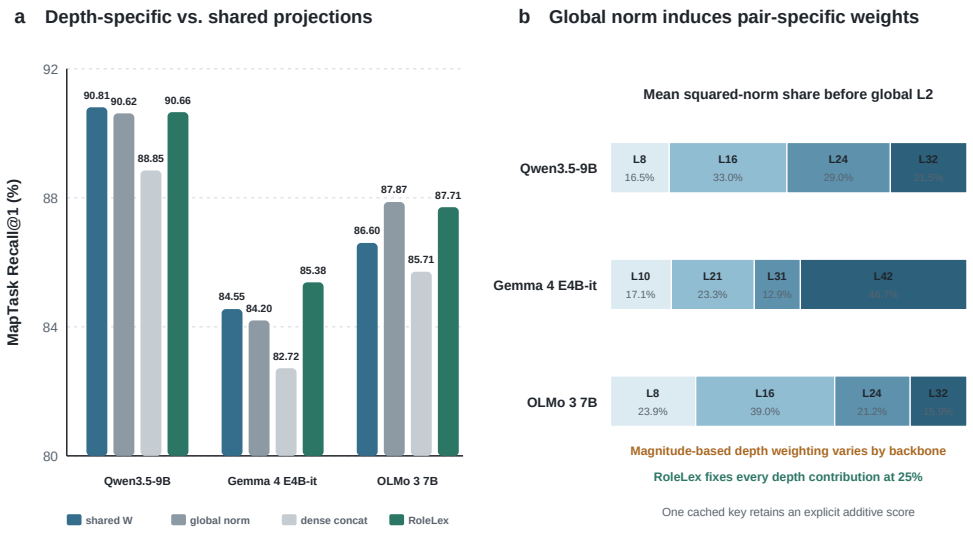


Figure 2: Left: four-seed means for RoleLex and shared-projection, dense, and global-normalized controls. Shared projection uses one quarter of RoleLex's parameters; dense concatenation uses four times as many. Right: projected energy shares under global normalization, where block magnitude induces an input-dependent depth weight.

The learned fixed-weight means are 23.8/27.9/24.0/24.3% for Qwen and 22.8/23.3/25.9/27.9% for Gemma. Relative to dense concatenation, RoleLex improves seed-averaged map-macro Recall@1 by 1.84/2.81/2.00 points for Qwen/Gemma/OLMo, with positive deltas on all 16 maps. Relative to scalar mixing, the corresponding differences are 0.10/3.15/1.59 points. Global normalization differs from RoleLex by $-0.03/-1.18/+0.16$ points: effectively tied on Qwen, worse on Gemma, and slightly better on OLMo. Sharing the projection across depth changes the corresponding scores by $+0.15/-0.83/-1.11$ points. This control separates the value of the blockwise multi-depth score from the additional capacity to learn a distinct map at each depth.

Table 11: Four-depth aggregation controls (Recall@1, percent). RoleLex, shared projection, scalar mixture, dense concatenation, and global normalization report four-seed means ± standard deviations; learned weights and state means are seed-3 diagnostics. All rows use one 256-dimensional cached key.

Backbone	Aggregation	Parameters	Overall	Different	Same
Qwen3.5-9B	RoleLex, uniform blocks	1.05M	90.66 ± 0.11	88.43 ± 0.13	98.27 ± 0.08
	Shared projection, separate blocks	0.26M	90.81 ± 0.30	88.58 ± 0.38	98.42 ± 0.08
	RoleLex, learned fixed weights	1.05M	90.64	88.39	98.30
	Scalar mixture → projection	1.05M	90.55 ± 0.12	88.28 ± 0.15	98.30 ± 0.03
	Normalized-state mean → projection	1.05M	90.74	88.52	98.30
	Raw-state mean → projection	1.05M	90.02	87.74	97.81
	Dense concatenation	4.19M	88.85 ± 0.10	86.26 ± 0.09	97.69 ± 0.13
	Block diagonal, global norm	1.05M	90.62 ± 0.20	88.34 ± 0.26	98.38 ± 0.02
Gemma 4 E4B-it	RoleLex, uniform blocks	0.66M	85.38 ± 0.17	81.99 ± 0.23	96.96 ± 0.18
	Shared projection, separate blocks	0.16M	84.55 ± 0.32	80.86 ± 0.38	97.17 ± 0.23
	RoleLex, learned fixed weights	0.66M	85.12	81.72	96.74
	Scalar mixture → projection	0.66M	82.23 ± 0.20	78.47 ± 0.21	95.11 ± 0.26
	Normalized-state mean → projection	0.66M	82.05	78.21	95.19
	Raw-state mean → projection	0.66M	80.95	77.24	93.63
	Dense concatenation	2.62M	82.72 ± 0.13	79.18 ± 0.14	94.81 ± 0.26
	Block diagonal, global norm	0.66M	84.20 ± 0.13	80.66 ± 0.13	96.31 ± 0.28
OLMo 3 7B	RoleLex, uniform blocks	1.05M	87.71 ± 0.13	84.63 ± 0.18	98.23 ± 0.17
	Shared projection, separate blocks	0.26M	86.60 ± 0.11	83.25 ± 0.17	98.07 ± 0.18
	Scalar mixture → projection	1.05M	86.12 ± 0.13	82.68 ± 0.14	97.87 ± 0.14
	Dense concatenation	4.19M	85.71 ± 0.12	82.31 ± 0.14	97.31 ± 0.11
	Block diagonal, global norm	1.05M	87.87 ± 0.22	84.81 ± 0.27	98.32 ± 0.16

A separate dense-control optimization sweep selects among two learning rates, three weight decays, and checkpoints through 24 epochs using only each fold's development map. It reaches 90.06 overall and 87.67 different-surface Recall@1 for Qwen, and 84.53/81.06 for Gemma, versus 90.59/88.38 and 85.18/81.77 for seed-3 RoleLex.

Factorization mechanism. Let $\tilde{h}$ denote a normalized state and partition an unrestricted dense map as $A = [A_1, \ldots, A_R]$. Its unnormalized kernel is

$$\left(\sum_r A_r \tilde{h}_i^{(r)}\right)^\top \left(\sum_s A_s \tilde{h}_j^{(s)}\right) = \sum_{r,s} \tilde{h}_i^{(r)\top} A_r^\top A_s \tilde{h}_j^{(s)}. \tag{11}$$

RoleLex removes the $r \neq s$ terms by construction and independently normalizes the retained diagonal blocks before addition. Table 12 analyzes all four seeds and sixteen folds. Effective rank is $\exp(H(p))$, where p is the normalized squared singular-value spectrum of the full $256 \times 4d$ map; the four RoleLex maps form one block-diagonal operator for this calculation. Although both operators have at most 256 output directions, RoleLex uses 203–222 effectively, while the dense alternative uses 96–123. Gemma is especially diagnostic: dense concatenation reaches the lower training loss but the worse held-out score, separating its deficit from simple under-optimization.

Table 12: Projection mechanism over 64 fold–seed runs per backbone. Dev and test are map-macro Recall@1; R/D denotes RoleLex/dense concatenation. Effective rank is computed from each learned projection and then averaged.

Backbone	Train InfoNCE R/D	Dev R/D	Test R/D	Effective rank R/D
Qwen3.5-9B	.0125/.0124	90.16/88.86	90.81/88.97	219/96
Gemma 4 E4B-it	.0156/.0107	85.24/82.55	85.47/82.66	203/96
OLMo 3 7B	.0123/.0124	87.66/86.01	87.67/85.67	222/123

The longer development-tuned dense sweep restores effective rank to 232 for Qwen and 226 for Gemma, but not the held-out score. Its map-macro Recall@1 reaches 90.27 and 84.45, still below RoleLex at 90.81 and 85.47; the corresponding micro comparisons are 90.06 versus 90.59 and 84.53 versus 85.18. Rapid rank collapse explains part of the matched-budget gap; a held-out difference remains even after dense optimization recovers full-rank use of the key.

The four independently decoded blocks choose different antecedents for 46.8–56.9% of query pairs for Qwen, 59.2–67.7% for Gemma, and 48.3–61.8% for OLMo. Each block is uniquely

correct on a nonzero share of queries (1.0–4.2% across backbone–depth combinations), so the selected depths carry decisions the others do not. The dense metric assigns 72.3–73.8% of its squared metric energy to off-diagonal depth blocks. Deleting those terms after training does not repair the model. The benefit comes from learning with independent coordinates and normalization from the start; suppressing a particular dense interaction after training does not recover it.

Table 13: Data efficiency under deterministic training-map subsampling (seed 3, Recall@1 percent). Reduced-data dense uses its stronger lower-rate schedule with development selection through 24 epochs; the 14-map row uses the full development-selected sweep. RoleLex uses its fixed 12-epoch schedule.

Backbone	Train maps	RoleLex	Dense	Δ	Different Δ
Qwen3.5-9B	2	77.04	75.00	+2.04	+2.19
	4	83.36	81.89	+1.47	+1.46
	8	87.60	87.09	+0.51	+0.48
	14	90.59	90.06	+0.53	+0.71
Gemma 4 E4B-it	2	73.13	68.99	+4.14	+4.00
	4	78.18	75.94	+2.24	+2.36
	8	82.67	81.51	+1.16	+1.11
	14	85.18	84.53	+0.65	+0.71

Table 14: Negative construction sensitivity for four-depth RoleLex (Recall@1, percent). Middle-depth hard mining is the primary protocol.

Backbone	Negatives	Overall	Different	Same
Qwen3.5-9B	Middle-depth hard	90.59	88.38	98.15
	Final-depth hard	90.18	87.74	98.54
	Random eligible	89.80	87.26	98.49
Gemma 4 E4B-it	Middle-depth hard	85.18	81.77	96.84
	Final-depth hard	84.63	80.96	97.18
	Random eligible	84.38	80.66	97.08
OLMo 3 7B	Middle-depth hard	87.51	84.41	98.10
	Final-depth hard	87.63	84.50	98.35
	Random eligible	86.56	83.12	98.35

Changing the mining source moves overall Recall@1 by at most 0.80 point for Qwen and Gemma and 1.07 points for OLMo.

Table 15: Four-seed stability across all 16 leave-one-map-out folds (percent). Values are seed means ± seed standard deviations.

Backbone	Key	Overall	Different	Same
Qwen3.5-9B	Fixed final depth	86.19 ± 0.19	83.48 ± 0.28	95.43 ± 0.23
	RoleLex	$\mathbf{90.66 \pm 0.11}$	$\mathbf{88.43 \pm 0.13}$	$\mathbf{98.27 \pm 0.08}$
Gemma 4 E4B-it	Fixed final depth	74.72 ± 0.14	71.22 ± 0.12	86.67 ± 0.32
	RoleLex	$\mathbf{85.38 \pm 0.17}$	$\mathbf{81.99 \pm 0.23}$	$\mathbf{96.96 \pm 0.18}$
OLMo 3 7B	Fixed final depth	81.86 ± 0.26	78.27 ± 0.33	94.12 ± 0.27
	RoleLex	$\mathbf{87.71 \pm 0.13}$	$\mathbf{84.63 \pm 0.18}$	$\mathbf{98.23 \pm 0.17}$

C.2 Layer and representation controls

Applying RoleLex to the mean state over the annotated span reaches 91.18 for Qwen and 84.04 for Gemma. We retain endpoint pooling because it is the stronger cross-backbone default and is available as soon as a mention ends. Full BGE encoder adaptation reaches 90.51.

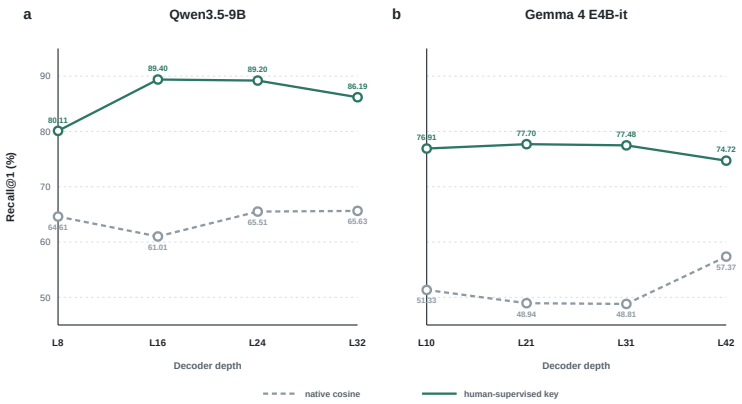


Figure 3: Decoder-depth profiles for native centered cosine and a supervised 256-dimensional single-depth key. The trained metric peaks before the final layer for both backbones.

Table 16: Single-depth analysis (Recall@1, percent). This diagnostic uses a 256-dimensional projection at each tested layer.

Backbone	Layer	Relative depth	Native cosine	Trained key	Gain
Qwen3.5-9B	8	25%	64.61	80.11	15.50
	16	50%	61.01	89.40	28.39
	24	75%	65.51	89.20	23.69
	32	100%	65.63	86.19	20.56
Gemma 4 E4B-it	10	24%	51.33	76.91	25.58
	21	50%	48.94	77.70	28.76
	31	74%	48.81	77.48	28.67
	42	100%	57.37	74.72	17.35

C.3 Retrieval strata

We stratify the held-out predictions by surface relation, expression type, antecedent distance, and eligible-memory size. The largest gains occur for distant antecedents and large candidate sets.

Table 17: Complete Qwen comparison on MapTask-RefMem (percent). Updates denotes trainable parameters; BGE updates its full encoder and the other encoders are frozen.

Representation	Updates	Overall	Different
Exact surface	0	34.94	0.00
Token Jaccard	0	63.68	0.00
Dynamic BM25	0	73.62	65.98
BGE expression	0	65.43	57.87
BGE causal context	0	61.32	60.40
Qwen3-Embedding-0.6B, frozen 256d	0	65.98	64.64
Context-free Qwen	0	60.15	50.90
Centered contextual Qwen	0	65.63	59.34
Maverick coreference	0	67.48	60.42
Prompting	0	65.77	59.44
Two-layer GELU, dev width	0.54–4.33M	83.14	80.23
Pairwise MLP, dev width	0.79–12.58M	74.30	70.67
Qwen3-Embedding metric head	0.26M	66.26	64.88
Fixed final depth, 256d	1.05M	86.27	83.63
RoleLex, four depths	1.05M	90.59	88.38
BGE-base, full fine-tune	109.48M	90.51	88.29

Table 18: Capacity and metric controls for the Qwen final-layer key (Recall@1, percent). The supervised rank sweep uses the canonical layer-24 hard-negative sets shared with the primary fixed-final comparison. Nonlinear widths are selected on the development map.

Method	Width	Overall	Different	Method	Width	Overall
Random projection	128	58.13	–	Supervised linear	32	82.82
Train-map PCA	128	54.97	–	Supervised linear	64	84.71
Two-layer GELU	dev	83.14	80.23	Supervised linear	128	85.78
Pairwise MLP	dev	74.30	70.67	Supervised linear	256	86.26

Table 19: Mention-boundary robustness (Recall@1, percent). Models are trained on exact endpoints; deterministic noise shifts 10% or 20% of test query and memory endpoints by one token.

Backbone	Test endpoints	Fixed final	RoleLex	Δ
Qwen3.5-9B	Exact	86.27	90.59	+4.32
	10% shifted	81.73	87.86	+6.13
	20% shifted	77.92	84.89	+6.98
	All shifted right	41.88	58.00	+16.12
Gemma 4 E4B-it	Exact	74.63	85.18	+10.55
	10% shifted	70.90	81.06	+10.17
	20% shifted	67.30	77.23	+9.94
	All shifted right	39.61	52.50	+12.89

Table 20: Final RoleLex stratification on MapTask-RefMem (percent). Final is the matched fixed-final-depth key and Δ is RoleLex minus Final.

		Qwen3.5-9B			Gemma 4 E4B-it		
Slice	Count	Final	RoleLex	Δ	Final	RoleLex	Δ
Different surface	7,031	83.63	88.38	+4.75	71.07	81.77	+10.70
No same-surface gold	5,919	81.64	86.59	+4.95	69.10	79.35	+10.25
Same surface	2,057	95.28	98.15	+2.87	86.78	96.84	+10.06
Pronoun	1,862	78.68	83.78	+5.10	69.66	76.05	+6.39
Demonstrative NP	235	77.02	81.70	+4.68	67.23	74.04	+6.81
Lexical NP	6,991	88.60	92.70	+4.11	76.20	87.98	+11.79
Distance 1–5	6,843	88.70	91.93	+3.23	80.58	87.91	+7.34
Distance 6–20	1,731	79.66	87.35	+7.68	60.31	78.57	+18.26
Distance 21–50	390	77.18	85.13	+7.95	48.21	72.31	+24.10
Distance ≥ 51	124	72.58	79.03	+6.45	29.03	66.94	+37.90
1–4 candidates	642	93.15	97.04	+3.89	90.19	93.30	+3.12
5–9 candidates	970	86.91	90.52	+3.61	77.53	85.15	+7.63
10–19 candidates	1,887	86.91	91.79	+4.88	75.62	85.59	+9.96
≥ 20 candidates	5,589	85.15	89.46	+4.31	72.00	84.11	+12.11

C.4 OneCommon

Table 21: OneCommon-RefMem results (four-seed means, percent). Comp. retains queries with at least one incorrect eligible candidate.

Backbone	System	Overall	Different	Comp.	Comp. different	Dialogue macro
–	Dynamic BM25	77.86	76.79	66.00	64.55	80.00
Qwen3.5-9B	Fixed final depth	86.29	85.83	78.94	78.35	88.64
	RoleLex	88.93	88.74	83.00	82.80	91.47
Gemma 4 E4B-it	Fixed final depth	82.33	81.76	72.86	72.13	84.95
	RoleLex	83.83	83.40	75.17	74.65	87.11
OLMo 3 7B	Fixed final depth	84.87	84.30	76.76	76.01	87.95
	RoleLex	88.60	88.16	82.50	81.92	91.80

The test set contains 917 retrieval queries. Of these, 866 have a differently worded preferred antecedent and 597 have at least one incorrect candidate. Across every backbone and all four seeds, RoleLex improves overall, different-surface, competitive, competitive-different, and dialogue-macro Recall@1.

Table 22: OneCommon aggregation controls (four-seed mean Recall@1, percent). Every method emits one 256-dimensional key. Shared applies one projection at all depths; dense is an unrestricted projection of concatenated states. O/D denote overall/different-surface retrieval.

Backbone	Slice	RoleLex	Shared	Scalar mix	Global norm	Dense
Qwen3.5-9B	O	88.93	88.71	88.90	88.85	88.66
	D	88.74	88.51	88.74	88.91	88.39
Gemma 4 E4B-it	O	83.83	83.26	84.43	83.51	83.48
	D	83.40	82.79	83.92	83.00	83.00
OLMo 3 7B	O	88.60	87.76	87.98	88.39	88.22
	D	88.16	87.38	87.59	88.02	87.79

Table 23: Cross-dataset zero-shot transfer of contextual-only RoleLex (four-seed means, percent). Training uses no labels or tuning from the target dataset.

Train → test	Backbone	Overall	Different	Same/competitive
MapTask → OneCommon	Qwen3.5-9B	75.44	74.74	62.27
	Gemma 4 E4B-it	74.02	73.04	60.09
OneCommon → MapTask	Qwen3.5-9B	60.18	52.56	86.19
	Gemma 4 E4B-it	55.36	47.46	82.38

C.5 Qualitative retrievals

The examples below include changed-surface rescues and cases where the added depth evidence overturns an otherwise correct final-depth match.

Table 24: Held-out MapTask examples. Each row gives the query, a human-linked earlier expression, and the two retrieved expressions.

Query	Human-linked expression	Fixed-final retrieval	RoleLex retrieval	Outcome
the start position	start position	the camera shop	start position	rescue
the camera shop	the camera shop's	the start position	the camera shop's	rescue
the telephone box	the telephone telephone box	the telephone telephone box	the alpine garden	regression
right	your yacht club	your yacht club	right	regression

C.6 Entity-memory geometry

For candidate keys $\mathcal{K}_{i,e}$ sharing evaluation identity e, the diagnostic forms a spherical centroid

$$c_{i,e} = \frac{\sum_{j \in \mathcal{K}_{i,e}} z_j}{\left\| \sum_{j \in \mathcal{K}_{i,e}} z_j \right\|_2}, \qquad \hat{e}_i = \arg\max_e z_i^\top c_{i,e}. \tag{12}$$

The grouping is used only for this analysis.

Table 25: Entity-memory geometry. Mention-max is ordinary retrieval after collapsing mentions with the same evaluation identity; centroid uses Equation 12.

Dataset	Backbone	Final mention-max	RoleLex mention-max	Final centroid	RoleLex centroid
MapTask	Qwen3.5-9B	86.27	90.59	87.54	91.85
MapTask	Gemma 4 E4B-it	74.64	85.18	76.20	85.68
MapTask	OLMo 3 7B	81.64	87.51	82.89	88.59
OneCommon	Qwen3.5-9B	86.29	88.93	86.12	89.12
OneCommon	Gemma 4 E4B-it	82.33	83.83	82.31	84.22

On MapTask, RoleLex raises target-versus-distractor centroid margin from 0.244 to 0.306 for Qwen, from 0.144 to 0.223 for Gemma, and from 0.211 to 0.276 for OLMo. OneCommon RoleLex target concentration is 0.968 for Qwen and 0.952 for Gemma, with centroid margins 0.558 and 0.489.