COMPUTATION OVER GEOMETRY: MEANING IDENTITY IS COMPUTED, NOT SHIPPED IN THE EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Meaning identity—whether two sentences say the same thing after the wording has changed—is treated throughout retrieval and RAG as a geometric fact about independently encoded sentence vectors. We test a different picture: identity is computed when both sentences occupy the same forward pass, not shipped as a cheaply readable property of either sentence's embedding. On a common 2,000-pair English PAWS test set, linear readouts of four frozen joint passes reach AUC 0.900–0.946 with 1,200 labels, versus 0.622–0.646 for full-width nonlinear readers of independent states. With all 49,258 training pairs, those readers improve to 0.736–0.786, and cross-attention over frozen token sequences reaches 0.868 on a 14B backbone, while the joint readouts remain 0.922–0.960. Across a separate overlap-matched survey of 30 decoder checkpoints, joint exceeds an independently selected product/difference probe in 29 rows and partner shuffle collapses to chance; a pair-disjoint replication reverses the sole GPT-2-large ordering, although its intervals overlap. The same joint-over-independent pattern recurs in bidirectional encoders, encoder–decoders, cross-language and nonce controls, and controlled role, negation, and quantity relations. It is not the generic consequence of scoring jointly: BGE rerankers enter the identity band, while MS-MARCO and Jina rerankers do not. Nor is it an immutable architectural limit: fine-tuning BGE or applying rank-16 LoRA to E5-Mistral can install a strong independent criterion by changing the encoder. The sharper result is route-dependent teachability. A 1.5B joint student and a trainable independent student receive the same 9,000 teacher-scored pairs and update budget, yet reach English PAWS AUC 0.942 and 0.760; six independently trained models also produce strongly concordant, symmetric, and graded joint scores. Across the tested models, readers, and supervision budgets, sentence geometry remains useful for aboutness, but propositional identity behaves as a recurring relation made cheap to read by computation over the pair.

1 INTRODUCTION

A sentence embedding is an invitation to treat meaning as a point. Once each sentence has a vector, "same meaning" becomes cosine, and the industrial stack—dense retrieval, duplicate-question detection, RAG—is built on that geometry (Reimers & Gurevych, 2019; Muennighoff et al., 2023; Gao et al., 2024). The invitation is so familiar that it is easy to miss the empirical claim it encodes: that identity of meaning is a property of a single sentence's representation, recoverable by comparing two such properties.

That claim can be false even if embeddings are useful. They can rank topical neighbours, cluster documents, and win STS (Cer et al., 2017) while still being nearly blind to the distinction PAWS was built to isolate: two sentences that share almost all of their words, one of which is a paraphrase and one of which is not (Zhang et al., 2019; Yang et al., 2019). If identity is not readily accessible in the vectors a system ships, then encode-first systems are looking in the wrong place for a large class of hard pairs.

We separate two pictures, on identical items, with a metric whose chance level is exactly 0.5:

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.

Table 1: English confirm AUC on the overlap-matched family survey. Every readout in a row scores the same n confirm pairs; *each* readout (including cosine) is reported at the layer that maximises its own development AUC, so no column is read at a layer chosen for another. Late fusion and shuffle are the two falsifiers of a computational reading; Ov. is overlap-only and chance $= 0.5$. Qwen2.5-32B joint 95% CI is $[0.884, 0.964]$; 14B is $[0.925, 0.971]$. Older / smaller families are in Table 4, Appendix E.

Model	n	joint layer	Ov.	Cos	Sep	Late	**Joint**	Shuf
Qwen2.5-7B-Inst	250	L21	.457	.569	.558	.517	**.961**	.509
Qwen2.5-14B	250	L36	.533	.584	.629	.548	**.950**	.496
Qwen2.5-1.5B-Inst	500	L21	.527	.585	.654	.499	**.947**	.518
Qwen2.5-7B	500	L21	.554	.556	.572	.516	**.941**	.513
Qwen2.5-3B	400	L27	.468	.545	.568	.555	**.937**	.469
Qwen2.5-32B	150	L48	.495	.609	.551	.532	**.927**	.488
Qwen2.5-1.5B	500	L21	.486	.598	.564	.521	**.916**	.506
Mistral-7B-v0.3	300	L16	.506	.556	.563	.508	**.915**	.489
Llama-3-8B	300	L16	.551	.530	.514	.487	**.897**	.542
Phi-2	300	L16	.501	.613	.573	.531	**.891**	.521
GPT-J-6B	250	L14	.480	.567	.508	.510	**.868**	.500
Qwen2.5-0.5B-Inst	500	L18	.486	.598	.622	.513	**.863**	.495
SmolLM2-1.7B	300	L18	.518	.521	.597	.495	**.843**	.502
OPT-6.7B	250	L24	.547	.625	.612	.561	**.829**	.466
GPT-Neo-1.3B	300	L12	.470	.591	.611	.550	**.815**	.486
Qwen2.5-0.5B	500	L18	.457	.560	.589	.589	**.789**	.507
TinyLlama-1.1B	400	L16	.515	.622	.641	.533	**.778**	.441
Phi-1.5	400	L12	.487	.577	.565	.524	**.778**	.506
GPT-2 XL	300	L24	.483	.492	.502	.517	**.756**	.470
SmolLM2-360M	400	L24	.524	.622	.604	.522	**.753**	.506

Geometry. Each sentence carries its meaning identity in its representation. Cosine, a pair probe, or late fusion of the two vectors should recover that identity with comparable supervision.

Computation. Identity is a relation the network computes when both sentences occupy the same context. It is then readable from the last token of a concatenated forward pass, while recovering it from separately encoded states requires substantially more supervision or a change to the encoder.

The pictures are not rhetorical. Late fusion gives a linear probe both vectors and no cross-attention, so it tests whether the geometry is merely misaligned with cosine. A full-width nonlinear reader and token-level cross-attention test whether richer functions of frozen independent states recover the relation. A shuffled-partner joint pass preserves format, language, and the marginal of sentence B while destroying the pairing; a probe that reads only the template would survive it.

We find computation over geometry across modern open families (Table 1, Figure 1). On overlap-matched English PAWS-X, Llama 3 8B, Mistral 7B, and Qwen2.5 from 1.5B to 32B yield joint confirm AUC 0.90–0.96 at development-selected depths, while separately encoded cosine, product/difference probes, and late fusion remain in 0.49–0.65; partner shuffle returns to ≈ 0.50. The GPT-2 lineage is weaker: large is the survey's sole row where joint does not exceed the independent product/difference probe (0.647 vs. 0.650), although a fixed pair-disjoint replication reverses that ordering; XL reaches 0.756. Outside that stack the same protocol holds for Phi-2, GPT-J-6B, OPT-6.7B, GPT-Neo, Pythia, and SmolLM2 (joint 0.75–0.90), for mean-pooled DeBERTa/RoBERTa encoders (DeBERTa-v3 0.92, nli-DeBERTa 0.94), and for encoder–decoders (Flan-T5-base 0.95; T5-small / BART-base 0.64–0.66) (Radford et al., 2019; Biderman et al., 2023; Black et al., 2021; Dubey et al., 2024; Jiang et al., 2023; Qwen Team, 2025). A nonce-substitution control keeps joint above late fusion with shuffle near chance on Qwen and on non-Qwen causals (Appendix K). Qwen base models can also *say* the answer (zero-shot verdict AUC 0.92–0.99 from 3B to 32B); Llama 3 and Mistral compute the relation at 0.90 while their verbal verdict stays ≈ 0.61.

Two further findings sharpen the result from a route gap toward a recurring relation operator. First, six independently trained models (Qwen, Llama, Mistral) produce strongly agreeing joint scores, including within each gold-label class, and the relation is approximately symmetric and graded (§7). Second, the operator is teachable: under matched teacher supervision and update counts, a 1.5B joint student approaches its teacher whereas a trainable independent route remains far below it on PAWS (§7).

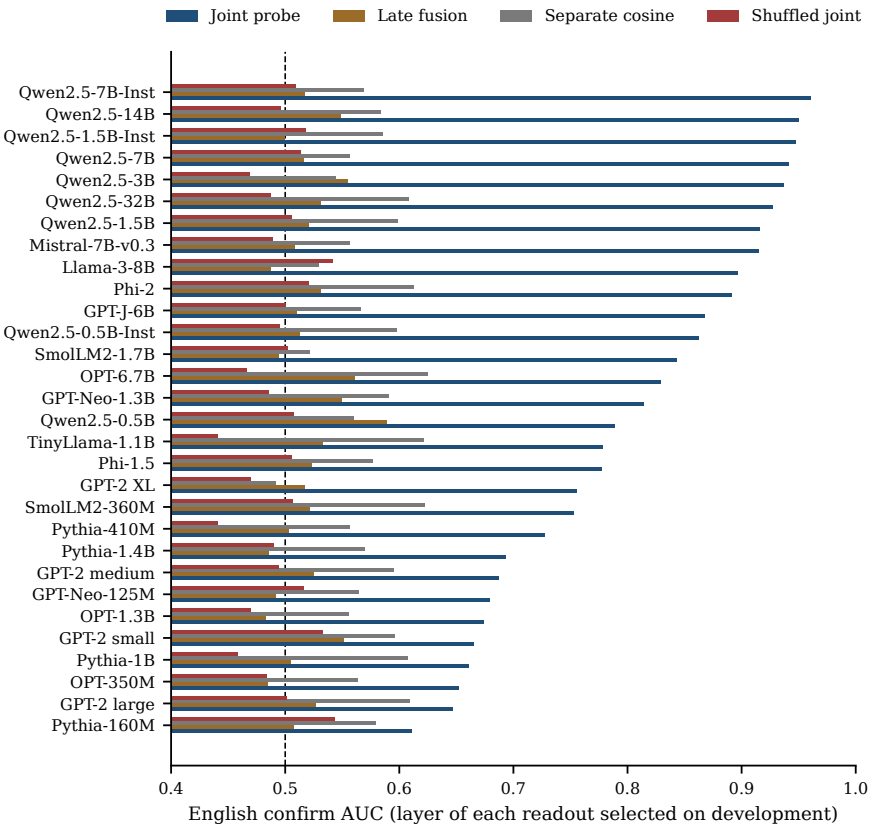


Figure 1: English confirm AUC on overlap-matched PAWS-X; each readout's layer is selected independently on development. Joint usually sits above late fusion, cosine, and shuffle. Rows combine Tables 1 and 4.

Contributions. We make the geometry–computation distinction experimentally separable within the same frozen backbone, using independently selected linear, nonlinear, token-level, shuffle, nonce, and adaptation controls. We establish its recurrence across causal, bidirectional, and encoder–decoder architectures, across datasets and languages, and under controlled compositional relations. We then show that the same route asymmetry governs learning from a common teacher: matched text, supervision, and updates produce a near-teacher joint student but a much weaker independent student. The key distinction from standard cross-encoder reranking is diagnostic: some joint rerankers enter the identity band while others remain with dense encoders, and changing only the route through the same frozen backbone creates a large supervision gap.

The implication is not that embeddings should be discarded. They remain effective for aboutness retrieval; the evidence instead warns against treating their shipped similarity as a sufficient certificate of propositional identity.

2 Related work

Sentence embeddings trained with siamese or contrastive objectives (Reimers & Gurevych, 2019; Gao et al., 2021; Wang et al., 2022; Xiao et al., 2024) dominate retrieval benchmarks (Thakur et al., 2021; Muennighoff et al., 2023). Recent work both questions and formalizes their geometry. Distances among alternative sentence representations need not predict their linguistic-task performance, because task-relevant information can reside in weighted combinations that cosine does not expose (Nastase & Merlo, 2025). Conversely, an information-theoretic account recovers interpretable feature directions and their composition in sentence embeddings (Tehenan, 2025). For CLIP-like unit-sphere embeddings, formal incompatibility results show that cosine geometry cannot simul-

taneously preserve content grouping, binding, relations, and negation; token-level similarity maps provide an alternative (Kang et al., 2025). Our claim is compatible with rich semantic geometry: it distinguishes geometry for sentence content and aboutness from cheap access to the binary relation that two independently represented propositions are identical.

Independent routes can be made substantially stronger. Specialized adaptation turns decoder models into text encoders (BehnamGhader et al., 2024); contrastive fine-tuning with hard negatives trains a multilingual bi-encoder directly on PAWS-X (Fedorova & Musatow, 2024); and lightweight supervised calibration of modern embeddings is competitive on broad paraphrase benchmarks (Rocci et al., 2026). Those studies ask how to build effective reusable embeddings. We separate the criterion present in shipped or frozen states from the criterion installed by task adaptation, and compare independent and joint students under common teacher-scored text and update counts. Closest to our system implication, Ralev et al. (2026) find that structure-targeted hard negatives improve local sensitivity but reduce out-of-domain dense retrieval, that MaxSim still misses structural near-misses, and that learned token-map verification is stronger. Kang et al. prove a geometric incompatibility for CLIP-like spaces; Ralev et al. document related retrieval consequences. Neither compares joint and independent routes through one frozen LM under matched supervision. Our frozen-backbone recurrence and matched-teacher students instead locate identity's initial accessibility and quantify route-dependent teachability under controlled adaptation. Cross-attentive teacher scores can be distilled into embedding students (Barkan et al., 2020); we hold teacher-scored text and update counts fixed while comparing joint and independent student routes.

The retrieval literature already knows that *cross*-encoders, which read the pair jointly, outperform bi-encoders on reranking (Nogueira & Cho, 2019; Khattab & Zaharia, 2020). That fact is usually treated as a compute/quality tradeoff. We show it is sharper: BGE rerankers reach the joint-probe band on overlap-matched PAWS, while MS-MARCO and Jina rerankers do not—joint scoring alone does not imply identity. We treat this as evidence about *where the relation is accessible*: at the joint probe's training size, tested readers over two frozen encodings remain much weaker, whereas a linear readout after pair computation is already strong. Late interaction (Khattab & Zaharia, 2020) still encodes independently, then scores a function of the two sequences; our late-fusion probe and MaxSim are its sentence- and token-level analogues.

Probing work asks what is linearly readable from LM states (Alain & Bengio, 2017; Tenney et al., 2019; Hewitt & Manning, 2019; Pimentel et al., 2020; Belinkov & Glass, 2019). The linear representation hypothesis concerns semantic directions *inside* a model (Park et al., 2024); sentence-level extensions map specified changes such as negation and conditionality to transferable rotations on embedding hyperspheres (Freenor & Alvarez, 2026). Such learned geometric operations can coexist with our result: organizing a specified transformation geometrically does not make generic similarity a certificate of identity for arbitrary sentence pairs. The Platonic Representation Hypothesis instead concerns convergence of representation geometry across models and modalities (Huh et al., 2024). Our cross-model analysis compares supervised relation scores on shared items, not coordinate alignment between hidden spaces. Recent probes on modern open models sharpen the same map: lexical identity is linearly strongest early and weakens with depth (Li & Subramani, 2026), while semantic and syntactic signals occupy partly separable mid-depth regimes (Acevedo et al., 2026), and lexical overlap continues to distort independent-embedding similarity under paraphrase stress tests (Rizwan et al., 2026). Our question is narrower: whether the *relation* "these two sentences express the same proposition" is cheaply readable from shipped independent points, or becomes accessible through computation over the pair.

PAWS and PAWS-X (Zhang et al., 2019; Yang et al., 2019) adversarialise bag-of-words overlap. Broader paraphrase evaluation has since moved beyond a single classification set (Michail et al., 2025); we stay with overlap-matched PAWS-style identity because that is exactly where shipped bi-encoder geometry is asked to certify sameness and fails. We go further and *match* overlap between paraphrase and non-paraphrase cells; overlap-only confirm AUC is near chance (0.45–0.55 across survey cells) and is reported alongside each main-table row.

3 PROTOCOL

Two complementary data protocols. The breadth survey samples PAWS-X English and Chinese pairs into discovery / development / confirm splits and matches lexical overlap of paraphrases and

non-paraphrases inside each cell. Confirm cells contain 150–500 pairs; Table 1 reports each reader at the layer selected by its *own* development AUC, with n shown per row. The matched learning curves instead use one common bank: 49,258 filtered English training pairs and the official 2,000-pair validation and 2,000-pair test splits, with nested subset seeds 17, 29, and 43. All readers at a given budget receive identical training indices and test pairs; external tests use 2,000 Chinese PAWS-X pairs, a fixed 2,000-pair QQP validation subset, and all 408 MRPC validation pairs (Wang et al., 2018; Dolan & Brockett, 2005). We report AUC; chance is 0.5 regardless of class balance.

Pair readouts. For each frozen causal LM we score identical pairs four linear ways. *Separate cosine*: last-token states from independent forwards, cosine. *Separate probe*: a logistic probe on $(\mathbf{h}_1 \odot \mathbf{h}_2, |\mathbf{h}_1 - \mathbf{h}_2|)$ after a PCA map fit on discovery. *Late fusion*: the same probe class on $[\mathbf{h}_1; \mathbf{h}_2]$—both vectors present, no mixing inside the LM. *Joint probe*: the two sentences are concatenated with a language-specific template that is *not* a yes/no question (`A: ... / B: ...`); we probe the last token of that joint pass. Probes are fit on discovery, and each readout's layer and regularisation are selected independently on development before confirm is examined. On the common bank we additionally fit a full-width independent MLP to $[\mathbf{h}_A; \mathbf{h}_B; \mathbf{h}_A \odot \mathbf{h}_B; |\mathbf{h}_A - \mathbf{h}_B|]$ with no PCA bottleneck, selecting depth, pooling, learning rate, and stopping point on validation.

Partner-shuffle and random-label controls. Sentence B is replaced by another sentence from the same language and split: format, language, and the marginal of B are preserved; pairing is destroyed. A probe that reads "two sentences appeared" rather than "these two belong together" will survive shuffle. Separately, on the common bank we permute the 1,200 *training labels* while retaining the original pairs and joint features; the three test AUCs are 0.505, 0.510, and 0.489. This checks readout learning under randomized supervision and is distinct from partner shuffle.

Depth. We record last-token states at a fixed grid of relative depths, plus the final-norm input (prenorm, captured by hook—HuggingFace `hidden_states[-1]` is already post-norm) and output (postnorm).

Generative channel, separately. A zero-shot yes/no verdict, length-normalised PMI, and conditional surprisal are scored on a smaller overlap-matched draw; unlike the joint probe, these ask the model to *speak*.

Encoders and rerankers. Nineteen public dense checkpoints (BGE, E5, GTE, MiniLM, mpnet, multilingual E5, E5-Mistral-7B) and six cross-encoders (three BGE rerankers, two MS-MARCO MiniLM rerankers, and Jina reranker-v2) are scored on a dedicated overlap-matched bank with n=900 confirm pairs per language, fixed sampling seeds, and English overlap matching (confirm overlap gap -0.002); bi-encoders and rerankers score the same items.

Readings fixed before confirm was examined. **(P1)** Computation: joint $\gg$ late fusion, and late fusion stays near the separate-probe band. **(P2)** Geometry: late fusion $\approx$ joint, both well above chance. **(P3)** Shuffle must fall to chance.

4 Shipped geometry does not make identity cheaply accessible

Dedicated encoders fail on overlap-matched confirm (Table 5, Appendix F). The strongest English cosine is 0.701 (BGE-large-zh); the strongest English-*trained* checkpoint is mpnet-base (0.653). Typical retrieval encoders sit at 0.55–0.65; scaling to E5-Mistral-7B does not break the band (0.613). A linear probe on frozen encoder vectors moves English BGE-small only from 0.588 to 0.607, and BGE-large-zh from 0.701 to 0.710: a linear reorientation does not close the gap.

Not every cross-encoder computes identity (Table 6, Appendix G). BGE-reranker-large (0.940), v2-m3 (0.902), and base (0.888) align with the joint-probe band of §5. Jina reranker-v2 (0.639) and MS-MARCO MiniLM rerankers (0.545–0.558) also read the pair jointly but stay in the dense band—they score passage relevance, not PAWS-style identity. The gap is therefore not simply "cross-encoder beats bi-encoder"; only some joint scorers expose a strong identity readout.

Independently encoded LM states tell the same story under fairer selection (Table 1; Figure 1). Each column now uses its own development-selected layer rather than the joint probe's layer. Across the

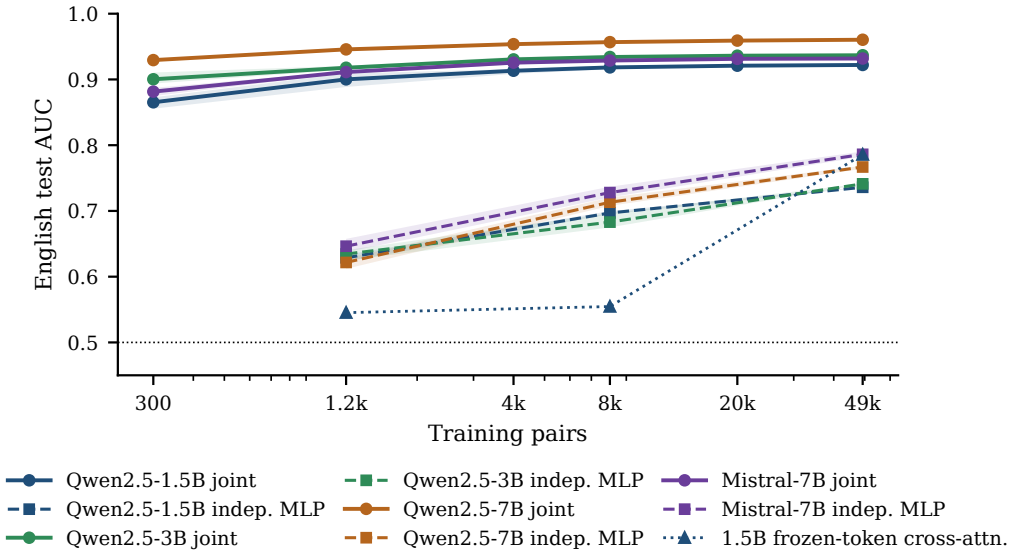


Figure 2: Same-test learning curves. Lines are three-seed means; bands are seed minima–maxima, not confidence intervals. Every method also uses 2,000 validation labels. The frozen-token reader is shown for Qwen2.5-1.5B. Lines connect measured budgets only.

30-checkpoint survey, separate cosine is 0.49–0.63, product/difference probing is 0.50–0.65, and late fusion is 0.48–0.59. Joint exceeds the product/difference probe in 29 of 30 rows; GPT-2-large is the small exception (0.647 versus 0.650). A fixed post-hoc replication using the same protocol and a pair-ID-disjoint bank (n=300 confirm) reverses the ordering: joint 0.626 ($[0.565, 0.689]$) versus separate 0.561 ($[0.494, 0.631]$), with shuffle 0.513. The intervals overlap, so we retain the original exception and interpret the two samples as evidence that GPT-2-large's gap is not stable. Late fusion is a useful *linear* control, not an upper bound on arbitrary functions of two vectors.

Identity might instead be recoverable from richer functions of the independent states. On the common bank, the full-width independent MLP reaches 0.622–0.646 at 1,200 labels, but remains 0.264–0.324 below the matched joint linear readout. At 49,258 labels it reaches 0.736–0.786 while joint reaches 0.922–0.960 (Figure 2); all 12 matched model-by-budget differences have paired test-sampling intervals above zero. ColBERT-style MaxSim of last-layer tokens reaches 0.62–0.71 on the separate overlap-matched bank. Scale does not close this gap: 14B MaxSim (0.628) is *lower* than 7B's, 32B MaxSim 0.694 is 0.23 below its development-selected joint 0.927, and Phi-2's MaxSim 0.701 remains below joint 0.891.

A stronger independent-encoding reader is a two-block bidirectional cross-attention module over *frozen* token sequences. On common Qwen2.5-1.5B features it reaches mean AUC 0.545, 0.555, and 0.786 at 1,200, 8,000, and 49,258 labels; the full-budget seeds span 0.748–0.860. At the same full budget, the 14B frozen-token reader reaches 0.868 (Table 11), the strongest independent frozen-backbone result we observe. This shows that useful material is retained, but the trained interaction module does not match the low-budget joint probe on average.

Fine-tuning the encoder itself is a different question. On the common bank, frozen BGE-large and E5-Mistral read with the same full-width head reach only 0.619–0.657 at 1,200 labels and 0.666–0.700 at 49,258. Updating the encoders closes much of the gap: full fine-tuning of BGE-large reaches 0.910 at 8,000 labels, and rank-16 q, v LoRA on E5-Mistral-7B reaches 0.897 at 1,200 and 0.946 at 8,000 (Table 2). These are not architecture rankings: pretraining, parameter count, and optimization differ. They instead locate the boundary of the claim—strong independent encoders can be *adapted* to the task, whereas the relation is already accessible to a small linear head after a frozen joint pass. MiniLM/BGE fine-tuning likewise fits PAWS (0.87–0.93) but reduces overlap-matched QQP performance and STS-B Spearman by about 0.25, illustrating that installing a task criterion into sentence geometry can tax general aboutness.

Table 2: Dedicated encoders on the common 2,000-pair English test; three-seed mean AUC. Trainable counts are encoder-side (all rows train a head). Joint rows freeze the backbone and train a $\leq$ 128-dimensional linear probe. Rows differ in pretraining and optimization; budgets are labels, not equal compute.

Encoder	Trainable	Adaptation	1,200	8,000
BGE-large (334M)	0	frozen	.619	.643
BGE-large (334M)	334M	full fine-tune	.661	.910
E5-Mistral (7.1B)	0	frozen	.657	.675
E5-Mistral (7.1B)	6.8M	LoRA (q, v)	.897	.946
Qwen2.5-1.5B (joint)	0	linear probe	.900	.918
Qwen2.5-7B (joint)	0	linear probe	.946	.957

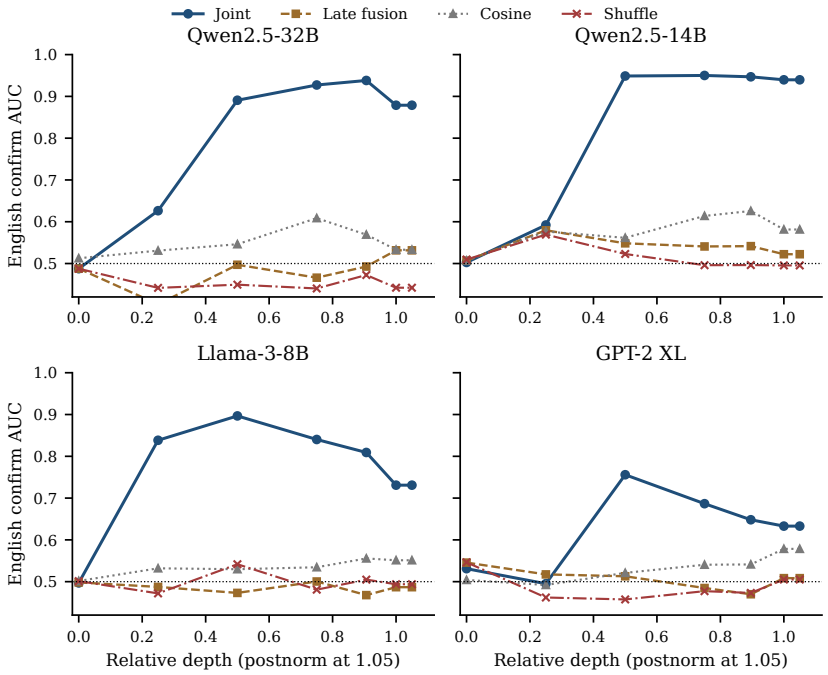


Figure 3: Depth of the four readouts. Joint rises at mid-depth; the other three stay near chance. Qwen14B keeps the signal into postnorm (0.950 $\rightarrow$ 0.940); Llama spends it (0.897 $\rightarrow$ 0.731).

5 Identity appears when the pair is computed

Table 1 (and Figure 1) reports English confirm AUC after selecting the layer of *each readout independently* on development; confirm is used only once for reporting.

Joint is not late fusion, and shuffle kills it. For Llama, Mistral, and Qwen $\geq$ 1.5B, joint exceeds late fusion by 0.37–0.46 AUC (Qwen2.5-14B: 0.950 versus 0.548); late fusion stays in the encoder band for these rows, supporting P1 rather than P2. At the same layer, shuffle AUC is 0.46–0.54: the probe reads the pairing, which shuffle destroys, so P3 holds.

Not a decoder-only family effect. Table 1 is dominated by Llama-style causal models. We therefore re-ran the same four readouts on the same overlap-matched English bank (n=200 confirm) on other causal stacks, bidirectional encoders, and encoder–decoders (Table 9, Appendix J). Every loaded model satisfies P1 and P3: joint exceeds late fusion, and partner shuffle returns near chance. Non-causal peaks sit in the main-table band—Flan-T5-base 0.953, nli-DeBERTa 0.940, DeBERTa-v3 0.921—with smaller T5/BART lower but same-signed (0.64–0.66). Nonce substitution of content words preserves the gap (Appendix K).

The wrapper is not the result. Four concatenation templates—`A:`/`B:`, "Sentence 1/2", a raw separator, a blank line—yield joint confirm AUC 0.919–0.950 on Qwen2.5-7B (shuffle 0.47–0.52), 0.933–0.961 on 14B, and 0.924–0.967 on 32B. Mistral 7B and Llama 3 8B at an off-peak layer on a smaller bank are more wrapper-sensitive (0.776–0.881 and 0.716–0.863), but every wrapper still beats late fusion and collapses under partner shuffle.

QQP is not a PAWS artefact. On overlap-matched QQP (English confirm), joint AUC is 0.88–0.91 for every $\geq$3B model we ran (Qwen2.5 3B–32B, Llama, Mistral, Phi-2; Qwen7B 0.910, 95% CI $[0.877, 0.943]$), while late fusion rises to 0.72–0.77: duplicate questions share *aboutness*, which PAWS's word-scramble matching forbids and separate vectors can see. To remove every single-sentence cue by construction we also built an *anchor-balanced* QQP bank: each anchor question appears exactly twice, once with a duplicate and once with an overlap-matched non-duplicate, so the first sentence carries no label information and lexical overlap alone scores 0.52. There, with the layer locked on development, joint AUC is 0.80–0.83 for all seven models (Qwen2.5 3B/7B/14B/32B, Llama 3 8B, Mistral 7B, Phi-2; 95% CIs $\approx \pm 0.05$), late fusion 0.55–0.60, separate cosine 0.58–0.62, and partner shuffle 0.42–0.46 (n=300 confirm; Table 8, Appendix I). Geometry can notice that two questions are about the same topic; certifying that they are the same question still takes the joint pass.

The computed direction transfers; the fitted geometry does not. Train the linear probe on PAWS joint states, lock layer and regulariser on PAWS dev, then apply the frozen probe zero-shot: on overlap-matched QQP confirm it scores 0.72–0.75 (Qwen7B / Llama / Mistral / Qwen14B); on *Chinese* PAWS confirm—different language, different wrapper—it scores 0.745–0.818, nearly the in-domain Chinese joint numbers (0.766–0.851); partner shuffle collapses it (0.48–0.53). The reverse direction is sharper: a probe trained only on QQP reads PAWS confirm at 0.834 / 0.848 / 0.724 / 0.696 (Qwen7B / Qwen14B / Llama / Mistral) without seeing a PAWS pair—the Qwen directions exceed the tested independent-encoding readouts on PAWS itself (best MaxSim 0.708), with Llama and Mistral at its level. One linear direction in mid-depth joint states carries the relation across datasets and across languages; the PAWS-tuned bi-encoder, by contrast, *lost* QQP accuracy relative to its own untuned checkpoint. Transfer is partial (0.72–0.75 vs 0.88–0.91 in-domain): part of any probe is dataset-specific.

Direct matched-supervision distillation. We give a joint and an independently encoding Qwen2.5-1.5B student the same 9,000 pairs, standardized consensus scores from Qwen-7B, Llama-3-8B, and Mistral-7B, the same 1,000 teacher-scored development pairs, and 1,689 optimizer updates per seed. Both update the same four intermediate blocks; the independent route uses mean-pooled sentence states and a full-width MLP. Checkpoint selection minimizes teacher-score development MSE; student updates and selection use no gold labels, although locating the teachers' coordinates uses labelled PAWS. On English PAWS, joint reaches 0.9420 versus 0.7599 independently, a paired difference of $+0.1821$ (95% CI $[0.1625, 0.2030]$); teacher-score Pearson correlation is 0.9285 versus 0.5677. Chinese PAWS gives $+0.1489$ $[0.1254, 0.1733]$. The smaller QQP and MRPC differences, $+0.0181$ and $+0.0431$, have intervals crossing zero. Thus the joint route learns the teacher relation more faithfully under this matched supervision and update budget; this comparison does not equate route FLOPs or prove that no independent architecture could do so.

The readout profile peaks at mid-depth. Peak joint is not the final residual (Figure 3). Llama and Mistral peak at layer 16 of 32; Qwen7B at layer 21 of 28; Qwen14B at layer 36 of 48; GPT-2 XL at layer 24 of 48. Postnorm joint is systematically *lower* than the mid-depth peak, but families differ in how much they overwrite: Llama 0.73 vs 0.90 (Qwen14B only 0.940 vs 0.950)—Qwen keeps what it computed, Llama spends it on next-token prediction. Either way, the final residual of a separately encoded string does not make the relation accessible to these readouts, and a final-norm edit that wrecks next-token prediction does not improve them: permuting the RMSNorm gain on Qwen2.5-7B yields KL 36 against the untouched head (top-1 agreement 0.4%) while separate cosine stays 0.558 versus 0.561; Llama, Mistral, 14B, and the smaller models remain in 0.55–0.62 under the same edits.

Chinese confirm (Appendix H) repeats the pattern at lower amplitude for multilingual models (joint 0.77–0.85 vs late fusion 0.54–0.60) and vanishes for GPT-2, which was not trained on Chinese.

Table 3: Controlled relation tasks. Mean candidate-ranking accuracy (%) across 18 panels and three reader seeds. Differences are percentage points with paired seed-and-fact bootstrap intervals.

Frozen backbone	Independent	Joint	Δ	95% CI
BGE-large	75.47	97.99	+22.52	[21.20, 23.93]
Qwen base	83.51	94.14	+10.63	[9.87, 11.37]
Consensus student	83.75	94.64	+10.89	[9.92, 11.87]
Gold student	84.20	93.89	+9.69	[8.85, 10.56]

6 Common supervision beyond natural paraphrase

Off-the-shelf heads need not implement the same objective. We therefore generate controlled candidate-ranking tasks that change agent/patient roles, the scope of negation across two events, or the assignment of quantities to two entities. Each fact contributes four pairs with labels $[1, 0, 0, 1]$; every anchor and candidate occurs with both labels, and correct and incorrect candidates have identical unigram counts. Crossing three relations, familiar versus nonce entities, and three expression styles yields 18 held-out panels over the same 96 fact indices. We freeze four backbones—BGE-reranker-large, base Qwen2.5-1.5B, and consensus- and gold-trained Qwen students—and fit joint and independent full-width MLPs with the same 1,024 labels.

All four intervals lie above zero. The largest gap occurs for BGE; its independent readers reach train AUC 0.71–0.85, so this row bounds accessibility under the tested optimization rather than an optimal independent ceiling. Base Qwen and the gold student fit both routes and retain 10.63- and 9.69-point advantages. When each relation class is excluded from training, development, early stopping, and layer selection, all 12 backbone-by-held-out-relation averages exceed 60% with lower intervals above chance, though only 60 of 72 individual vocabulary/style panels meet that criterion. The results support transferable structure in this restricted grammar, not universal expression-invariant correctness.

7 A recurring relation signal across independently trained models

Each of six models (Qwen2.5 1.5B/3B/7B/14B, Llama 3 8B, Mistral 7B) gets its joint direction located *once*, on PAWS English discovery, then frozen; all six score the same held-out items. Their item scores agree beyond the labels (mean pairwise Spearman 0.855; cross-family 0.834), their errors co-occur, and swapping the pair preserves the score (Spearman 0.89–0.93). Applied zero-shot to SNLI, every direction orders paraphrase $>$ entailment $>$ neutral $>$ contradiction $>$ non-paraphrase in the mean. Transfer to QQP/MRPC is partial, and consensus does not uniformly beat the best teacher; Appendix B reports the full agreement, error, symmetry, label-audit, and teacher-sensitivity analyses.

8 Conclusion

Across causal, bidirectional, and encoder–decoder families, high-overlap meaning identity becomes cheaply readable when a pretrained model computes over the pair. The result survives nonlinear and token-level readers of frozen independent states, partner shuffle, nonce substitution, controlled relations, and matched-supervision distillation; independently trained families expose a strongly concordant, symmetric, and graded signal. Independent routes can learn the criterion with enough supervision and encoder adaptation, but that does not make it a property of the representations those systems ship. The claim concerns accessibility under tested readers and budgets, not information-theoretic absence; Appendix A gives the complete scope and limitations. Embeddings organize useful neighbourhoods of wording and aboutness; joint computation exposes propositional identity. Conflating those two resources mistakes computation for geometry.

AI USE STATEMENT

Under author direction, generative AI coding assistants assisted with protocol critique, implementation, rule-based synthetic-data generation, data cleaning, and experiment monitoring. They also assisted with literature search, figure and table generation, manuscript drafting and editing, and consistency checks. The reported controlled-task labels are determined by explicit construction rules and checked by a parser; natural-data evaluations use public dataset labels, not free-form AI judgments. Checks on AI-assisted artifacts included unit tests, prediction-based metric recomputation, and cross-checks between manuscript values and archived records. The authors take responsibility for the final text, claims, code, and data artifacts, including errors introduced through AI assistance.

ETHICS STATEMENT

The study uses public model weights and research datasets PAWS-X, QQP, MRPC, SNLI, and STS-B. No human subjects, private data, or deployed user logs are involved. The supplement attributes these datasets and does not relicense benchmark text; downstream use remains subject to each source's terms. The practical claim—that bi-encoders should not certify paraphrase-level identity—is about system design, not about ranking people.

REPRODUCIBILITY STATEMENT

Data construction (overlap matching, splits), the four readouts, the shuffle control, and the prenorm hook are specified in §3 and the supplementary code. Confirm scores are not used for fitting, hyperparameter choice, or layer selection. Model identifiers are the public HuggingFace names in Table 1; the cross-architecture record is stored under the `paws_cross_arch.json` key in `paper/data/survey.json`. Figures 1 and 3 are generated from the same JSON dumps as the tables.

REFERENCES

Santiago Acevedo, Alessandro Laio, and Marco Baroni. Differential syntactic and semantic encoding in LLMs. In *ICML*, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *ICLR Workshop*, 2017.

Oren Barkan, Noam Razin, Itzik Malkiel, Ori Katz, Avi Caciularu, and Noam Koenigstein. Scalable attentive sentence pair modeling via distilled sentence embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020. doi: 10.1609/aaai.v34i04.5722.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *COLM*, 2024. URL `https://arxiv.org/abs/2404.05961`.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *TACL*, 7:49–72, 2019.

Stella Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. *ICML*, 2023.

Sid Black et al. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. 2021.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval*, 2017.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *International Workshop on Paraphrasing*, 2005.

Abhimanyu Dubey et al. The Llama 3 herd of models. *arXiv:2407.21783*, 2024.

Inessa Fedorova and Aleksei Musatow. Cross-lingual paraphrase identification. *arXiv preprint arXiv:2406.15066*, 2024. doi: 10.48550/arXiv.2406.15066.

Michael Freenor and Lauren Alvarez. Mapping semantic & syntactic relationships with geometric rotation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qEmKKvYr07.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. *NAACL*, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024. URL https://arxiv.org/abs/2405.07987.

Albert Q. Jiang et al. Mistral 7B. *arXiv:2310.06825*, 2023.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is CLIP ideal? no. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22436–22446, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kang_Is_CLIP_ideal_No._Can_we_fix_it_Yes_ICCV_2025_paper.html.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *SIGIR*, 2020.

Michael Li and Nishant Subramani. Model internal sleuthing: Finding lexical identity and inflectional features in modern language models. In *ACL*, 2026.

Andrianos Michail, Simon Clematide, and Juri Opitz. PARAPHRASUS: A comprehensive benchmark for evaluating paraphrase detection models. In *COLING*, pp. 8749–8762, 2025.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *EACL*, 2023.

Vivi Nastase and Paola Merlo. Testing the assumptions about the geometry of sentence embedding spaces: The cosine measure need not apply. *arXiv preprint arXiv:2509.01606*, 2025. doi: 10.48550/arXiv.2509.01606.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. In *arXiv:1901.04085*, 2019.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *ICML*, 2024.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *ACL*, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *OpenAI*, 2019.

Radoslav Ralev, Aditeya Baral, Iliya Sotirov Zhechev, Jen Agarwal, and Srijith Rajamohan. Training for compositional sensitivity reduces dense retrieval generalization. In *ICLR Workshop on Scientific Methods for Understanding Deep Learning*, 2026. URL https://openreview.net/forum?id=F5ueiEGP4l.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP*, 2019.

Hammad Rizwan, Muhammad Umair Haider, Nishant Subramani, Mona T. Diab, A. B. Siddique, and Hassan Sajjad. On the persistent effects of lexicality in large language models. *arXiv preprint arXiv:2606.02750*, 2026.

Giovanni Rocci, Andrianos Michail, Andreas Loizidis, Simon Clematide, and Juri Opitz. Efficient paraphrase detection with embeddings. In *Proceedings of KONVENS 2026*, 2026. URL `https://www.korpuslab.uni-hamburg.de/en/konvens2026/59.pdf`.

Matthieu Tehenan. Semantic geometry of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11993–12004, 2025. doi: 10.18653/v1/2025.findings-emnlp.641.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *ACL*, 2019.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP Workshop BlackboxNLP*, 2018.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv:2212.03533*, 2022.

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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *EMNLP*, 2019.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *NAACL*, 2019.

A Scope and limitations

The primary construct is high-overlap PAWS-style paraphrase, not every notion of meaning identity. The common-bank curves use full official test splits but do not isolate lexical overlap; the breadth survey does, but its model-specific cells are smaller and support within-row recurrence rather than a matched-data cross-model leaderboard. Reader comparisons establish accessibility under specified architectures, searches, and supervision budgets, not information-theoretic absence or an optimal independent ceiling. The full-width MLP and frozen-token reader learn substantially with more data, and adapted E5/BGE encoders can approach the joint readout. Conversely, those adaptation rows differ in pretraining, parameter count, trainable parameters, and compute; equal label counts are not equal resource budgets. All learning-curve configurations also consume 2,000 validation labels, so the results do not establish a fixed equal-performance sample-complexity multiplier.

Controlled tasks come from a restricted grammar; matching unigram counts removes one shortcut, not every surface regularity, and some independent readers fail the training-fit diagnostic. The direct distillation comparison matches teacher-scored texts and updates but not route FLOPs; teacher localization still uses gold labels. Transfer to QQP/MRPC is partial, and the consensus contrast depends on its pool and checkpoint rule. Finally, the experiments identify a robust readout pattern, not a unique attention-level causal mechanism.

B CROSS-MODEL RELATION DETAIL

If the readout reflects a recurring relation rather than only a dataset-specific criterion, models trained by different organisations on different corpora should produce concordant scores. On the same 250 PAWS confirm pairs, pairwise Spearman between the six models' item scores averages 0.855; restricted to cross-family pairs (Qwen vs Llama vs Mistral) it is 0.834. Agreement is not explained only by the binary label: within the negative and positive classes, mean rank correlation is 0.655 and 0.663, respectively.

Errors co-occur; consensus is not uniformly better. Where one model errs, the others err: $P(\text{B wrong} \mid \text{A wrong})$ exceeds the base rate by 2.9–$5.7\times$ on PAWS, and the seven items that all six models get wrong are three orders of magnitude more frequent than independent errors would allow. Averaging scores does not automatically beat the best single model: the models share residuals as well as signal.

The relation is symmetric and graded. Swapping the sentences barely moves the score ($\text{score}(A, B)$ vs $\text{score}(B, A)$ Spearman 0.89–0.93 on PAWS), and the forward–swap discrepancy carries no label information (AUC 0.50–0.61). Applied zero-shot to SNLI, every one of the six directions orders PAWS paraphrase $>$ entailment $>$ neutral $>$ contradiction $>$ PAWS non-paraphrase, monotonically in the mean, with entailment-vs-contradiction AUC 0.79–0.98.

Dataset transfer and label disagreement. On overlap-matched QQP confirm, all six models contradict the published label on 18 items ($57\times$ the independence rate). Six of them (2.4% of the bank) are confident ($|z| > 1$); the pairs, with published labels and consensus z, are in Appendix O. Consensus AUC against those published labels is 0.773. Transfer of the frozen PAWS coordinate is partial on overlap-matched MRPC (0.48–0.61), and where it transfers less the models also agree less (frozen-coordinate QQP: pairwise Spearman 0.34–0.84, error lift 1.3–$2.3\times$).

Consensus versus one teacher. In a fixed comparison, consensus- and single-Qwen-teacher joint students receive the same pool, loss, updates, and teacher-MSE checkpoint rule. Consensus is slightly lower on English PAWS (0.9427 vs 0.9480) but higher on QQP (0.7856 vs 0.7489); at their common third-epoch endpoint the PAWS interval includes zero while the QQP gain remains. The robust result is the direct joint-versus-independent transfer, not an unconditional advantage for multiple teachers.

C SCALE AND FAMILY

Two scale stories are easy to confuse: computation as a Qwen idiosyncrasy, versus computation as the default organisation of modern LMs, which older English LMs grow more slowly; Llama 3 and Mistral—not Qwen, not instruction-tuned—close the first. Instruction tuning is not the cause either: Qwen2.5-1.5B base already reaches 0.916 (instruct 0.947; 7B-Instruct 0.961 with development-selected late fusion 0.517).

Inside Qwen2.5 the development-selected joint readout shows a rising trend followed by an early plateau: 0.789 (0.5B) $\to$ 0.916 (1.5B) $\to$ 0.937 (3B) $\to$ 0.941 (7B) $\to$ 0.950 (14B) $\to$ 0.927 (32B, 95% CI $[0.884, 0.964]$), with development-selected late fusion at 32B 0.532. Beyond 1.5B, the joint readout remains high while the tested independent linear readouts do not show a corresponding trend.

The older families are weaker and less monotone (Table 1, lower block): GPT-2 scores $0.665/0.688/0.647/0.756$ (small/medium/large/XL), including the large-model exception noted above. Several other lineages show a higher joint than late-fusion readout at their largest tested size: TinyLlama-1.1B and Phi-1.5 split joint 0.778 vs late ≈ 0.54, Phi-2 reaches 0.891; SmolLM2 goes $0.753 \to 0.843$; GPT-Neo $0.680 \to 0.815$ and GPT-J-6B 0.868; OPT reaches 0.829 at 6.7B; Pythia remains 0.611–0.727.

Table 4: English confirm AUC, older/smaller families; same protocol and layer selection as Table 1. GPT-2 large is the one survey row whose joint probe does not exceed its separate probe.

Model	n	joint layer	Ov.	Cos	Sep	Late	**Joint**	Shuf
Pythia-410M	400	L12	.488	.556	.569	.503	**.727**	.441
Pythia-1.4B	300	L12	.489	.570	.572	.485	**.693**	.490
GPT-2 medium	400	L12	.486	.595	.613	.525	**.688**	.494
GPT-Neo-125M	400	L9	.498	.564	.600	.492	**.680**	.517
OPT-1.3B	300	L18	.509	.556	.579	.483	**.674**	.470
GPT-2 small	500	L9	.527	.596	.652	.552	**.665**	.533
Pythia-1B	300	L8	.483	.607	.611	.504	**.661**	.458
OPT-350M	400	L18	.452	.563	.560	.485	**.653**	.483
GPT-2 large	300	L27	.531	.609	.650	.527	**.647**	.501
Pythia-160M	400	L6	.488	.579	.590	.508	**.611**	.544

D THE MOUTH IS OPTIONAL

A joint probe reads a state; a chatbot user reads tokens. On a smaller overlap-matched draw (n=150–300 confirm), the channels split by family. Qwen base models can *say* the answer (zero-shot verdict AUC 0.915–0.987 from 3B to 32B; PMI stays 0.47–0.61). Llama 3 8B and Mistral 7B expose strong joint readouts (0.897 / 0.915) while their verdicts remain 0.614 / 0.606; the tested GPT-2 and Phi models show the same channel gap (GPT-2 verdicts 0.47–0.51 against joint up to 0.756; Phi 0.56–0.61 against 0.778–0.891). Prompting is not a substitute for the joint state: PMI does not recover it, and conditional surprisal sits at 0.55–0.76 across models, unrelated to the joint probe.

E OLDER AND SMALLER FAMILIES

F DEDICATED BI-ENCODERS ON OVERLAP-MATCHED CONFIRM

Table 5: Bi-encoder cosine AUC on overlap-matched PAWS-X confirm (n=900, fixed seeds, English overlap gap -0.002). Chance $= 0.5$. STS-B Spearman on the validation split. Full list sorted by English AUC.

Encoder	English	Chinese	STS-B
BGE-large-zh	0.701	0.651	0.784
mpnet-base	0.653	0.481	0.881
BGE-large-en	0.642	0.481	0.897
BGE-base-zh	0.640	0.549	0.773
gte-large	0.640	0.480	0.899
multilingual-E5-large	0.640	0.509	0.888
E5-large-v2	0.638	0.470	0.896
gte-base	0.624	0.478	0.896
E5-Mistral-7B-Instruct	0.613	0.542	0.762
MiniLM-L6-v2	0.613	0.487	0.867
MiniLM-L12-v2	0.612	0.487	0.875
BGE-base-en	0.609	0.479	0.895
E5-base-v2	0.594	0.471	0.886
BGE-small-en	0.588	0.480	0.889
paraphrase-MiniLM-L12	0.588	0.485	0.875
multilingual-E5-base	0.583	0.498	0.879
E5-small-v2	0.574	0.475	0.878
multilingual-E5-small	0.557	0.494	0.875
BGE-small-zh	0.555	0.484	0.771

Linear probes on frozen encoder vectors (discovery train, confirm report) add at most 0.01–0.03 AUC (BGE-large-zh probe 0.710).

Table 6: Cross-encoder AUC on the same overlap-matched confirm bank (n=900). Chance = 0.5. Joint input, but not all models compute identity.

Reranker	English	Chinese
BGE-reranker-large	0.940	0.883
BGE-reranker-v2-m3	0.902	0.845
BGE-reranker-base	0.888	0.857
Jina reranker-v2-multilingual	0.639	0.604
MS-MARCO MiniLM-L12	0.558	0.507
MS-MARCO MiniLM-L6	0.545	0.506

G CROSS-ENCODERS ON OVERLAP-MATCHED CONFIRM

H CHINESE CONFIRM

At the English peak layer unless noted, Chinese joint AUC is 0.868 (Qwen2.5-32B, peak at L48), 0.851 (Qwen2.5-14B, peak at L24), 0.806 (Qwen2.5-7B), 0.787 (Qwen2.5-3B), 0.795 (Llama-3-8B), 0.766 (Mistral-7B), with late fusion 0.54–0.60 and shuffle $\approx$ 0.45–0.55. GPT-2 XL Chinese joint is 0.57. The pattern transfers to Chinese in the multilingual models tested but not in the English-trained GPT-2 checkpoints.

Table 7: Chinese confirm AUC at the English peak layer (14B uses its own Chinese peak, L24; 32B uses L48). Chance = 0.5.

Model	Cos	Late	Joint	Shuf
Qwen2.5-32B	.541	.531	.868	.510
Qwen2.5-14B	.529	.598	.851	.550
Qwen2.5-7B	.504	.541	.806	.453
Qwen2.5-3B	.491	.544	.787	.528
Llama-3-8B	.533	.577	.795	.533
Mistral-7B	.545	.552	.766	.487
Qwen2.5-1.5B-Inst	.491	.548	.786	.515
GPT-2 XL	.488	.510	.568	.482

I ANCHOR-BALANCED QQP BANK

Each anchor question q contributes exactly two pairs, $(q,$ duplicate, $1)$ and $(q,$ non-duplicate, $0)$, with q as the first sentence in both; the non-duplicate partner is chosen among q's QQP non-duplicates to match the lexical overlap of the duplicate (gap < 0.10), and splits are anchor- and text-disjoint (1,200 discovery / 300 development / 300 confirm pairs). Overlap-only AUC on confirm is 0.519. Layer and regulariser are locked on development; all columns are confirm AUC at that layer.

Table 8: Anchor-balanced QQP confirm (n=300). Joint 95% bootstrap intervals in brackets.

Model	Layer	Joint	Late	Sep	Cos	Shuf
Qwen2.5-3B	L27	.813 [.766, .857]	.576	.581	.580	.432
Qwen2.5-7B	L21	.813 [.765, .860]	.546	.629	.597	.445
Qwen2.5-14B	L36	.807 [.762, .855]	.581	.609	.605	.428
Qwen2.5-32B	L48	.815 [.769, .861]	.603	.625	.583	.426
Llama-3-8B	L16	.805 [.753, .854]	.569	.633	.585	.463
Mistral-7B	L16	.831 [.781, .876]	.585	.654	.616	.456
Phi-2	L16	.798 [.745, .846]	.588	.626	.614	.422

J CROSS-ARCHITECTURE REPLICATION

Table 1 is dominated by Llama-style causal models. To test whether P1–P3 are architecture-specific we re-ran the same four readouts on the same overlap-matched English bank (n=200 confirm, 800 discovery; overlap-only confirm AUC ≈ 0.49) on three families that are *not* Qwen/Llama/Mistral: other causal stacks; bidirectional encoders (mean pool over tokens); and encoder–decoders (encoder stack only, mean pool). Causal models use last-token states as in Table 1; encoder and encoder–decoder rows use mean pooling over the encoder hidden states. Every successful load satisfies joint $\gg$ late fusion and shuffle near chance.

Table 9: Cross-architecture English confirm AUC (n=200) at the layer of peak joint probe. Chance $= 0.5$. Joint 95% bootstrap intervals in brackets for the strongest non-causal rows.

Family	Model	Cos/Sep	Late	Joint	Shuf
causal	Phi-2	.528	.576	.871 [.823, .914]	.445
causal	GPT-J-6B	.549	.581	.899 [.855, .937]	.463
causal	OPT-6.7B	.539	.570	.833 [.778, .884]	.424
causal	GPT-Neo-1.3B	.486	.551	.795	.482
causal	Pythia-1.4B	.455	.574	.750	.476
causal	SmolLM2-1.7B	.563	.548	.818	.472
encoder	nli-DeBERTa-v3	.677	.567	.940 [.909, .966]	.550
encoder	DeBERTa-v3-base	.607	.558	.921 [.881, .956]	.514
encoder	RoBERTa-base	.574	.588	.691	.520
enc–dec	Flan-T5-base	.543	.539	.953 [.927, .977]	.505
enc–dec	T5-small	.589	.553	.656	.522
enc–dec	BART-base	.576	.570	.640	.494

K NONCE SUBSTITUTION CONTROL

Published PAWS strings may appear in pretraining. As a decontamination control we rebuild the overlap-matched English bank by replacing content words *consistently inside each pair* with invented tokens (labels unchanged; relative overlap preserved). A probe that only recalled published wording should collapse; a probe that computes the pairing need not. Table 10 reports the same four readouts on that nonce bank (n=200 confirm). Joint remains above late fusion on every model; shuffle stays near chance.

Table 10: Nonce-substituted English confirm AUC (n=200). Chance $= 0.5$.

Model	Late	Sep	Joint	Shuf
Qwen2.5-0.5B	.490	.558	.676	.485
Qwen2.5-1.5B	.541	.539	.796	.515
Qwen2.5-3B	.510	.528	.841	.530
Phi-2	.560	.527	.763	.488
GPT-Neo-1.3B	.503	.540	.775	.478
Pythia-1.4B	.525	.564	.722	.492
OPT-6.7B	.506	.570	.691	.494
GPT-J-6B	.544	.551	.836	.465

L READERS TRAINED ON THE FULL PAWS-X TRAIN SPLIT

Frozen token states are cached once per model at two depths (0.75 of depth, "mid", and the final layer); readers are scored on overlap-matched confirm cells built with the same fixed seeds as Table 1 (default n=300 here; cross-file comparisons must match `bank_meta`). The cross-attention reader

is the two-block module of §4 trained on the leak-filtered PAWS-X English train split (49,375 of 49,401 pairs kept) at increasing sizes; learning rate and epoch are selected on development.

Table 11: Confirm AUC of the frozen-sequence cross-attention reader as a function of training pairs. Depth selected on development. On the same 49k pairs, logistic late fusion of the two sentence vectors stays 0.52–0.58; an MLP on the pooled pair reaches 0.72–0.75.

Model	1.2k	4k	8k	20k	49k
Llama-3-8B	.550	.499	.557	.515	.695
Mistral-7B	.512	.564	.584	.530	.697
Qwen2.5-3B	.504	.529	.551	.513	.700
Qwen2.5-7B	.470	.516	.461	.523	.680
Qwen2.5-14B	.486	.547	.511	.561	.868

M WHY LAST-TOKEN OF A CONCATENATION IS NOT A PROMPT

The joint template never contains the words "paraphrase", "same", or "yes". We do not score an answer token. The generative verdict in Appendix D is a different experiment on purpose.

N FINAL-NORM FOOTNOTE

HuggingFace `hidden_states[-1]` is post-final-norm. A hook on the norm input is required to talk about prenorm at all. Double-normalising that tensor is a silent invalidation of any cosine-vs-readout comparison.

O UNANIMOUS CONFIDENT CONTRADICTIONS OF QQP GOLD

Six items of the overlap-matched QQP confirm bank (2.4%) are contradicted by all six models with high consensus confidence ($|z| > 1$). Published labels and consensus z are below.

- gold = not duplicate, $z = +1.61$: "Can you help me solve the problem below?" / "Can you help me solve this problem?"
- gold = not duplicate, $z = +1.61$: "What are the things I can do to become more responsible?" / "What are some ways to become more responsible?"
- gold = not duplicate, $z = +1.14$: "What is actually mep (mechanical electrical plumbing)?" / "What is MEP (mechanical electrical plumbing)?"
- gold = not duplicate, $z = +1.30$: "If Donald Trump becomes president, how will he make America great again?" / "Will Donald Trump 'make America great again'?"
- gold = not duplicate, $z = +1.11$: "How does nail polish remove nail polish?" / "How do you remove nail polish?"
- gold = duplicate, $z = -1.31$: "Is it wise to pursue an MS in the USA after Trump's win?" / "Can you share merits and demerits of MS international students in USA after election result, as Hillary Clinton's Green card or Trump's job offers?"

Consensus AUC against the published labels is 0.773 (best single model 0.778). The six models agree with one another on these items.

P COMMON-BANK DIFFERENCES AND STRONG ENCODERS

For the common learning curves, joint-minus-independent-MLP paired 95% intervals are $[.252, .293]$, $[.199, .244]$, and $[.164, .209]$ for Qwen2.5-1.5B at 1,200/8,000/49,258 labels; $[.262, .306]$, $[.230, .272]$, and $[.173, .218]$ for 3B; $[.303, .346]$, $[.224, .264]$, and $[.173, .216]$ for 7B; and $[.242, .288]$, $[.179, .223]$, and $[.124, .168]$ for Mistral-7B. Intervals resample shared-sentence connected components and condition on the three fitted seeds; they do not resample training runs.

Table 12: Dedicated-encoder arms on all four common-bank tests; three-seed mean AUC. PAWS English is primary.

Encoder	Adaptation	Pairs	PAWS-en	PAWS-zh	QQP	MRPC
BGE-large	full FT	1,200	.661	.555	.803	.674
BGE-large	full FT	8,000	.910	.626	.832	.690
BGE-large	frozen	49,258	.666	.549	.606	.694
E5-Mistral	LoRA	1,200	.897	.786	.867	.761
E5-Mistral	LoRA	8,000	.946	.809	.842	.793
E5-Mistral	frozen	49,258	.700	.614	.812	.727

Q DIRECT DISTILLATION AND CONTROLLED TRANSFER

Table 13: Direct distillation under common teacher supervision. Three-seed mean AUC; intervals are for joint minus independent.

Test	n	Joint	Independent	Δ	95% CI
PAWS English	2000	.9420	.7599	+.1821	[.1625,.2030]
PAWS Chinese	2000	.8245	.6756	+.1489	[.1254,.1733]
QQP	2000	.7728	.7547	+.0181	[-.0022,.0381]
MRPC	408	.7204	.6773	+.0431	[-.0242,.1051]

For leave-one-relation-out controlled transfer, the 12 backbone-by-relation mean accuracies are: BGE, role/negation/quantity $97.54/79.40/99.65$; base Qwen $60.36/76.33/98.26$; consensus student $60.62/72.34/99.83$; and gold student $68.37/66.87/97.77$. Every paired lower interval exceeds 50%, but panel-level performance is heterogeneous, especially for nonce-role expressions.