# LLM-SEMANTICLAN: ANCHORED PUBLIC LATENT COMMUNICATION VIA KV CACHES ACROSS HETEROGENEOUS LARGE LANGUAGE MODELS

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

## ABSTRACT

Latent communication is the key technique to coordinate heterogeneous large language models for composite tasks, which directly exchanges information between internal states such as KV caches instead of texts to avoid slow sequential decoding and redundant prefilling. However, existing latent communication either maps caches between one pair of models at a time or relies on a shared space defined only by adapters trained together, leaving no fixed coordinates that a new model of another scale, family, architecture, or modality could align to. To address these issues, we formalize Public Latent Communication (PLC), which links every model via a shared latent space like a local area network (LAN). We realize PLC in LLM-SemanticLAN, a lightweight and high-efficiency framework that anchors the coordinates of this space in closed form by ridge regression before training, using targets from a frozen reference model. LLM-SemanticLAN employs model-specific codecs to map native KV caches of LLMs into the compact shared representations of this space for transfer and reconstruction, with codecs of only 1.0%–3.6% of the base model parameters and latent representations 64× smaller on average than the native KV caches at the same BF16 precision. Extensive evaluations across scales, attention architectures, families, and modalities verify its strong generalization, showing that aligning a new model with only a single existing member in the shared latent space, supervised only by the receiver's behaviour, enables zero-shot communication across all unpaired members. Without benchmark-specific training, LLM-SemanticLAN achieves near-lossless parity with text-to-text communication and superior vision-to-text performance, while delivering 12–35× speedups in end-to-end latency and 2.2–4.3× FLOPs reductions.

## 1 INTRODUCTION

Multi-model systems combine models of different architectures, capabilities, and modalities, as when a specialist hands its solution to a generalist or a vision model tells a text model what it saw. Natural language is a universal means of communication, yet it incurs autoregressive generation at the sender and full re-encoding at the receiver. Latent communication instead passes internal states such as KV caches, but these states are written in coordinates private to each model.

Passing a cache from one model to another would let the receiver use what the sender read without reading it again, and within one model such reuse already works (Gim et al., 2024; Yao et al., 2025; Hu et al., 2025). Across models, however, the cache of one model cannot be read by another, even between models of identical cache shape (§3.2). Existing latent communication methods learn a separate cache translator for each sender-receiver pair (Liu et al., 2026e; Heo et al., 2026; Qu et al., 2026; Fu et al., 2026), so $N$ models need $N(N-1)$ translators and every payload remains a full cache (Figure 1a). A shared latent space with per-model adapters removes this quadratic cost, and a new model joins by training its adapters from random weights against several frozen existing models (Dery et al., 2026) (Figure 1b). However, such a space is defined only by the adapters trained together, so nothing fixes its coordinates in advance. A new model has no starting point it can compute from its own cache, and a model with another tokenizer, attention type, or modality has no guarantee of mapping into the space at all.



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.

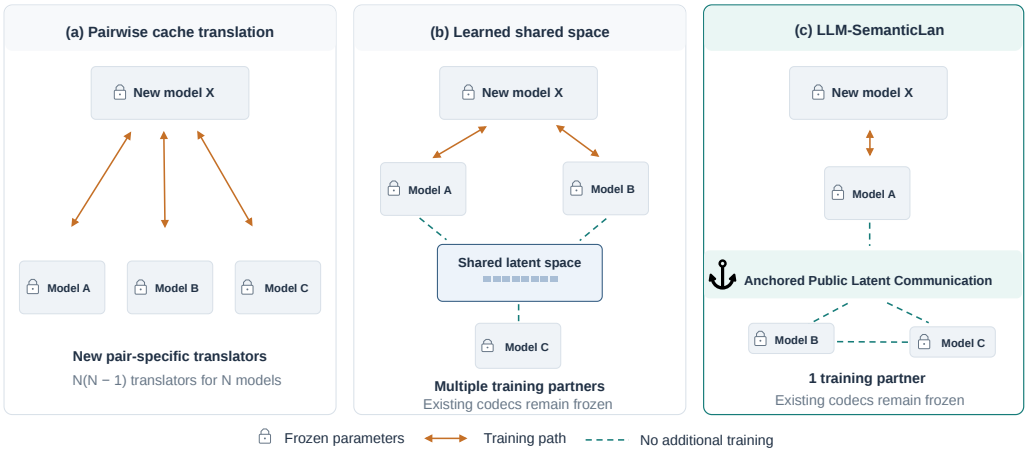



Figure 1: Connecting a model by pairwise maps, a shared space, or LLM-SemanticLAN.

We formalize *public latent communication* (PLC), which links every model via a shared latent space like a local area network (LAN), and realize it in **LLM-SemanticLAN**, which anchors the coordinates of this space before any training (Figure 1c). Each model contributes one lightweight codec, an encoder that maps its native KV cache into an ordered sequence of compact latent vectors and a decoder that reconstructs a prefix cache in its own format. LLM-SemanticLAN builds its codecs in two stages. First, the initial models build the space together and solve their codecs in closed form by ridge regression. The targets come from a frozen reference model that takes no part in communication. The coordinates therefore exist before any training and favour no model. Second, the reference is dropped, and the codecs are trained only to match the receiver's behaviour on the passage itself, with no cache loss. A new model solves its codec against the latent vectors of one existing model, which already lie in the anchored coordinates, while every other model and codec stays frozen.

We evaluate frozen models spanning model scale, domain specialization, families, attention architectures, and input modalities. Averaged over two held-out sets built for this work and three public QA benchmarks excluded from codec training (§4.1), sender-receiver pairs never trained retain 94.3 to 100.3 percent of the accuracy of text-to-text communication, and pairs of models added independently, which were never trained together, reach 100.3 percent (Figure 3, Table 1). On held-out questions built from COCO images (Lin et al., 2014), the representations of a visual encoder trained with one text receiver outperform written descriptions, and they also work on two public image benchmarks and for text receivers it was never trained with (§4.3). Text and images therefore travel through the same semantic space. Each transmitted latent representation is 64× smaller than a native cache on average (Appendix D), and LLM-SemanticLAN answers 12–35× faster with 2.2–4.3× fewer FLOPs than text-to-text communication. These gains rely directly on our initialization and supervision designs. Under the same training budget, codecs starting from random weights answer 0.8 percent of questions compared to 85.3 percent from our closed-form initialization, while replacing behaviour matching with continuation alone drops accuracy by 10.8 points.

Our main contributions are:

- We formalize PLC and realize it in LLM-SemanticLAN, which anchors the latent space in closed form by ridge regression before training and trains the codecs to match the receiver's behaviour.
- We show that a new model can communicate with all models in the latent space by training only its own codec against a single existing model. This generalizes across scale, domain specialization, family, attention architecture, and modality, where pairs never trained together retain 94.3 to 100.3 percent of the accuracy of text-to-text communication.
- We demonstrate that LLM-SemanticLAN is efficient and lightweight. It answers 12–35× faster with 2.2–4.3× fewer FLOPs than text-to-text communication, using codecs that hold 1.0 to 3.6 percent of model parameters and latent representations 64× smaller than native caches on average.

## 2 Related Work

**KV-cache reuse and transfer.** Reusing attention states reduces repeated context processing, but differences between models complicate direct reuse. Prompt Cache, CacheBlend, and EPIC reuse context within one model (Gim et al., 2024; Yao et al., 2025; Hu et al., 2025), while DroidSpeak extends reuse across fine-tuned variants (Liu et al., 2026e). For distinct models, closed-form methods fit cache mappings (Heo et al., 2026; Qu et al., 2026), whereas C2C and XKV learn cache fusion (Fu et al., 2026; Liu et al., 2026a). Beyond cache compatibility, LCF-X studies agents with different contexts (Rossi et al., 2026), and Chen et al. (2026) examine transfer without receiver access to the input. In our setting, the receiver sees the question but not the original passage, so the transferred representation must convey the evidence needed to answer it.

**Shared latent spaces.** Shared Latent Spaces let models reuse their read and write mappings across partners. Dery et al. (2026) demonstrate a shared KV space, new-member admission with existing adapters frozen, and transfer along untrained routes. C2C also explores shared representations with sender projections and receiver fusion modules (Fu et al., 2026). LLM-SemanticLAN instead fixes the coordinates of the space before any training, with closed-form maps onto targets from a frozen reference, so a new model aligns with one existing member rather than joining a joint training run.

**Hybrid-attention models.** Hybrid models require reuse of recurrent states as well as KV. HYPIC composes cached segment states, while LinearKV combines initialization from a matched segment with selective recomputation (Liu et al., 2026c;d). We transfer only full-attention KV across models and use fixed-prefix initialization for recurrent and convolutional states, which retains strong performance in our evaluations (Section 4.4).

**Vision-language communication.** Visual communication must preserve information that text descriptions may omit. Vision Wormhole aligns codecs through a shared reference space and transmits messages through VLM visual inputs (Liu et al., 2026b); $L^2$-VMAS maintains shared perceptual and reasoning memories (Yu et al., 2026). We study transferring into an existing text-model KV space, where established text decoders remain reusable across modalities.

## 3 Method

### 3.1 Public latent communication: definition and measurement

We consider a pool $\mathcal{M} = \{m_1, \ldots, m_N\}$ of frozen language models whose KV caches live in incompatible spaces, differing in depth, head count, head dimension, tokenizer, attention type, or input modality. Prefilling a $T$-token passage $x$ in member $m_i$ yields its KV cache $C_i(x)$ across all layers and heads, allowing $m_i$ to continue generation without rereading $x$. The public latent space is a per-token representation $\mathcal{H} = \mathbb{R}^{T\times d}$ whose width $d$ is a small fraction of the number of cache scalars per token in any member, and that carries no layer, head, or position structure. A model with $L$ layers, $H$ key-value heads, and head dimension $d_h$ stores $2LHd_h$ cache scalars per token, so each latent vector is $2LHd_h/d$ times smaller than the native cache it replaces. Each member owns a codec, an encoder $E_i$ from its cache into the latent space and a decoder $D_i$ from the latent space back to a full-depth cache in its own format (Figure 2). Communication from $m_i$ to $m_j$ uses

$$\hat{C}_j = D_j\big(E_i(C_i(x))\big), \qquad u_i = E_i(C_i(x)) \in \mathbb{R}^{T\times d}, \tag{1}$$

as the receiver's prefix cache, after which $m_j$ continues without ever reading $x$. The latent space is first built by a few initial models whose codecs are trained together, and every other model is added later by training its own codec against one of them (§3.5). The latent space belongs to no member (§3.3), so any member reaches any other by composing two codecs, and adding a model costs one codec trained against a single existing member (§3.5).

We measure communication quality by the receiver's performance across five conditions. Throughout evaluation, the sender processes only the passage and the receiver observes only the question. The five setups comprise answering with *No context*, reading the passage directly (*Receiver reads passage*), receiving the latent representation of the passage (*LLM-SemanticLAN*), receiving the latent representation of a different passage as a content control (*Mismatched passage*), and reading a note the sender writes based on the raw passage through text-to-text communication (*T2T*).

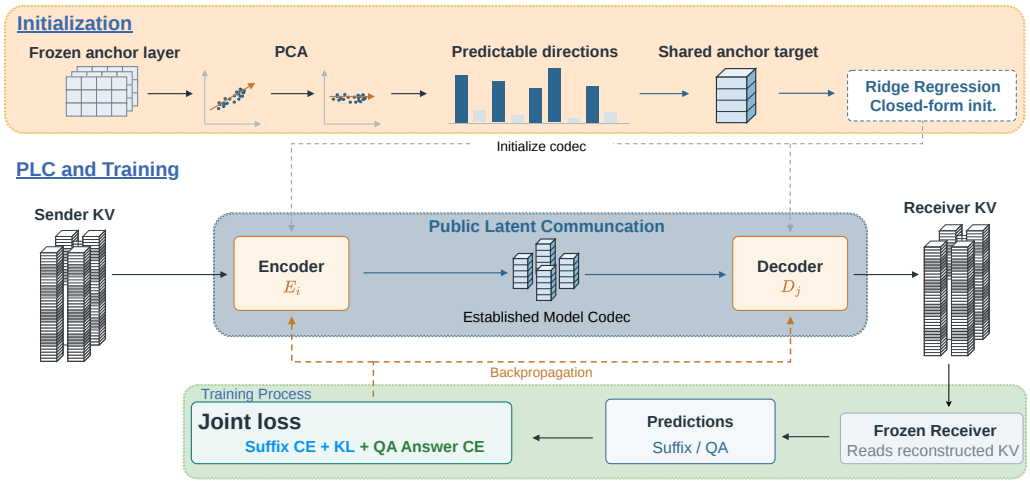



Figure 2: Overview of LLM-SemanticLAN. Latent coordinates are anchored in closed form by ridge regression before training, using targets from a frozen reference. Codecs carry the sender's cache to the frozen receiver, whose suffix and answer predictions train both codecs.

### 3.2 From linear maps to a trained latent space

Linear algebra can fix the latent coordinates but cannot build a working codec alone. Since the design below rests on this point, we test it before introducing the method, with the three models, Qwen3-4B, Qwen3-8B, and Qwen3-14B, on Web, our held-out questions on FineWeb-Edu passages (Penedo et al., 2024) (§4.1). Qwen3-8B and Qwen3-4B have caches of identical shape, yet Qwen3-4B answers only 42.6 percent of the questions from the unchanged cache of Qwen3-8B, against 95.1 percent when it reads the passage itself. A pairwise closed-form ridge map from Qwen3-14B to Qwen3-4B (Heo et al., 2026) reaches 79.2 percent, but such a map covers one direction of one pair, is undefined across tokenizers, and grows quadratically with the number of models. Routed through the latent space instead, the same closed form starts near the No context level and reaches 87.5 percent after training, 8.3 points above the pairwise map. The codecs are therefore nonlinear and trained from a closed-form initialization (Appendix D.2).

### 3.3 The latent space and the codecs

**Anchored coordinates.** The upper panel of Figure 2 anchors the latent coordinates to targets from a frozen reference model that takes no part in communication. An external reference gives each initial model a target it can solve against on its own, so all start in one coordinate system and none is privileged. For token $t$, let $\boldsymbol{r}_t \in \mathbb{R}^{d_{\text{ref}}}$ be its activation in one layer of the reference (Appendix D), standardized with the calibration mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ and rotated into principal directions,

$$\boldsymbol{h}_{\text{ref},t} = \boldsymbol{P}\,\text{diag}(\boldsymbol{\sigma})^{-1}\big(\boldsymbol{r}_t - \boldsymbol{\mu}\big), \qquad \boldsymbol{P} \in \mathbb{R}^{d \times d_{\text{ref}}}, \tag{2}$$

where $\boldsymbol{P}$ applies the PCA rotation and keeps $d$ directions. These are the directions that the caches of the initial models predict best by ridge regression on held-out tokens (Raghu et al., 2017), and $d$ is set where this predictability drops (Appendix D). After initialization neither the reference nor $\boldsymbol{P}$ is used again, so the capacity of the latent space is never bounded by what the reference can represent.

**Closed-form initialization.** Each codec starts from two linear maps solved by ridge regression on calibration passages (Appendix D.1). The encoder's input projection maps the model's selected cache features to the anchor targets $\boldsymbol{H}_{\text{ref}}$, and the decoder's output projection takes the anchor targets back to every layer and head of the model's own cache. Every initial model solves against the same targets, so all their codecs start in one coordinate system. Because the residual blocks start at zero, the codec at step zero is exactly this linear map, the starting point of §3.2.

**Encoder and Decoder.** The lower panel of Figure 2 follows one passage from sender to receiver. The encoder $E_i$ reads the keys and values of the $L_s$ sender layers whose masking raises the contin-

uation loss most (Liu et al., 2026e; Men et al., 2024) (Appendix E), removes the rotary embedding from the keys, and concatenates them into one vector per token. An input projection maps this vector into the $d$-dimensional latent space, and two residual blocks of windowed self-attention, initialized to output zero, let each latent vector draw on its neighbouring tokens. The decoder $D_i$ mirrors this, with two such blocks and an output projection to the keys and values of every layer and head of the receiver. Keys are output without rotary embedding and rotated at the prefix positions, so each receiver's decoder handles its own depth, head layout, and prefix offset.

### 3.4 Training objective

The codecs are trained by what the receiver does with a reconstructed cache, never by how closely that cache matches the receiver's own, so Eq. 3 contains no loss on the cache itself, no distillation of attention outputs, and no regularization toward the reference, a point-wise term being redundant beside a functional one and harmful when added (Dery et al., 2026). As in the lower panel of Figure 2, for a direction $i \to j$ and a passage $x$, the frozen receiver $m_j$ reads $\hat{C}_j$ of Eq. 1 as its prefix, the gradient flows through all of its layers into $D_j$, the latent representation, and $E_i$, and the sender is never loaded. The objective is

$$\mathcal{L} = \mathrm{KL}_w + \mathrm{CE}_w^{\text{suffix}} + \mathrm{CE}^{\text{answer}}, \tag{3}$$

three terms on the behaviour of the receiver, with the subscript $w$ denoting the information weighting below and all constants listed in Appendix D.

**Continuation.** The suffix cross entropy evaluates the receiver on the trailing text of the passage, providing a free self supervised signal without external annotations.

**Behaviour matching.** The KL term minimizes the distributional gap between two runs of the same receiver at every suffix token, one reading the reconstructed cache and the other reading the original text. The answer term is the cross-entropy of gold answers to questions about the passage. The answer loss supervises the receiver on passage queries, where copy prompts requiring verbatim span repetition compel the latent space to retain precise numerical and entity values.

**Information weighting.** The KL and suffix terms are weighted per token by how much the passage lowers the receiver's cross-entropy at that token, so the loss concentrates on the names, numbers, and relations the passage introduces rather than on text the receiver would predict anyway. Tokens the passage does not help receive no extra weight, and the exact form of the weight and its constants are given in Appendix D.

After the initial models are trained, their codecs are frozen. Every model added later aligns its own codec to this fixed latent space (§3.5).

### 3.5 Adding a model: align with one member, communicate with all

A model joins the defined latent space by training only its own codec against one existing model in the space, its partner. Its linear maps are solved by ridge regression against the latent vectors its partner writes, its residual blocks start at zero. These vectors lie in the coordinates anchored in §3.3, so aligning with one model places the added model in the same coordinates as all the others. Eq. 3 is applied to the two directions between the added model and its partner, while the partner's codec and every model in the latent space stay frozen. Every pair between the added model and the other members stays untrained, and these pairs are the test of whether the latent space is public. A domain-specialized model joins under this rule using the standard codec. For hybrid-attention receivers, we use *fixed linear-state initialization*. The recurrent and convolutional states are precomputed by the receiver from a single fixed prefix unrelated to the current passage and reused as the initialization for every example. This provides states produced by the model's normal computation, while passage-specific information is transferred through full-attention KV. The initialized states are then updated token by token as the receiver processes the question and generates its answer.

**Tokenizer.** Introducing a model with a distinct tokenizer alters the sequence layout of latent vectors without disrupting their semantic content. Both tokenizers cover the same bytes, so the latent positions a token of the added model spans follow from byte offsets and involve no learning. When a token spans several positions of its partner's grid, the encoder writes each position's vector with a small network dedicated to that position within the token. In the reverse direction, the decoder

applies an identity-initialized, position-specific projection to each vector and sums the outputs, preserving order that averaging would lose (Appendix F).

**Vision.** The vision model joins as a sender, and the only trained component is a visual encoder. It turns the vision model's own cache of the image into a fixed number of latent positions, which the frozen decoder of any text model reads as a prefix cache. The visual encoder is trained against a single text model with the KL and cross-entropy terms of Eq. 3. It is trained on two kinds of examples in equal parts. One kind is a question about the image with its answer, and the other is the instruction to describe the image with the description written for it as the target, which takes the place of the passage continuation. The KL teacher is the vision model itself, answering the same question or writing the same description while looking at the image.

# 4 Experiments

## 4.1 Experimental setup

**Models.** Seven models take part. Qwen3-4B-Instruct-2507, Qwen3-8B, and Qwen3-14B (Yang et al., 2025) initialize the latent space jointly. Llama-3.1-8B (Grattafiori et al., 2024), II-Medical-8B, Qwen3.5-9B, and Qwen3-VL-8B (Bai et al., 2025) join afterwards. Each added model trains only its own codec against one existing model in the defined latent space (§3.5).

**Training data.** The codecs train on 7,955 Web passages, drawn from FineWeb-Edu (Penedo et al., 2024), and 5,969 synthetic records with 63,868 questions. The encoder of Qwen3-VL-8B trains on 19,081 COCO images with 127,219 generated questions. Appendix G describes the detailed workflow of data construction.

**Evaluation sets.** Synthetic and Web are held-out runs of the training generators. Each Web passage is a page cut into a prefix and a suffix that reuses its facts. deepseek-v4-flash writes seven question types per passage. Every question must fail closed-book and pass open-book, and Web questions must also pass a world-knowledge gate. NewsQA (Trischler et al., 2017), HotpotQA (Yang et al., 2018), and the out-of-domain part of MRQA (Fisch et al., 2019) were never used in training, 500 questions each. COCO questions (Lin et al., 2014) are written from held-out validation images, never from the description another stronger frontier model pre-wrote for each image (Appendix G.4). RealWorldQA (xAI, 2024) and CV-Bench (Tong et al., 2024), used as released, were never used in training either. Table 10 lists every set in detail and Appendix G gives the complete pipelines.

**Conditions.** The five conditions of §3.1 run with the prompts of Appendix G.6. The transferred T2T note defined in §3.1 is generated from a fixed prompt, with its output length determined dynamically by the sender. For images the T2T message is the description the sender writes at its natural length. Pairwise cache maps serve only the pairs they were trained on, and no prior method reaches an untrained pair, so T2T is the reference throughout.

**Scoring and cost.** Deepseek-v4-flash judges whether an answer matches the gold answer. The judge prompts for text and image answers are given in Appendix G.6. Table 1 reports accuracy relative to the same receiver reading the T2T note, with absolute performance values detailed in Appendix A. Latency and FLOPs are measured end to end over the five phases of Appendix H.

## 4.2 Main results

**Overview.** One frozen latent space lets every model read the passage of every other model, including pairs never trained together, at the accuracy of T2T and a fraction of its cost (Figure 3, Table 1). Every text row of Table 1 averages 94 to 101 percent of T2T, untrained rows included. A mismatched passage stays at or below the No context level on every set, so what a receiver gains above that level comes from the passage it received.

**Trained and untrained pairs.** Figure 3 orders the pairs by how closely their codecs were trained together. On Web, the initial models, whose codecs were trained together, reach 93.9. Initial models sending to added models reach 94.5, and added models sending to initial models reach 96.7, although the codec of each added model was trained only against Qwen3-8B. The added models, each trained against Qwen3-8B alone and never with each other, still reach 98.8 on Web and 97 to 104 on the other four sets. Within the blocks between initial and added models, the six directions trained with

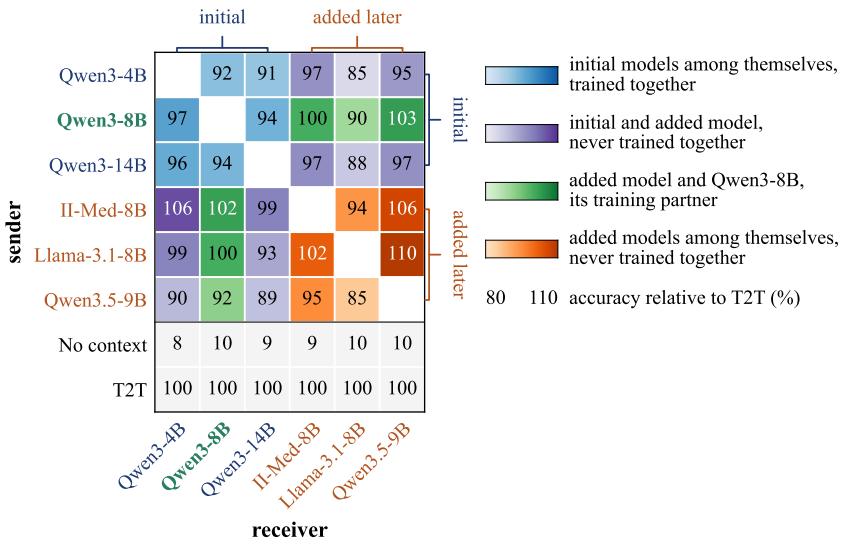



Figure 3: Every model communicates with all others through one latent space. Each cell is the accuracy on Web of the receiver reading the sender's passage. Colours mark whether the two codecs of a pair were trained together, and the bottom rows give No context and T2T on the same scale.

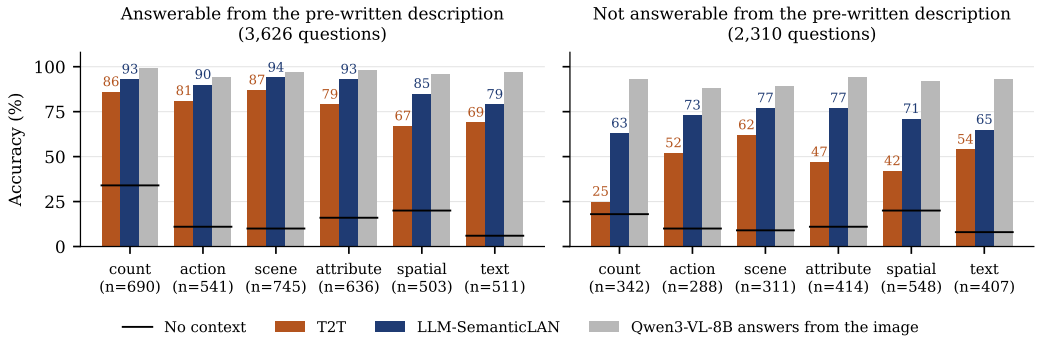



Figure 4: Accuracy on held-out COCO questions of six types, with Qwen3-14B as receiver. The left panel holds the questions that Qwen3-VL-8B can answer from the pre-written description of the image, and the right panel holds the rest. N is the number of each type of questions.

Qwen3-8B lead the untrained directions beside them by only 2 to 4 points. Whether a pair was trained together therefore does not predict how well it communicates, which makes the latent space public and lets each model join with a single codec (§3.5).

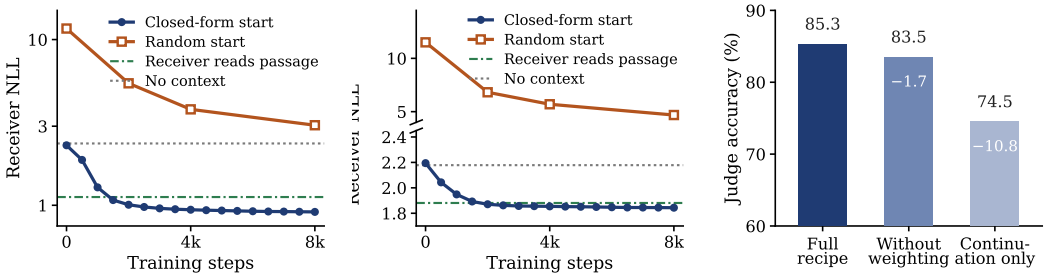
**Comparison with T2T.** The T2T note, written once without the question, is the strongest thing a sender can do without latent communication, and LLM-SemanticLAN matches it. On Synthetic it exceeds the note in every row except the untrained directions of Llama-3.1-8B, at 97 to 108, and on Web and the three public sets it retains 88 to 104 percent. Cells above 100 do not mean that LLM-SemanticLAN adds information, since Llama-3.1-8B and II-Medical-8B write shorter and more fragmentary notes than the Qwen models (Appendix B).

**Cost.** A codec holds 1.0 to 3.6 percent of its model's parameters. With $d = 1{,}024$, the ratio of §3.1 is 72 for Qwen3-4B and Qwen3-8B and 80 for Qwen3-14B (Appendix D). End to end, LLM-SemanticLAN answers 12–35× sooner with 2.2–4.3× fewer FLOPs. Both ratios grow with passage length, because the note grows with the passage and T2T must write it token by token and then have the receiver read it again (Appendix H).

Table 1: Main results. Top block covers text members. Bottom block covers Qwen3-VL-8B sending image states to text readers. Each cell gives a receiver's accuracy with LLM-SemanticLAN as a percentage of its accuracy with T2T. † marks untrained directions and bold marks cells at or above T2T. In-distribution sets are held out, and out-of-distribution sets were never used in training. Latency and FLOPs give the speedup and savings of LLM-SemanticLAN over T2T, ranging over the row's directions (Appendix H). Appendix A lists every direction.

| Property, members | Dirs. | In-distribution | | Out-of-distribution | | | Mean | Latency | FLOPs |
|---|---|---|---|---|---|---|---|---|---|
| | | Web | Synthetic | NewsQA | HotpotQA | MRQA | | faster | fewer |
| *LLM-SemanticLAN accuracy relative to T2T (%)* | | | | | | | | | |
| *Initial models, whose codecs were trained together* | | | | | | | | | |
| Scale, Qwen3-4B/8B/14B | 6 | 93.9 | **103.2** | 94.5 | 93.1 | 96.4 | 96.2 | 21–25× | 2.2–4.3× |
| *Added models, each trained against Qwen3-8B only* | | | | | | | | | |
| Domain, II-Medical-8B | 2 | **101.0** | **106.3** | 96.6 | 97.8 | **100.7** | **100.5** | 20–23× | 3.0–3.4× |
| | 4† | 99.9 | **105.8** | 98.7 | 97.9 | 99.3 | **100.3** | 18–24× | 2.5–4.1× |
| Family, Llama-3.1-8B | 2 | 94.8 | **103.8** | **102.8** | 88.6 | 94.4 | 96.9 | 13–32× | 2.6–2.8× |
| | 4† | 91.0 | 96.9 | **103.7** | 88.0 | 91.9 | 94.3 | 12–35× | 2.3–3.4× |
| Attention, Qwen3.5-9B | 2 | 97.2 | **107.5** | 97.8 | 95.1 | 95.5 | 98.7 | 19–29× | 3.0–3.2× |
| | 4† | 92.8 | **103.7** | 97.3 | 94.9 | 95.9 | 96.9 | 19–29× | 2.5–3.8× |
| Among added models | 6† | 98.8 | **103.6** | **104.1** | 96.9 | 98.0 | **100.3** | 20–29× | 3.2–3.4× |
| *Reference accuracy (%)* | | | | | | | | | |
| No context | – | 8.2 | 2.4 | 8.4 | 23.0 | 44.9 | 17.4 | – | – |
| Mismatched passage | – | 4.7 | 2.5 | 5.7 | 18.1 | 37.0 | 13.6 | – | – |
| T2T, note | 6 | 92.6 | 94.7 | 85.1 | 91.6 | 89.0 | 90.6 | 1 | 1 |

| Property, members | Dirs. | In-distribution | Out-of-distribution | | Mean | Latency | FLOPs |
|---|---|---|---|---|---|---|---|
| | | COCO | RealWorldQA | CV-Bench | | faster | fewer |
| *LLM-SemanticLAN accuracy relative to T2T (%)* | | | | | | | |
| *Added as a sender, trained against Qwen3-14B only* | | | | | | | |
| Modality, Qwen3-VL-8B | 1 | **123.9** | **111.7** | **119.5** | **118.4** | 10–62× | 1.1–3.6× |
| | 2† | **102.8** | 89.2 | 88.5 | 93.5 | | |
| *Reference accuracy (%)* | | | | | | | |
| No context | – | 15.3 | 33.5 | 48.5 | 32.4 | – | – |
| Mismatched image | – | 12.3 | 39.7 | 52.7 | 34.9 | – | – |
| T2T, description | 3 | 61.8 | 57.0 | 63.1 | 60.6 | 1 | 1 |




(a) Synthetic held-out (b) Web held-out (c) Loss terms at step 8,000

Figure 5: Ablations of the latent space training. (a, b) Receiver negative log-likelihood of held-out continuations as the codecs train from the closed form of §3.3 and from random weights, with the receiver reading the passage and the receiver without context as references. (c) Judge accuracy on Web at step 8,000, mean over the six directions among the initial models.

### 4.3 Adding a vision model

On COCO, LLM-SemanticLAN beats the T2T description for all three receivers, including the two never trained with the visual encoder (Table 1).

**Questions answerable from the description.** Each held-out COCO question is labelled by whether Qwen3-VL-8B can answer it from the pre-written description alone, which separates what a description carries from what it leaves out (Appendix G.4). On questions the pre-written description can answer, LLM-SemanticLAN matches or exceeds the T2T description in every question type (Figure 4). On the rest, the T2T description falls to 46.5 while LLM-SemanticLAN reaches 71.0. Qwen3-VL-8B answering from the image reaches 91.8, and the gain over the description lies in counts, attributes, and spatial relations. LLM-SemanticLAN therefore keeps what a description carries and recovers more than half of what it leaves out.

**Out-of-distribution benchmarks and cost.** On RealWorldQA and CV-Bench, Qwen3-14B, the receiver trained with the visual encoder, beats the T2T description on all six tasks, including depth and distance, which never occur in training (Appendix I). On these benchmarks the two receivers never trained with the visual encoder reach 88 to 89 percent of the description. Because the sender decodes no text, LLM-SemanticLAN answers 10 to 62 times sooner with 1.1 to 3.6 times fewer FLOPs (Appendix H).

### 4.4 Hybrid attention with fixed linear-state initialization

We evaluate fixed linear-state initialization in frozen Qwen3.6-27B while keeping passage-specific full-attention KV unchanged. Replacing the recurrent and convolutional states with the fixed-prefix initialization retains 100.3% of passage-state performance on average across Web, Synthetic, and the pooled HotpotQA, NewsQA, and DROP evaluation. Mean retention is 96.8% across the four non-code LongBench tasks. Matched code-completion controls further show strong performance after processing subsequent input. These results support transferring passage information through full-attention KV without separately translating recurrent states, the approach adopted for our Qwen3.5-9B member. Full results and controls appear in Appendix F.2.

### 4.5 Ablations

**Closed-form initialization.** Figure 5(a, b) follows the receiver's negative log-likelihood of held-out continuations as the codecs of the initial models train from the closed form of §3.3 and from random weights under the same objective and schedule. The closed form already places the codecs where the receiver stands without any context, 2.30 against 2.36 on Synthetic, so training only has to cover the distance to the receiver reading the passage, which it does within 1,500 steps on Synthetic and 2,000 on Web before settling at 0.91 after 8,000. From random weights the loss starts at 11.65, far outside that band, and after 8,000 steps it is still 3.04, above the receiver without context, while the codecs answer 0.8 percent of the Web questions that the closed-form start answers at 85.3 (Table 4). Training from the closed form only adjusts coordinates the closed form has already fixed, and the same loss trained from random weights never finds them.

**Loss terms.** Figure 5(c) removes parts of Eq. 3 for the six directions among the initial models at step 8,000 on Web. With the continuation term alone the receiver is asked only to continue the passage, and accuracy falls from 85.3 to 74.5, so continuing the text is not the same as holding its content, and the behaviour-matching terms are what carry the answers. Information weighting adds a small gain, 1.7 points at 8,000 steps and about the same at 2,000 and 4,000, while the ten-point gap left by the continuation term alone does not close with more training (Table 4).

## 5 Conclusion

We formalize public latent communication (PLC), in which heterogeneous language models communicate through a public latent space rather than translating between each pair. LLM-SemanticLAN realizes PLC by anchoring the coordinates of this space before any training, with closed-form maps onto targets from a frozen reference, and refines its codecs by the behaviour of the receiver. A model trained against a single existing member then communicates with every member it was never paired with, across scale, domain, family, attention architecture, and modality, at the accuracy of text-to-text communication and a fraction of its cost. These results suggest that PLC can turn a shared latent space from a model-specific adapter into common infrastructure for direct multi-model knowledge exchange.

### AI USE STATEMENT

We used deepseek-v4-flash, qwen3.8-flash, and kimi-k3 to generate and filter synthetic passages, questions, and image descriptions, and deepseek-v4-flash as the evaluation judge, as documented in Appendix G and Appendix G.6. We also used AI assistants for literature retrieval, feedback on research methodology, experimental design and result interpretation, manuscript drafting and revision, figure preparation, and implementation and debugging of experiment and analysis code. Their assistance included formulating loss functions and checking mathematical expressions for consistency and correctness. The authors reviewed the AI-assisted material and tested the code. We take responsibility for the final content of this work, including its text, claims, and artifacts.

### REPRODUCIBILITY STATEMENT

Appendix D lists every constant of the codecs, the closed-form initialization, and training, Appendix D.1 gives the closed-form solutions, Appendix G describes the construction of every dataset with the verbatim prompts of Appendix G.6 and the leakage checks of Appendix G.5, and Appendix H gives the cost protocol. All member models are publicly released checkpoints. Code for the codecs, the admission procedure, and the evaluation harness will be released with the evaluation sets.

## REFERENCES

Shuai Bai et al. Qwen3-vl technical report, 2025.

Andrei Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pp. 21–29, 1997.

Siyi Chen, Xiaoyan Zhang, Meng Wu, Jonathan Tremblay, Valts Blukis, Stan Birchfield, Rene Vidal, Alvaro Velasquez, Sijia Liu, and Qing Qu. See what i see, know what i think: Dense latent communication across heterogeneous agents, 2026. URL https://arxiv.org/abs/2606.13594.

Lucio M. Dery, Zohar Yahav, Henry Prior, Qixuan Feng, Jiajun Shen, and Arthur Szlam. Latent space communication via K-V cache alignment, 2026. URL https://arxiv.org/abs/2601.06123.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/474ada926b331d78f06d95e8913111cc-Abstract-Conference.html.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. In *Proceedings of Machine Learning and Systems*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/a66caa1703fe34705a4368c3014c1966-Abstract-Conference.html.

Aaron Grattafiori et al. The llama 3 herd of models, 2024.

Taekyung Heo, Rasoul Shafipour, Ritchie Zhao, Maximilian Golub, Mohammad Mahdi Kamani, Ritika Borkar, Makesh Tarun Chandran, Pantea Zardoshti, and Bita Darvish Rouhani. Cross-model KV cache transfer in LLM families: A closed-form linear mapping for prefill reuse, 2026. URL https://arxiv.org/abs/2608.03893. Version 1, 4 August 2026.

Junhao Hu, Wenrui Huang, Weidong Wang, Haoyi Wang, Tiancheng Hu, Zhang Qin, Hao Feng, Xusheng Chen, Yizhou Shan, and Tao Xie. EPIC: Efficient position-independent caching for

serving large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 24391–24402. PMLR, 2025. URL https://proceedings.mlr.press/v267/hu25j.html.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

Jiyao Liu, Qi Zhang, Yaoyi Jia, Ziwen Kan, and Song Wang. Dual-cache latent space communication between heterogeneous language models, 2026a. URL https://arxiv.org/abs/2608.20617.

Xiaoze Liu, Ruowang Zhang, Weichen Yu, Siheng Xiong, Liu He, Feijie Wu, Hoin Jung, Matt Fredrikson, Xiaoqian Wang, and Jing Gao. The Vision Wormhole: Latent-space communication in heterogeneous multi-agent systems. *arXiv preprint arXiv:2602.15382*, 2026b. URL https://arxiv.org/abs/2602.15382.

Yifei Liu, Juntong Wu, Yang Liu, Junhao Hu, Minghao Li, Xiaoxu Chen, and Weihang Chen. HYPIC: Accelerating hybrid-attention LLM serving with position-independent caching, 2026c. URL https://arxiv.org/abs/2607.01299.

Yirui Liu, Ruoling Qi, Longwen Wang, Xuaner Wu, Jian Chen, Yuxin Jin, Jiawei Shao, and Xuelong Li. LinearKV: One cached state suffices for position-independent caching in hybrid LLMs, 2026d. URL https://arxiv.org/abs/2608.11231.

Yuhan Liu, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, and Esha Choukse. DroidSpeak: KV cache sharing across fine-tuned model variants. In *23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26)*, pp. 319–338. USENIX Association, 2026e. URL https://www.usenix.org/conference/nsdi26/presentation/liu-yuhan.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect, 2024.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024.

Xingyu Qu, Siyuan Lu, Zhiyu Chen, Sheng Wang, and Tao Lin. Cachebridge: Efficient cross-model KV cache transfer, 2026. URL https://arxiv.org/abs/2609.00891.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://arxiv.org/abs/1706.05806.

Maximillian Rossi, Prajwal Raghunath, and Eugene Wu. Latent Cache Flow: Model-to-model communication without text, 2026. URL https://arxiv.org/abs/2605.22863.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017.

xAI. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v, 2024. RealWorldQA benchmark release.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. CacheBlend: Fast large language model serving for RAG with cached knowledge fusion. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 94–109, 2025.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Bo Yin, Cheng Yang, Yongbo He, Yihao Hu, Jiangning Zhang, Cheng Tan, Xiaobin Hu, and Shuicheng Yan. Dual latent memory for visual multi-agent system. *arXiv preprint arXiv:2602.00471*, 2026. URL https://arxiv.org/abs/2602.00471.

## A PER-DIRECTION RESULTS

Table 2 lists the absolute judge accuracy of every text direction behind Table 1, LLM-SemanticLAN beside the T2T note of the same sender on each set, together with each receiver alone. Table 3 does the same for the visual sender. Both are generated from the same cell file as Table 1.

## B WHERE LLM-SEMANTICLAN AND THE NOTE DIFFER

This appendix breaks down the comparison between LLM-SemanticLAN and T2T that Table 1 summarizes, on the six founding directions of Web and Synthetic. On Synthetic LLM-SemanticLAN exceeds the note because a note loses what the latent representation keeps. On coreference questions the founding members read 96.6 through the latent space and 78.6 through the note, whose writer drops who did what to whom, and relation questions show the same pattern by five points. On Web and the public sets LLM-SemanticLAN retains 88 to 104 percent of the note. The loss sits in exact values and in questions that span sentences, where it trails the note by 11 and 13 points, whereas facts and coreference are even and counts and relations lose 5 to 6. A member's self-loop, its own state through its own codec pair, bounds the whole chain. On Web it lies 5 to 11 points below the member reading the passage itself and within 2 points on Synthetic, so the remaining loss on natural text comes from reconstructing the passage from its latent representation rather than from alignment between members. Long digit strings are where reconstruction fails. On Web LLM-SemanticLAN reads 82.4 percent of answers with 10 to 14 digits and 56.5 percent of the 31 answers with 15 digits or more, against 96.2 and 90.9 through the note, while on Synthetic every digit bucket is above 96 on both routes.

Cells above 100 in Table 1 mark directions in which LLM-SemanticLAN carries more of the passage than the sender's own note. On Web this happens when Llama-3.1-8B or II-Medical-8B sends, and the cause lies in the note rather than in the latent representation. II-Medical-8B writes prose notes about 30% shorter than the passage, and Llama-3.1-8B is the clearer case. Under the same frozen instruction every writer answers in bullet points, but Llama-3.1-8B writes notes about a quarter shorter and more fragmentary than the Qwen writers, keeping the numbers while dropping who did what to whom and the antecedents of pronouns, so relation and coreference questions lose about 20 points and T2T into Qwen3-8B reaches 79.6 against 93.8 for a note written by Qwen3-14B. The loss sits in the note rather than in the reader, since Llama-3.1-8B reading its own notes scores 76.4 with the same deficit on relation questions. LLM-SemanticLAN transfers the passage representation and does not depend on the sender's writing habits. In absolute terms the latent representation sent by Llama-3.1-8B still trails those sent by the Qwen members by about 13 points on Web.

Table 2: Per-direction results behind Table 1, text members. Absolute judge accuracy (%) of the receiver, LLM-SemanticLAN (Latent) beside the T2T note written by the same sender (Text), on every set; Web is the merged set of 1,459 questions. A check marks a trained direction and every other direction was never trained. The reference block gives each receiver with no context and reading the passage itself.

| Direction | Trained | Web | | Synthetic | | NewsQA | | HotpotQA | | MRQA | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Latent | Text | Latent | Text | Latent | Text | Latent | Text | Latent | Text |
| *Founding members* | | | | | | | | | | | |
| Qwen3-4B → Qwen3-8B | ✓ | 86.4 | 94.2 | 97.4 | 95.5 | 78.2 | 84.3 | 85.6 | 92.4 | 85.8 | 89.6 |
| Qwen3-4B → Qwen3-14B | ✓ | 84.9 | 93.6 | 97.5 | 95.8 | 81.7 | 84.5 | 84.8 | 92.8 | 85.6 | 89.2 |
| Qwen3-8B → Qwen3-4B | ✓ | 87.5 | 89.8 | 97.8 | 93.2 | 80.8 | 82.3 | 84.8 | 89.8 | 84.6 | 87.6 |
| Qwen3-8B → Qwen3-14B | ✓ | 86.9 | 92.2 | 97.7 | 94.9 | 79.7 | 87.2 | 84.2 | 93.0 | 86.2 | 90.0 |
| Qwen3-14B → Qwen3-4B | ✓ | 87.9 | 91.9 | 98.1 | 94.1 | 81.5 | 85.9 | 86.6 | 90.6 | 85.0 | 88.2 |
| Qwen3-14B → Qwen3-8B | ✓ | 87.8 | 93.8 | 98.1 | 94.8 | 80.4 | 86.5 | 85.6 | 91.0 | 87.2 | 89.2 |
| *Llama-3.1-8B* | | | | | | | | | | | |
| Qwen3-4B → Llama-3.1-8B | | 78.3 | 92.6 | 91.7 | 94.8 | 78.2 | 80.4 | 79.8 | 91.4 | 86.0 | 89.0 |
| Qwen3-8B → Llama-3.1-8B | ✓ | 81.4 | 90.7 | 96.2 | 94.6 | 80.6 | 82.4 | 81.8 | 92.6 | 85.6 | 89.6 |
| Qwen3-14B → Llama-3.1-8B | | 80.5 | 92.0 | 93.1 | 95.3 | 80.2 | 85.5 | 85.0 | 91.0 | 84.8 | 89.6 |
| Llama-3.1-8B → Qwen3-4B | | 74.5 | 75.3 | 89.7 | 83.6 | 71.4 | 64.5 | 73.5 | 83.6 | 71.6 | 84.4 |
| Llama-3.1-8B → Qwen3-8B | ✓ | 79.5 | 79.6 | 94.8 | 89.5 | 74.9 | 69.5 | 78.0 | 87.8 | 81.8 | 87.8 |
| Llama-3.1-8B → Qwen3-14B | | 73.5 | 79.0 | 76.3 | 88.7 | 73.8 | 65.3 | 73.9 | 88.8 | 79.8 | 87.2 |
| *II-Medical-8B* | | | | | | | | | | | |
| Qwen3-4B → II-Medical-8B | | 85.9 | 88.9 | 97.5 | 90.7 | 78.4 | 79.5 | 84.4 | 86.8 | 84.8 | 86.4 |
| Qwen3-8B → II-Medical-8B | ✓ | 86.5 | 86.6 | 97.3 | 89.6 | 76.4 | 81.1 | 85.4 | 87.0 | 83.2 | 83.8 |
| Qwen3-14B → II-Medical-8B | | 86.3 | 88.6 | 97.2 | 89.7 | 80.8 | 78.1 | 86.4 | 78.6 | 84.6 | 83.6 |
| II-Medical-8B → Qwen3-4B | | 87.5 | 82.5 | 97.2 | 92.3 | 79.5 | 80.2 | 86.0 | 90.8 | 84.8 | 84.4 |
| II-Medical-8B → Qwen3-8B | ✓ | 88.6 | 86.8 | 98.0 | 94.3 | 80.4 | 81.2 | 88.2 | 90.6 | 87.4 | 85.6 |
| II-Medical-8B → Qwen3-14B | | 86.6 | 87.2 | 96.5 | 94.6 | 78.6 | 83.9 | 83.0 | 92.4 | 85.4 | 87.6 |
| *Qwen3.5-9B* | | | | | | | | | | | |
| Qwen3-4B → Qwen3.5-9B | | 86.2 | 90.3 | 96.9 | 89.4 | 80.1 | 77.1 | 83.8 | 85.4 | 82.4 | 85.0 |
| Qwen3-8B → Qwen3.5-9B | ✓ | 87.3 | 85.1 | 97.6 | 84.9 | 80.6 | 76.4 | 85.2 | 86.4 | 84.0 | 86.0 |
| Qwen3-14B → Qwen3.5-9B | | 87.9 | 90.8 | 97.5 | 90.7 | 80.8 | 78.6 | 86.0 | 87.2 | 84.4 | 85.2 |
| Qwen3.5-9B → Qwen3-4B | | 84.2 | 93.4 | 96.4 | 96.2 | 78.9 | 85.4 | 84.0 | 91.2 | 81.6 | 86.8 |
| Qwen3.5-9B → Qwen3-8B | ✓ | 86.1 | 93.7 | 97.0 | 96.9 | 77.3 | 85.7 | 83.6 | 91.2 | 83.8 | 89.8 |
| Qwen3.5-9B → Qwen3-14B | | 83.4 | 93.8 | 96.1 | 97.4 | 77.8 | 86.3 | 85.0 | 93.6 | 83.8 | 89.6 |
| *Added members among themselves* | | | | | | | | | | | |
| Llama-3.1-8B → II-Medical-8B | | 76.0 | 74.3 | 92.6 | 86.7 | 72.8 | 67.4 | 73.3 | 84.0 | 77.6 | 83.0 |
| Llama-3.1-8B → Qwen3.5-9B | | 81.2 | 74.0 | 85.0 | 78.4 | 77.3 | 64.0 | 82.0 | 78.0 | 82.0 | 80.6 |
| II-Medical-8B → Llama-3.1-8B | | 79.1 | 84.4 | 93.7 | 93.7 | 80.0 | 83.0 | 84.0 | 86.2 | 85.2 | 86.6 |
| II-Medical-8B → Qwen3.5-9B | | 87.5 | 82.2 | 97.1 | 93.3 | 79.7 | 74.4 | 85.6 | 86.2 | 83.4 | 83.2 |
| Qwen3.5-9B → Llama-3.1-8B | | 79.6 | 93.3 | 95.0 | 96.9 | 81.7 | 85.2 | 83.8 | 90.6 | 85.4 | 89.6 |
| Qwen3.5-9B → II-Medical-8B | | 85.3 | 89.7 | 97.1 | 93.3 | 78.4 | 81.2 | 85.4 | 85.4 | 82.6 | 83.6 |
| *Reference: receiver alone* | | | | | | | | | | | |
| Qwen3-4B, no context | | 6.8 | | 1.3 | | 7.7 | | 18.6 | | 41.6 | |
| Qwen3-4B, reads the passage | | 95.2 | | 98.0 | | 92.9 | | 91.4 | | 88.2 | |
| Qwen3-8B, no context | | 8.8 | | 2.9 | | 7.9 | | 23.6 | | 47.1 | |
| Qwen3-8B, reads the passage | | 96.0 | | 98.7 | | 88.5 | | 91.2 | | 90.4 | |
| Qwen3-14B, no context | | 8.0 | | 1.8 | | 8.6 | | 25.0 | | 48.4 | |
| Qwen3-14B, reads the passage | | 95.6 | | 98.6 | | 87.9 | | 92.8 | | 90.0 | |
| Llama-3.1-8B, no context | | 9.3 | | 2.9 | | 9.0 | | 29.6 | | 45.2 | |
| Llama-3.1-8B, reads the passage | | 93.6 | | 97.6 | | 87.0 | | 92.4 | | 89.8 | |
| II-Medical-8B, no context | | 7.8 | | 3.4 | | 8.6 | | 16.6 | | 40.2 | |
| II-Medical-8B, reads the passage | | 89.9 | | 96.8 | | 80.2 | | 75.2 | | 84.6 | |
| Qwen3.5-9B, no context | | 8.6 | | 2.0 | | 8.4 | | 24.8 | | 46.6 | |
| Qwen3.5-9B, reads the passage | | 94.9 | | 98.5 | | 85.0 | | 83.8 | | 87.6 | |

Table 3: Per-receiver results behind the bottom sub-table of Table 1. Qwen3-VL-8B sends latent representations of images and each receiver answers from them (Latent) or from the written description of the same image (Text); multiple-choice answers are forced to an option letter under both conditions.

| Reader | Trained | COCO | | RealWorldQA | | CV-Bench | |
|---|---|---|---|---|---|---|---|
| | | Latent | Text | Latent | Text | Latent | Text |
| Qwen3-VL-8B → Qwen3-14B | ✓ | 82.4 | 66.5 | 62.2 | 55.7 | 73.0 | 61.1 |
| Qwen3-VL-8B → Qwen3-8B | | 58.4 | 62.4 | 51.6 | 59.3 | 57.8 | 64.2 |
| Qwen3-VL-8B → Qwen3-4B | | 63.4 | 56.6 | 51.1 | 55.9 | 55.7 | 64.1 |
| No context (Qwen3-14B) | | 15.3 | | 33.5 | | 48.5 | |

## C Loss terms by question type

Table 4 gives the loss-term variants of §4.5 at 2,000, 4,000, and 8,000 steps on Web and at 2,000 steps on Synthetic, with the questions whose answer is a number or code copied from the passage listed separately. The gap of the continuation-only variant does not close with training. On Web it trails the full recipe by 10.9, 9.4, and 10.8 points at the three steps, and on exact values the gap widens from 7.1 to 9.7. The random start reaches 0.8 on Web after 8,000 steps. Information weighting is a small and stable gain of 1.4 to 1.7 points on Web at 4,000 and 8,000 steps, with intervals that exclude zero, while on exact values the difference is no longer significant.

Table 4: Loss terms by step and question type. Judge accuracy (%) on answerable questions, mean over the six directions among the initial models, difference to the full recipe in parentheses. Exact value marks the questions whose answer is a number or code copied from the passage. Synthetic was scored at 2,000 steps.

| | Web, 8,000 steps | | Web, 4,000 steps | | Web, 2,000 steps | | Synthetic, 2,000 steps | |
|---|---|---|---|---|---|---|---|---|
| Variant | All | Exact value | All | Exact value | All | Exact value | All | Exact value |
| Full recipe | 85.3 | 74.6 | 82.5 | 69.5 | 78.4 | 62.5 | 92.4 | 91.0 |
| Without weighting | 83.5 (−1.7) | 72.5 (−2.0) | 81.2 (−1.4) | 67.9 (−1.5) | 75.7 (−2.7) | 58.9 (−3.6) | 90.4 (−2.0) | 86.7 (−4.3) |
| Continuation only | 74.5 (−10.8) | 64.9 (−9.7) | 73.1 (−9.4) | 63.2 (−6.2) | 67.4 (−10.9) | 55.3 (−7.1) | 88.5 (−3.9) | 90.3 (−0.7) |
| Random start | 0.8 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 |
| No context | 8.1 | 2.0 | 8.1 | 2.0 | 8.1 | 2.0 | 2.0 | 0.0 |
| Reads passage | 95.3 | 96.9 | 95.3 | 96.9 | 95.3 | 96.9 | 98.5 | 99.7 |

## D Training and initialization settings

Table 5 lists the constants used for the founding codecs and for each added model; the Method section names only the structural ones.

The information weighting of §3.4 multiplies the KL and suffix terms at token $t$ by

$$w_t = 1 + \alpha \, \min\big(\max(\Delta_t, 0), \, \Delta_{\max}\big), \qquad \Delta_t = \ell_t^{\text{no passage}} - \ell_t^{\text{passage}}, \tag{4}$$

where $\ell_t$ is the receiver's cross-entropy at token $t$ with and without the passage, so a token the passage does not help has $\Delta_t \le 0$ and weight one.

Table 6 applies the ratio of §3.1 with $d = 1{,}024$ to every model that communicates through the latent space.

### D.1 Closed-form solutions

For member $m_i$, let $\boldsymbol{Z}_i \in \mathbb{R}^{n \times D_i}$ stack the selected cache features over $n$ aligned calibration tokens. Each row contains the counter-rotated keys and unchanged values of the sixteen selected layers, and for the founding members $D_i = 32{,}768$. Let $\boldsymbol{H}_{\text{ref}} \in \mathbb{R}^{n \times d}$ contain the corresponding reference projections, and let $\boldsymbol{C}_i^{(\ell,h)} \in \mathbb{R}^{n \times 2d_{\text{head}}}$ contain the pre-RoPE keys and unchanged values of receiver layer $\ell$ and KV head $h$.

Table 5: Constants of the codecs, the closed-form initialization, and training. Founding members are Qwen3-4B, 8B, and 14B. Rows for the added models list only what differs from the founding recipe.

| Item | Value |
|---|---|
| Latent width $d$ | 1,024 of the 2,048 standardized principal directions of Qwen3-1.7B layer 22, ranked by held-out ridge $R^2$ averaged over the three initial models (474,000 training and 53,000 held-out tokens); predictability falls away sharply after about 800 directions; retained directions have mean $R^2$ 0.61 and minimum 0.44, the best discarded direction 0.44, and the retained set covers 81.6% of the standardized variance |
| Encoder input | 16 functional layers, K and V, keys counter-rotated; 32,768 dims per token |
| Encoder / decoder blocks | 2 windowed self-attention residual blocks each (8 heads, window 128, MLP ratio 2), zero-initialized outputs |
| Decoder output | all layers and heads of the receiver (73,728 or 81,920 values per token), keys re-rotated at receiver positions |
| Codec parameters | 142.7M per member (54.6M encoder, 88.2M decoder); 151.1M for the 14B member |
| Closed-form calibration | 500 passages (≈0.5M tokens); ridge coefficient chosen on validation, insensitive over four orders of magnitude |
| Loss weights | $\Delta_{\max} = 1.5$ nats; $\alpha = 3.26$, chosen so that tokens with at least one nat of gain carry half of the total weight; KL, suffix, and answer terms unweighted relative to each other |
| Direction sampling | one of 9 directions per micro-batch (3 self-loops, 6 cross); 16 passages per step (12 continuation, 4 question) |
| Training data | 13,924 passages (5,969 synthetic, 7,955 Web lineage); 4 questions per passage share one reconstructed cache; copy prompts mined at the 65th percentile of partner token loss |
| Optimizer | AdamW ($\beta$ = 0.9, 0.95), weight decay 0.01, grad clip 1.0; lr $10^{-5}$ for closed-form layers, $10^{-4}$ for new blocks; 200 warmup steps, cosine |
| Steps | 16,000, cosine schedule restarted every 4,000 steps; about 15 h per 4,000 steps on three A100 GPUs |
| Llama-3.1-8B joining | partner Qwen3-8B; byte-offset correspondence; up to 8 sub-positions; per-sub-position ridge init and two-layer split networks (hidden 1,024, zero-initialized output) |
| Qwen3.5-9B joining | 8 full-attention layers ([3, 7, 11, 15, 19, 23, 27, 31]) carried by the latent space; fixed recurrent and convolutional state from one 256-token passage |
| Qwen3-VL-8B visual encoder | 217M parameters; first linear 73,728→2,560 per image position (ridge start), 2 query blocks; 256 latent positions; partner Qwen3-14B; KL + CE on answer rows and description rows, half each; 4 questions per image share one prefix |

Table 6: Cache scalars per token and size ratio to one latent vector. Qwen3.5-9B counts only its eight full-attention layers, since its recurrent state is fixed and not transmitted. Mean is unweighted over the seven models.

| Model | $L$ | $H$ | $d_h$ | $2LHd_h$ | Ratio |
|---|---|---|---|---|---|
| Qwen3-4B | 36 | 8 | 128 | 73,728 | 72 |
| Qwen3-8B | 36 | 8 | 128 | 73,728 | 72 |
| Qwen3-14B | 40 | 8 | 128 | 81,920 | 80 |
| II-Medical-8B | 36 | 8 | 128 | 73,728 | 72 |
| Llama-3.1-8B | 32 | 8 | 128 | 65,536 | 64 |
| Qwen3.5-9B | 8 | 4 | 256 | 16,384 | 16 |
| Qwen3-VL-8B | 36 | 8 | 128 | 73,728 | 72 |
| Mean | | | | | 64 |

Write $\widetilde{\boldsymbol{Z}}_i$, $\widetilde{\boldsymbol{H}}_{\mathrm{ref}}$, and $\widetilde{\boldsymbol{C}}_i^{(\ell,h)}$ for the corresponding mean-centered matrices. The encoder and decoder weights are

$$\boldsymbol{W}_E^{(i)} = \left(\widetilde{\boldsymbol{Z}}_i^\top \widetilde{\boldsymbol{Z}}_i + \lambda_E \boldsymbol{I}\right)^{-1} \widetilde{\boldsymbol{Z}}_i^\top \widetilde{\boldsymbol{H}}_{\mathrm{ref}}, \tag{5}$$

$$\boldsymbol{W}_D^{(i,\ell,h)} = \left(\widetilde{\boldsymbol{H}}_{\mathrm{ref}}^\top \widetilde{\boldsymbol{H}}_{\mathrm{ref}} + \lambda_D \boldsymbol{I}\right)^{-1} \widetilde{\boldsymbol{H}}_{\mathrm{ref}}^\top \widetilde{\boldsymbol{C}}_i^{(\ell,h)}. \tag{6}$$

The corresponding biases recover the feature means:

$$\begin{aligned} \mathbf{b}_E^{(i)} &= \bar{\mathbf{h}}_{\mathrm{ref}} - \left(\boldsymbol{W}_E^{(i)}\right)^\top \bar{\mathbf{z}}_i, \\ \mathbf{b}_D^{(i,\ell,h)} &= \bar{\mathbf{c}}_i^{(\ell,h)} - \left(\boldsymbol{W}_D^{(i,\ell,h)}\right)^\top \bar{\mathbf{h}}_{\mathrm{ref}}. \end{aligned} \tag{7}$$

A single streaming pass accumulates the counts, first moments, Gram matrices, and cross-products required by these systems. The $d \times d$ decoder system is shared across all layers and heads, so it is factorized once and solved with multiple right-hand sides.

The zero-output residual blocks leave this linear route unchanged at step zero.

### D.2 Linear baselines of Section 3.2

Table 7 gives the measurements behind §3.2. Every row uses Qwen3-4B as receiver on the answerable Web questions, scored by the judge of §4.1. The unchanged cache of Qwen3-8B is placed in Qwen3-4B without any mapping, which the two models allow because they share layer count, head count, and head dimension. The pairwise map follows Heo et al. (2026), with one ridge regression per target layer and head at $\lambda = 0.01$, keys mapped before rotary embedding, and all 40 layers of Qwen3-14B as sources, the number chosen on held-out continuation loss. The two LLM-SemanticLAN rows use the codecs of Qwen3-14B and Qwen3-4B at closed-form initialization and after training.

Table 7: Linear baselines of Section 3.2, Qwen3-4B receiver on the answerable Web questions, judge accuracy (%).

| Condition | Accuracy (%) |
|---|---|
| No context | 7.2 |
| Qwen3-4B reads the passage | 95.1 |
| Unchanged cache of Qwen3-8B | 42.6 |
| Pairwise ridge map from Qwen3-14B | 79.2 |
| T2T from Qwen3-14B | 92.0 |
| LLM-SemanticLAN from Qwen3-14B, closed-form initialization | 3.3 |
| LLM-SemanticLAN from Qwen3-14B, after training | 87.5 |

## E Layer selection

Figure 6 shows which layers each initial model selects as encoder input and why. For each member the cache of one layer at a time is zeroed over the passage positions of 100 held-out Web passages and the continuation loss is scored, with retention defined as $100\,(\ell_{\text{none}} - \ell_{\text{masked}})/(\ell_{\text{none}} - \ell_{\text{full}})$, where the three losses are those with the full cache, with the layer masked, and with no passage at all, so that 100 means the layer costs nothing when masked and 0 means it costs as much as the whole passage. The deep layers carry the passage in all three members, masking the first seven layers together retains 86 to 92 percent, and the sixteen layers whose masking costs most are the encoder's input, layers 7, 9, 13, 22, and 24 to 35 for Qwen3-4B, layers 7, 13, 19, and 22 to 34 for Qwen3-8B, and layers 13, 14, 21, 23, 24, and 27 to 37 for Qwen3-14B. The seventeenth layer raises the loss by at most 0.013 nat in any member.

## F Adding heterogeneous members

This appendix gives the mechanism of the three extra components of §3.5.

### F.1 Tokenizer correspondence

Over our English and Chinese passages the Llama and Qwen segmentations disagreed on at most 0.025% of spans. Ordinary words are one-to-one in 99.9% of cases, but Llama groups digits three at a time, so only 11.9% of digit tokens are, and one Llama token may have to produce up to eight

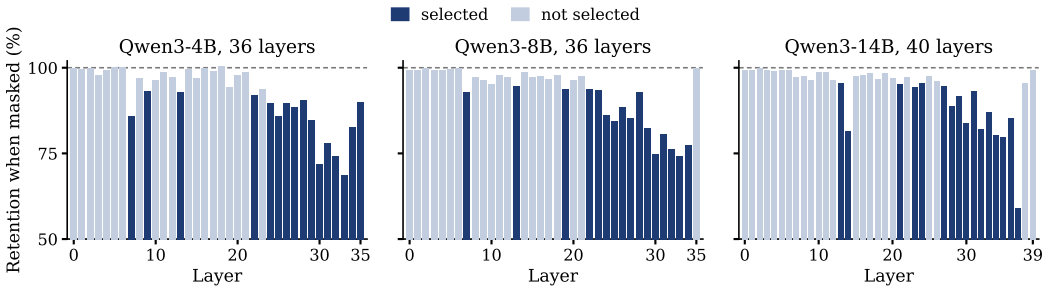



Figure 6: Retention of the continuation likelihood when the cache of a single layer is masked, on 100 held-out Web passages. Dark bars are the sixteen layers selected as encoder input.

latent vectors. For sub-position $k$ a small two-layer network with a zero-initialized output writes the $k$-th vector from the token's representation, starting from a per-sub-position ridge solution against the latent vectors Qwen3-8B writes. These networks are indexed by position within the token and not by token identity, so unseen strings, numbers, and names pass through the same networks. In the reverse direction the decoder passes each covered latent vector through the projection of its position within the token, initialised to the identity, and sums the results, so that the digits of a three-digit token keep their order where a plain average across the covered positions would discard it. The split is nonlinear because recovering the digits of a three-digit token is a lookup, which a linear average across the covered positions cannot perform without losing their order. The untrained directions in Table 1 and the Web set, which lies outside the training corpus, test whether this generalizes.

### F.2 Hybrid attention with a fixed donor state

**Setup.** We evaluate state replacement in frozen Qwen3.6-27B. Both conditions retain elementwise-identical, passage-specific native KV caches in all 16 full-attention layers. The remaining 48 layers use either recurrent and convolutional states computed from the current passage (*passage state*) or states obtained from one fixed, unrelated 256-token donor passage (*fixed donor*). The donor is selected independently of evaluation performance from data disjoint from the evaluation passages. It supplies an initialization produced by the model's normal computation; the installed states then update normally during subsequent input processing and generation. No codec or cross-model transfer is used in this control.

**QA and long-context evaluation.** Table 8 reports passage-state and fixed-donor performance. For each evaluation, retention is $100\, S_{\text{fixed}}/S_{\text{passage}}$. Averaging the retention values for Web, Synthetic, and the pooled public QA evaluation gives 100.3%. Across the four non-code LongBench tasks, mean task-level retention is 96.8%.

The public QA evaluation comprises 500 examples each from HotpotQA, NewsQA, and DROP. Both conditions use greedy decoding with a 64-token budget, suppress thinking tags, and score the first nonempty answer line with the same LLM judge. Passage-state and fixed-donor accuracy are respectively 88.20% and 89.60% on HotpotQA, and 63.80% and 59.80% on DROP. NewsQA has 398 and 419 confirmed correct answers, with two judgments missing in each condition. The aggregate in Table 8 uses the 1,498 examples judged in both conditions. Keeping the full 1,500-example denominator instead gives accuracy bounds of 77.20–77.33% and 77.73–77.87%. The Web and Synthetic evaluations retain their original protocols and are reported separately.

LongBench uses the official prompts, generation budgets, and task-specific scorers, with 200 HotpotQA, 200 2WikiMQA, 150 MultiFieldQA-en, and 200 MultiNews examples per condition.

**Code completion.** We evaluate all 500 examples of LCC and RepoBench-P with 1,102 input tokens processed after state installation. For each example, the passage-state and fixed-donor conditions use identical token sequences and native full-attention KV. Table 9 reports the official code-similarity scores on unmodified generations. Because this metric is sensitive to leading thinking output, we also report a diagnostic that removes at most one complete leading `<think>` block.

Table 8: Fixed-donor state control in Qwen3.6-27B. QA entries are LLM-Judge accuracy (%). LongBench entries use token F1 for QA and Rouge-L for MultiNews. Retention is relative to passage-specific states. The public QA aggregate includes 1,498 judged pairs.

| Evaluation | Passage state | Fixed donor | Retention (%) |
|---|---|---|---|
| Web QA ($n = 2{,}954$) | 93.70 | 91.84 | 98.01 |
| Synthetic QA ($n = 4{,}771$) | 95.95 | 98.16 | 102.30 |
| Public QA (1,498 judged) | 77.30 | 77.84 | 100.69 |
| HotpotQA (LongBench) | 70.70 | 66.63 | 94.24 |
| 2WikiMQA | 69.71 | 65.30 | 93.67 |
| MultiFieldQA-en | 58.22 | 58.94 | 101.24 |
| MultiNews | 23.95 | 23.51 | 98.16 |

Under this diagnostic, the fixed donor retains 94.70% of passage-state performance on LCC and 95.90% on RepoBench-P.

**Numerical checks.** The code-completion evaluation verifies exact token-sequence reconstruction, full-attention KV equality, and state installation. Across the matched-boundary audit, two of 2,000 direct-versus-split native next-token KL checks exceeded the original 0.02 threshold, with a maximum of 0.0433. Both retained the same top-1 prediction. A documented amendment retained 0.02 as a warning threshold and used 0.10 as an abort threshold, without changing experimental inputs, generation settings, or scoring.

**Application to the hybrid member.** These results support using a model-generated donor state as a reusable initialization while passage information is carried by full-attention KV. For Qwen3.5-9B, the hub translates the caches of its eight full-attention layers, while its other 24 layers initialize their recurrent and convolutional states from a fixed donor. Those states evolve normally as the receiver processes the question and generates its answer. The Qwen3.6-27B control isolates state replacement; the Qwen3.5-9B admission evaluation measures the complete cross-model communication pipeline.

### F.3 Visual writer

Its input is the vision-language model's cache of the image at native resolution, about 260 to 300 image positions for a COCO photograph, each carrying the keys and values of all 36 layers with rotary encoding removed, 73,728 values per position. A per-position linear layer, started from a ridge solution against the latent vectors Qwen3-14B writes, maps each position to the latent width and adds a term for its row and column in the image grid. Then 256 learned query vectors pass through two blocks of cross-attention over the image positions, self-attention among the queries, and a feed-forward layer. Every residual branch is initialised to zero, so the writer starts out image-independent and training adds only the dependence on the image. The output is a fixed sequence of 256 latent positions, which the frozen decoder of any text member turns into a prefix cache installed layer by layer before the receiver reads the question. Each training step draws answer rows and description rows in equal measure, and four questions about one image share one written state, so the writer has to write the whole image rather than the answer to one question.

## G Datasets, prompts, and evaluation protocol

Figure 7 summarizes the two construction pipelines and Table 10 lists every evaluation set. Writer and judge roles are split across model vendors. On the text side deepseek-v4-flash writes the synthetic passages and every question, qwen3.8-flash judges passages and questions, and deepseek-v4-flash is the evaluation judge. On the image side qwen3.8-flash writes descriptions, questions, and answer keys, kimi-k3 answers questions blind and judges description sentences, and deepseek-v4-flash judges answer equivalence.

Table 9: Fixed-donor code-completion results with 1,102 input tokens processed after state installation, using 500 examples per task. Raw denotes official scoring of unmodified generations; clean removes one complete leading thinking block as a diagnostic.

| | Raw | | Clean | | |
|---|---|---|---|---|---|
| Task | Passage | Fixed | Passage | Fixed | Clean retention (%) |
| LCC | 47.98 | 54.39 | 70.70 | 66.95 | 94.70 |
| RepoBench-P | 45.12 | 52.90 | 66.86 | 64.12 | 95.90 |

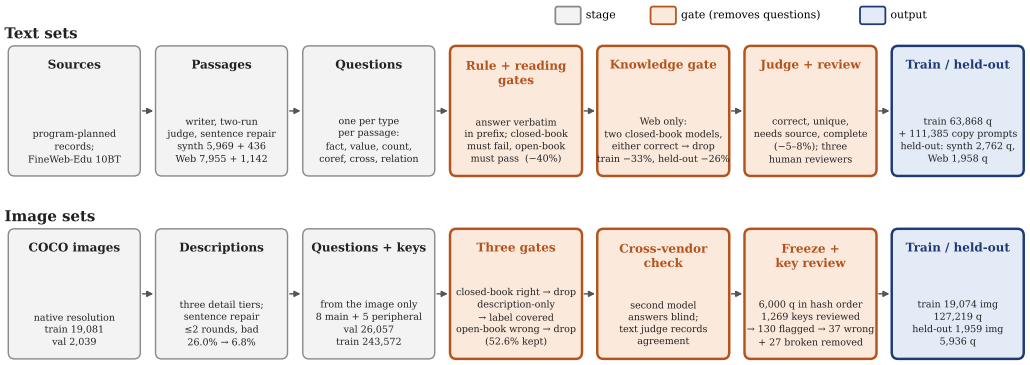



Figure 7: Construction of the text sets (top) and the image sets (bottom). Boxes with a darker frame are gates that remove questions, and the counts give passages and questions entering or leaving each stage. The text pipeline serves both the training corpus and the held-out evaluation sets, which are generated by the same procedure in independent runs and checked for overlap as described in §G.5.

### G.1 Synthetic records

The text pipeline of Figure 7 starts from a program that fixes the setting of each passage: one of twelve scenarios (insurance, construction, student records, legal, retail, clinical, maritime, research, agriculture, logistics, equipment, sports), three persons with name, age, and role, and nine planted values. The values cover account numbers, amounts, percentages, dates, measurements, ratios, identifiers, and times of day, with the value family, unit, magnitude, and institution bound to the scenario. A language model then writes a six-paragraph prefix that introduces every person, organisation, and value, and a four-paragraph suffix that reuses at least five of the values character for character. The suffix may neither name the persons in full nor restate the background, so it cannot be read without the prefix. A passage judge checks every planted value and the plausibility of the document in two independent passes at temperature 0, and a defect counts only with a verbatim quote from the document. Passages that fail on a single sentence are repaired at the sentence level rather than discarded (640 training passages with 1,419 sentences, 38 passages dropped) and rechecked. The question writer then produces one question of each type per passage: fact, precise value, count, coreference, cross-sentence, relation, and, for the held-out set only, unanswerable. Rules require the answer to appear verbatim in the prefix and a precise value to have at least six digits and to be neither a date, an interval, nor an address. The two values of a cross-sentence question must lie in different sentences, and the stem of a coreference question may not contain the answer's surname. Two gates follow. A closed-book pass on the question alone must fail and an open-book pass with the prefix must succeed, which together remove about 40% of the generated questions. A per-question judge removes another 5 to 8% as incorrect, ambiguous, answerable without the source, or incomplete, and three reviewers read the survivors.

The training set holds 5,969 passages with 29,001 questions and the held-out set 436 passages with 2,762 questions, of which 733 are precise values spread over four digit buckets. Training and held-out passages come from the same generator in independent runs, so scenarios, name pools, and value labels are shared, and §G.5 reports the overlap checks.

Table 10: Evaluation sets. Median passage length is measured with the Qwen3-8B tokenizer on the passage that the sender encodes. Scored questions exclude the unanswerable questions of the held-out sets.

| Set | Source | Passages | Questions (scored) | Scoring | Median tokens |
|---|---|---|---|---|---|
| Synthetic | program-planned records, held-out | 436 | 2,762 (2,404) | judge | 758 |
| Web | FineWeb-Edu passages, held-out | 1,142 | 1,958 (1,459) | judge | 820 |
| NewsQA | MRQA release | 372 | 500 (455) | judge | 677 |
| HotpotQA | MRQA release | 500 | 500 | judge | 285 |
| MRQA | MRQA out-of-domain dev, four subsets | 445 | 500 | judge | 371 |
| MedQA (Appendix J) | USMLE four-option test split | 500 | 500 | first token A–D | 149 |
| COCO | COCO validation images, held-out | 1,959 img | 5,936 | judge | – |
| RealWorldQA | xAI release | 765 img | 765 | rule (428 choice) + judge | – |
| CV-Bench | Cambrian-1 release | 2,638 img | 2,638 | rule | – |

### G.2 Web passages

Web passages come from the 10BT sample of FineWeb-Edu (Penedo et al., 2024). Pages of worked mathematics are dropped as a whole and near duplicates are removed with MinHash (Broder, 1997). Each page is cut at sentence boundaries into a prefix of about 760 tokens and a suffix, and is kept only when the suffix reuses a number or a person name of the prefix. Questions are written, rule-filtered, and judged by the pipeline of §G.1. Because a web fact may be public knowledge, a world-knowledge gate, applied to Web only in Figure 7, puts every question closed-book to qwen3.8-flash and deepseek-v4-flash with abstention forbidden. A question is removed when either answer matches the gold (for count questions, when both do) or when a strict judge twice calls the fact common knowledge. The gate removes 26% of held-out and 33% of training questions, most heavily among relation questions. The training set holds 7,955 passages and 34,867 questions (4.38 per passage, 27.6% precise values). The held-out set holds 1,142 passages with 1,958 questions: 499 unanswerable, 382 count, 357 fact, 215 relation, 136 cross-sentence, 245 precise value, and 124 coreference. It was reviewed in three rounds, and any batch of new questions with more than 10% errors was dropped as a whole.

**Copy prompts.** For both corpora a partner model scores every training prefix token by token. Contiguous spans of high surprisal (above the 65th percentile, 2 to 16 tokens, at most 8 per passage) become prompts of the form `Copy exactly the text that immediately follows "<left context>".`, where the six-token left context must be unique within the passage. This yields 47,752 synthetic and 63,633 Web prompts, used only in training.

### G.3 Public sets

NewsQA and HotpotQA use 500 questions each from the MRQA-formatted releases and are not filtered for closed-book answerability: the No context row of Table 1 reports that share directly. NewsQA drops 45 questions whose gold answer none of the four receivers reproduces when reading the passage, leaving 455. MRQA takes 125 questions from each of the four out-of-domain development sets of the 2019 shared task, DuoRC (ParaphraseRC), BioASQ, TextbookQA, and RelationExtraction, 500 questions on 445 passages sampled by the same hash rule and likewise unfiltered. Its No context row is high because BioASQ and TextbookQA are partly answerable from world knowledge, whereas DuoRC and RelationExtraction are not. One or two verdicts per cell blocked by the judge provider's content filter are excluded from the denominator. MedQA uses the four-option USMLE test split, with the clinical vignette as the passage and the question sentence with its options as the query. It is scored by the first generated token restricted to A to D. One HotpotQA passage has no byte-level correspondence for Llama-3.1-8B, so its directions use 499 questions.

### G.4 IMAGE SETS

The image pipeline of Figure 7 starts from COCO at native resolution: 19,081 training images and 2,039 validation images with no overlap. The validation pool keeps images with at least one annotated object covering 1% of the frame. For each validation image qwen3.8-flash writes descriptions at three levels of detail (48.8, 60.7, and 87.9 words on average). A judge of a different vendor, kimi-k3, marks each sentence as supported, contradicted, or unverifiable, and contradicted sentences are repaired in at most two targeted rounds. A blind paired test by a third model scores the repair as reducing hallucinated sentences from 26.0% to 6.8%. Questions and their answer keys are written by qwen3.8-flash looking at the image and never at the description, eight main and five peripheral questions per image in six types (attribute, spatial relation, scene, counting, action, text in the image). This gives 26,057 questions with free-form short answers. Three gates then run with Qwen3-VL-8B answering under three amounts of information, each verdict decided by deepseek-v4-flash against the key. A question answered correctly without the image is removed, a question answerable from the pre-written description of middle detail (60.7 words) alone is labeled covered, and a question the model answers wrongly while looking at the image is removed, leaving 13,699 questions (52.6%). A second answerer, kimi-k3, answers every question blind, and the same text judge records whether it agrees with the key. From the survivors 6,000 questions on 1,962 images are drawn in hash order and frozen. A review by qwen3.8-max of 1,269 keys (all 967 disagreements and 302 agreements) flags 130, of which an adjudicator of a different vendor, Claude Opus 5, confirms 37 wrong keys and 27 broken questions, leaving 5,936 questions on 1,959 images. Training images receive one description each and pass the same gates, giving 127,219 questions. RealWorldQA and CV-Bench are converted to the same format with the options folded into the question.

### G.5 LEAKAGE AND SANITY CHECKS

Every data release carries a manifest verifying that held-out and training passages are disjoint, that question identifiers are unique, and that the held-out files are byte-identical across releases. Held-out values of eight or more digits occur zero times in training, and 8-gram containment between held-out and training passages is 4.4% for the synthetic sets and 0.8% for the Web sets. Training and evaluation images share no identifier.

### G.6 PROMPTS

All prompts are frozen before any evaluation and reproduced here verbatim.

**T2T note.** Every sender receives, as a single user message through its own chat template without a system message, the sentence below followed by two line breaks and the passage, and writes greedily (vLLM, seed 0, bf16; thinking disabled where the template has the switch) under a 2,048-token cap. A note reaches the cap only when the writer fails to stop on its own and its output degenerates, and such a note is truncated there. The Llama-3.1 template inserts its default system header when no system message is given.

**Prompt 1. T2T note**

```
Read the passage and write notes that preserve every fact, name, number, identifier,
date, quantity and relation, so that someone who has not read the passage can answer
any question about it. Do not add commentary.
```

**Receiver query.** Receivers get no chat template. After the passage, the note, or the injected latent representation, the query below is appended and the answer is decoded greedily for at most 64 tokens and cut at the first line break.

**Prompt 2. Receiver query**

```
\n\nQuestion: {question}\nAnswer:
```

**Text judge.** One user message, JSON output, temperature 0, thinking disabled, at most 64 output tokens.

**Prompt 3. Text answer judge**

```
You are grading a short answer against a gold answer.

Question: {question}
Gold answer: {gold}
Model answer: {pred}

Is the model answer equivalent to the gold answer (same fact/value/meaning, regardless
of wording or language)? An empty, evasive, or contradicting answer is not equivalent.
Extra correct detail is fine.
Output JSON only: {"equivalent": true} or {"equivalent": false}
```

**Text corpora.** The nine messages below build the synthetic and FineWeb-Edu corpora, each sent as a single user message with a JSON reply and no system prompt, with per-call fields shown as <<...>>.

*Synthetic passage writer.* deepseek-v4-flash, one user message, JSON reply. Genre, scenario, situation, people, organisations, locations, and exact values are filled per passage, and the character targets are what the code derives from 760 and 520 target tokens.

**Prompt 4. Synthetic passage writer**

```
Write a fictional <<genre>> set in <<scenario>>. The situation is <<situation>>.
Output JSON: {"part1": "...", "part2": "..."}.

The numbers below are ordinary, correct records. Unless the situation itself is a
disagreement or a shortage, NOTHING is wrong with them: do not write about
discrepancies, mismatches, audits, investigations or anything failing to add up.
Equally, do not write that anything was checked and found correct, consistent,
verified or free of problems -- the values are simply used in the course of the
work, the way a real document uses them.

Use ONLY the people, organisations, locations and exact values below. Invent the events
that connect them. Never use the name of a real person, company or place.

People:
  - <<name>>, <<age>>, <<title>>
  - ...
Organisations:
  - <<organisation>>
  - ...
Locations:
  - <<location>>
  - ...
Exact values (reproduce each one CHARACTER FOR CHARACTER, never round or reformat):
  - <<label>>: <<value>>
  - ...

part1 - 6 paragraphs, 4 to 6 sentences each, about 2952 characters in total and
NEVER fewer than 2331 characters. Introduce every person, organisation and value
above, naturally, as part of the narrative. Do not stop early.

part2 - 4 paragraphs, 4 to 6 sentences each, about 2020 characters in total and
NEVER fewer than 1476 characters. Later developments. Hard requirements:
  (1) reuse at least 5 of the exact values above, character for character;
    (2) NONE of these full names may appear anywhere in part2: <<full name 1>>; <<
  full
  name 2>>; ....
      Refer to those people only by surname or pronoun;
  (3) do NOT re-explain who anyone is and do NOT restate part1's background; the new
      developments must depend on part1 to be understood.
```

**Prompt 4. Synthetic passage writer (continued)**

```
Style: every exact value must appear inside running prose, never in a bulleted or
numbered list. Do NOT mention every person in every paragraph -- most paragraphs
should involve only one or two of them, and the order should change between
paragraphs. Do not begin two sentences in the same paragraph with the same word.
Separate paragraphs with a blank line. Use plain ASCII punctuation only - no
non-breaking hyphens or typographic dashes, and always put a space after a full stop.
Output JSON only.
```

*Question writer.* deepseek-v4-flash. The held-out sets request all seven types, and the training sets omit the unanswerable block.

**Prompt 5. Text question writer**

```
You write QA probes for an information-transfer experiment. Produce exactly 7
questions from the passage below, EXACTLY ONE of each of these types:

  - fact: one concrete factual detail. The answer must be a short span copied
    verbatim from the passage.
  - count: how many of something. The answer must be a small integer (1-99) or an
    English number word.
  - precise_value: an exact value carrying at least 6 digits -- an amount, an
    account or tracking number, a long identifier, a full date. The answer must be
    that value copied CHARACTER FOR CHARACTER from the passage.
  - relation: how two named entities are related. The answer must be a short span
    copied verbatim from the passage.
  - coreference: the question MUST contain a third-person pronoun (he, she,
    they, him, her, them, his, their) standing in for a person the passage
    introduced earlier, and must ask which named person that is. Do not name
    that person anywhere in the question. Example shape: "Two paragraphs after
    the batch was released, someone overrode the hold; whose decision was
    recorded as his?" The answer is the name, copied verbatim.
  - cross_sentence: a question that needs two separate facts from two different
    sentences. The two facts must be AT LEAST TWO SENTENCES APART in the passage --
    if both appear in the same sentence the question is rejected. Give the answer as
    "answer_parts", a list of the two spans, each copied verbatim from the passage.
    Leave "answer" as the two parts joined by "; ".
  - unanswerable: a question about a plausible but ABSENT detail -- something a
    reader would expect this kind of document to contain but which this passage
    never states. Set "answer" to "not stated" and set "probe" to the exact thing
    being asked about (an entity or attribute name); that probe string must not
    appear in the passage.

Rules:
  - Every question must require this passage. None may be answerable from general
    world knowledge.
  - No two questions may begin with the same three words.
  - Write the questions in English.
  - Do not write open-ended, summary, opinion or topic-only questions.

Return strict JSON: {"questions": [{"question_type": "...", "question": "...",
"answer": "...", "answer_parts": [...], "probe": "..."}]}
("answer_parts" only for cross_sentence, "probe" only for unanswerable.)

Passage:
<<<
<<passage>>
>>>
```

*Passage judge.* qwen3.8-flash at temperature 0, two independent passes per synthetic passage.

**Prompt 6. Synthetic passage judge**

```
You are auditing ONE document from a synthetic corpus. The corpus exists to train a
system that must carry LONG, PRECISE values (17-digit file numbers, exact amounts,
alphanumeric codes) through a lossy channel. Long or unusual-looking values are
INTENDED. Never mark a value as a problem because it is long, rare, or has many
digits. Judge only whether a value belongs in THIS document and whether its size
fits its stated role.

DOCUMENT METADATA
  scenario: <<scenario>>
  genre: <<genre>>
  situation: <<situation>>

VALUES THE WRITER WAS TOLD TO USE (idx | role | value)
  <<idx>> | <<role>> | <<value>>
  ...

DOCUMENT, PART 1
<<<
<<part 1>>
>>>
DOCUMENT, PART 2 (later developments)
<<<
<<part 2>>
>>>

PART A -- judge each value with two questions, in order.
    fits_document: in a real <<genre>> from <<scenario>>, is there ANY realistic
  reason
  such a
      quantity or identifier would be recorded?
      FAIL examples: a power output in megawatts inside a sports-league fixture record;
      a blood-serum concentration in a law firm's case memo; "a sampling ratio of
      146:155" that no event in the text needs.
      PASS examples: a 17-digit case file number in a legal memo; EUR 4,000,377.34 as a
      contract amount; 3965.8 GB of survey data in a research log.
  fits_role: only if fits_document is true -- does the magnitude suit the role and
      unit?
      FAIL examples: "a measured reading of 55003.1 km" for a compliance sample;
      "31662.7 kN" given as the weight of cargo; USD 482 million invoiced for 13
      components.
      PASS examples: a 47.3% pass rate; 46582.1 kg on a weighbridge.
  For every false, copy the EXACT fragment of the document (verbatim, at least 6
  words) where that value is used, into "quote". A false without a verbatim quote
  is discarded.

PART B -- classify the document. Pick one label per line. Classify, do not judge.
  narrative_type: routine_operation | planning | review_or_audit |
            anomaly_investigation | disagreement | incident_response | training |
      handover | other
  person_reference_style: natural | every_paragraph_names_everyone
  prose_quality: natural | slightly_stiff | list_like

PART C -- defects. true means OK. Every false needs a verbatim quote.
  no_internal_contradiction   dates in an impossible order, a person acting before
                              being introduced, a quantity contradicting another
  suffix_continues_prefix     part 2 continues the same document and events, not a
                              fresh start with re-introductions
  language_pure               one language throughout (unit symbols and codes are fine)
  no_meta_leak                no "as an AI", no "Part 1"/"Part 2", no placeholder, no
                              echo of instructions
  values_woven_in             values appear inside running prose, not in a bulleted or
```

**Prompt 6. Synthetic passage judge (continued)**

```
                              numbered list

BE CONSERVATIVE: if unsure, answer true. Do not add requirements beyond these.

Return strict JSON and nothing else:
{"values": [{"idx": 0, "fits_document": true, "fits_role": true, "why": "", "quote":
""}, ...],
 "narrative_type": "...", "person_reference_style": "...", "prose_quality": "...",
 "defects": {"no_internal_contradiction": {"ok": true, "why": "", "quote": ""},
             "suffix_continues_prefix": {"ok": true, "why": "", "quote": ""},
             "language_pure": {"ok": true, "why": "", "quote": ""},
             "no_meta_leak": {"ok": true, "why": "", "quote": ""},
             "values_woven_in": {"ok": true, "why": "", "quote": ""}}}
Judge every value; keep the idx order.
```

*Per-question judge.* qwen3.8-flash, one message per passage carrying all its questions. A question is kept only when every flag is true.

**Prompt 7. Per-question judge**

```
You are checking questions written from the passage below. The passage is all
anyone answering them will ever see.

PASSAGE
<<<
<<passage>>
>>>

QUESTION TYPES
  fact           a concrete factual detail, answer copied from the passage
  count          how many of something; answer is a small integer or number word
  precise_value  an exact value of >=6 digits (amount, account/tracking number,
                 long identifier, full date), copied character for character
  relation       how two named entities stand to each other.  A role or title IS an
                 acceptable answer here ("the site engineer for X").
  coreference    the question uses a third-person pronoun for a person introduced
                 earlier and asks who that is; the person must NOT be named in the
                 question
  cross_sentence needs two separate facts from two different sentences; the answer
                 has two parts
  unanswerable   asks about a plausible detail the passage never states; the correct
                 answer is "not stated"

QUESTIONS
- qid=<<qid>>  type=<<question_type>>
  Q: <<question>>
  A: <<answer>>  [parts=[...] for cross_sentence; probe=... for unanswerable]
- ...

ALREADY MACHINE-CHECKED before you see them -- you do not need to re-verify, but do
report anything the machine check could only approximate:
  * the answer appears verbatim in the passage (except cross_sentence/unanswerable)
  * precise_value answers carry at least 6 digits; count answers are small integers
  * coreference questions contain a third-person pronoun and do not contain the
    answer string or the person's surname -- BUT the machine cannot see a person
    referred to by title or description, so judge that yourself
  * cross_sentence answers have two parts, both verbatim, and the two do not sit
    together in a single sentence

BE CONSERVATIVE. If you are genuinely unsure about a flag, mark it true. Only mark a
```

**Prompt 7. Per-question judge (continued)**

```
flag false when you can point at the specific problem. Do not invent additional
requirements beyond the definitions above.

For EACH question, judge these things independently. Answer strictly true/false.
  answer_correct  the given answer is what the passage actually says. False if it is
                  subtly wrong, paraphrased into something different, or if the question
                  presupposes something the passage does not state.
                  If false: copy into "quote" the EXACT passage fragment (verbatim, at
                  least 6 words) that shows the real answer. A false without a verbatim
                  quote is discarded.
  answer_unique   exactly one answer is defensible. False ONLY when the passage holds
                  two or more items that fit the question's own wording equally well and
                  the question does not say which is meant (for example "what is the
                  reference number" when three different reference numbers appear).
                  The mere presence of other numbers or names in the passage is NOT a
                  reason to mark this false.
                  If false: copy into "quote" the verbatim fragment holding the OTHER
                  item that fits equally well. No quote, verdict discarded.
    needs_passage   the question cannot be answered from general world knowledge, and
  cannot
                  be answered from its own wording. False if it is self-answering.
  well_formed     grammatical, in one language, not awkward or truncated.
    coref_resolvable  ONLY for type=coreference (answer true for every other type):
  reading
                  the passage, does the pronoun in the question point to exactly one
                  person, including when that person is referred to by a title or
                  description rather than a name? False if it is ambiguous or the
                  pronoun has no antecedent in the passage.

Do NOT judge whether the question matches its declared type, and do NOT judge whether
the passage's facts are realistic -- both are checked elsewhere.

Return strict JSON and nothing else:
{"verdicts": [{"qid": "...", "answer_correct": true, "answer_unique": true,
"needs_passage": true, "well_formed": true, "coref_resolvable": true,
"reason": "", "quote": ""}]}
Fill "reason" with a few words ONLY when at least one flag is false. Keep the order of
the
questions. Judge every question.
```

*FineWeb-Edu segment judge.* qwen3.8-flash, applied to every FineWeb-Edu segment before questions are written.

**Prompt 8. FineWeb-Edu segment judge**

```
You are checking ONE segment of web text that will be used as training material.
The segment was cut from a longer page into PART 1 and PART 2. Judge only the
hygiene questions below. Do not judge whether the content is true, interesting,
or well written.

PART 1
<<<
<<part 1>>
>>>
PART 2
<<<
<<part 2>>
>>>

For each check answer true (OK) or false (defect). Every false MUST carry a verbatim
fragment of the segment (at least 6 words) in "quote"; a false without a verbatim
```

**Prompt 8. FineWeb-Edu segment judge (continued)**

```
quote is discarded.
  single_document          the whole segment is one continuous document or article, not
                           two or more unrelated pieces joined together
  no_boilerplate           no navigation menus, cookie notices, "share this", "related
                           posts", copyright footers, ad text, login prompts, or
                           repeated template lines inside the body
  suffix_continues_prefix  PART 2 continues the same text as PART 1 (same topic, same
                           document), it is not a different article
    clean_start              PART 1 begins at a sensible point (a sentence or heading),
  not
                           mid-sentence or in the middle of a list item
  clean_end                PART 2 ends at a sentence boundary, not cut mid-sentence
  language_pure            one language throughout (isolated names, titles, code or
                           formulas are fine)
  prose_layout             mostly running prose or a coherent article; not a table dump,
                           a list of links, a form, or a wall of numbers
  no_garbage               no mojibake, control characters, HTML tags, or repeated
                           junk strings

Also classify (pick one, do not judge):
    doc_type: article | news | encyclopedic | forum_or_comment | academic |
  product_or_ad |
            listing_or_table | other

BE CONSERVATIVE: if unsure, answer true. Do not add requirements beyond these.
Return strict JSON and nothing else:
{"checks": {"single_document": {"ok": true, "why": "", "quote": ""}, ... all eight ...},
 "doc_type": "..."}
```

*World-knowledge gate, closed-book answering.* run with both qwen3.8-flash and deepseek-v4-flash at temperature 0, questions batched in one message.

**Prompt 9. World-knowledge gate: closed-book answering**

```
You have NO passage. Answer each question from your own knowledge. You MUST give your
single best guess for every question; never say unknown, not sure, cannot determine, not
provided, or leave anything blank -- if you are unsure, still commit to the most likely
specific answer. Ignore any wording like 'according to the passage' or 'given in the
passage' and just answer the underlying question. Answer each question with a short span
(a name, number, date or short phrase; for questions asking for several things give all
parts separated by ';'). Return JSON {"answers": ["...", "..."]} in question order,
exactly one string per question.

1. <<question 1>>
2. <<question 2>>
...
```

*World-knowledge gate, equivalence check.* qwen3.8-flash, only for pairs not already decided by the exact or numeric rule.

**Prompt 10. World-knowledge gate: equivalence check**

```
You compare a candidate answer with the reference answer to a question. Decide for each
item whether the candidate states the SAME fact as the reference. Count as same: a
paraphrase, a different formatting, unit or rounding of the same value, a name written
differently (with or without title, initials, first name), or a longer answer that
clearly contains the reference fact. Count as different: a different entity, value or
date, an answer that misses one of the parts when the reference has several parts, or a
vague or generic answer that does not commit to the reference fact. When the reference
names a specific entity, organisation, value or mechanism, the candidate must name that
```

**Prompt 10. World-knowledge gate: equivalence check (continued)**

```
same specific thing; a generic category word that merely overlaps with the reference
(for example 'a representative' for 'Forest Service Representative', or 'they are
related' for a specific relationship) is different. Return JSON {"same": [true, false,
...]} with exactly one boolean per item, in order.

1. Question: <<question>>
   Reference: <<gold answer>>
   Candidate: <<closed-book answer>>
2. ...
```

*World-knowledge gate, common-knowledge check (main verdict).* qwen3.8-flash, two shuffled passes.

**Prompt 11. World-knowledge gate: common-knowledge main verdict**

```
Each item below is a question and its reference answer, both written from a web passage
that you cannot see. Decide whether the reference answer is PUBLIC / COMMON KNOWLEDGE: a
famous person, event, place or date; a definition; a textbook fact; a standard
scientific, technical, legal or institutional fact; or any well-documented fact that a
well-read person or a strong language model could state correctly WITHOUT reading the
passage. Judge the specific answer, not just the topic: if the topic is famous but this
particular value, name or figure is only found in that passage (a study's own numbers, a
local organisation's details, a narrative detail), answer no. Return JSON {"items":
[{"id": 1, "common": true, "why": "..."}, ...]} with one entry per item.

1. Question: <<question>>
   Reference answer: <<gold answer>>
2. ...
```

*World-knowledge gate, common-knowledge check (strict verdict).* qwen3.8-flash, two shuffled passes.

**Prompt 12. World-knowledge gate: common-knowledge strict verdict**

```
Each item below is a question and its reference answer, both written from a web passage
that you cannot see. Decide whether a strong language model would very likely produce
this EXACT reference answer from memory alone, without the passage. Answer yes only for
genuinely encyclopedic facts about well-known subjects: major historical events and
their dates, famous people's biographical facts, standard scientific constants,
definitions and textbook facts, well-known institutions and their roles, widely
published statistics that are stable across sources. Answer no for: figures specific to
one study, report, survey or article; local or niche details (a specific organisation's
phone number, an address, coordinates, a small place's history, a minor person's
biography); values that differ between sources or change over time; counts or
measurements that belong to this passage's own narrative; anything about obscure
entities. Return JSON {"items": [{"id": 1, "common": true, "why": "..."}, ...]} with one
entry per item.

1. Question: <<question>>
   Reference answer: <<gold answer>>
2. ...
```

**Image descriptions.** The three tiers share one instruction and differ in the requested length and coverage. The simple tier:

**Prompt 13. Image descriptions: simple tier**

```
Look at this image and reply with JSON only:
{"caption": "..."}

caption: two or three sentences for someone who cannot see the image. State only
what is visibly present - the main subjects, how many, colours, what they are
doing, and where things sit relative to each other. Do not mention art style,
photo quality, resolution, mood, or anything you cannot literally see, and do
not speculate about meaning or intent.
```

The middle tier asks for two or three sentences and the detailed tier for four or five, and both add the following after the main subjects:

**Prompt 14. Image descriptions: middle and detailed tier addition**

```
then also the background, the edges of the frame, small objects,
what people are wearing, materials and surfaces, and any words you can
actually read on signs, labels or clothing. If you cannot read or make out
something clearly, leave it out rather than guessing at it.
```

**Image questions.** Main questions, eight per image:

**Prompt 15. Image questions: main questions**

```
Look at this image and reply with JSON only:
{"qa": [{"q": "...", "a": "...", "kind": "..."}]}

Ask exactly 8 question-answer pairs testing whether someone has really seen this
image. Ask about the main subjects, how many of them there are, their colours,
what they are doing, and where things sit relative to each other.

Rules:
- the answer must be a short phrase (a word, a number, a colour, a name) that can
  be read off the image;
- write every number as Arabic digits ("2", never "two");
- do not ask anything answerable by common sense without seeing the image;
- do not ask about anything you are unsure of;
- use the kind "count" only when the answer is a number; if the answer is not a
  number, label the question with one of the other kinds instead;
- spread them over different kinds, and label each with "kind" from:
  attribute (colour, material, size), count, spatial (where things are relative
  to each other), action (what someone or something is doing), text (words
  visible in the image), scene (what kind of place or setting).
```

Peripheral questions, five per image, use the same rules with this opening:

**Prompt 16. Image questions: peripheral opening**

```
Ask exactly 5 questions about details that a short one-sentence description of
this image would NOT mention - the background, the edges of the frame, small
objects, what someone is wearing, materials and surfaces, writing on signs or
labels or clothing, what a secondary figure is doing. Do not ask about the main
subject or the most obvious thing in the picture.
```

**Description sentence judge.**

**Prompt 17. Description sentence judge**

```
You are checking a written description of this image, sentence by sentence.

The sentences of the description are numbered below. For EACH sentence, decide:

- "supported": everything the sentence asserts can be found in the image.
- "contradicted": at least one thing the sentence asserts differs from what the
  image actually shows - wrong object, wrong colour, wrong count, wrong text or
  number, wrong action, wrong position, or an object that simply is not there.
  This is the only label meaning the description is WRONG.
- "unverifiable": you cannot tell either way - the thing is too small, too
  blurred, cut off, occluded, or outside the frame. Neither confirmable nor
  refutable.

Rules:
- Judge only against what is visible. Do not use world knowledge to fill gaps,
  and do not assume a claim is true because it sounds plausible.
- THRESHOLD. The question is whether a person looking at the image would say the
  description is wrong, not whether they would have chosen the same words. A
  looser but fair word for the same thing is "supported": a recliner called an
  armchair, a maroon jacket called red, a trailer called a truck, a countertop
  called a counter. Flag "contradicted" only when the claim points at something
  the image does not show: a different object, a different colour family, a
  different count, different text or digits, a different action or position, or
  an object that is absent.
- If a sentence makes several claims: label it "contradicted" if ANY claim is
  contradicted; otherwise "unverifiable" if any claim is unverifiable;
  otherwise "supported".
- For every sentence that is not "supported", name the exact component at fault
  in "component" - the specific object, count, colour, word/number, or action,
  quoted from the sentence - and state in "actual" what the image really shows
  (one short clause; for "unverifiable" say why it cannot be checked). If more
  than one component in the sentence is at fault, name them all.
- Read any text or digits in the image carefully before judging a claim about
  words or numbers; these are the claims most often wrong. Be sure which part of
  the object you are looking at before calling a colour wrong (a white candle
  can stand in a black holder).
- Do not be lenient about invented objects. An object that is not in the image
  is "contradicted", not "unverifiable".

Reply with JSON only, no prose, no code fence:
{"sentences": [{"i": 1, "verdict": "supported", "component": "", "actual": ""},
               {"i": 2, "verdict": "contradicted", "component": "...", "actual":
  "..."}]}
Return exactly %d entries, one per numbered sentence, in the same order.

Sentences:
%s
```

**Image answer judge.** Used for the agreement flag, the three gates, and every image-set condition, with Answer A the key and Answer B the answer under test. System message, then user message:

**Prompt 18. Image answer judge: system message**

```
You compare two short answers to the same question about an image and decide
whether they mean the same thing.
```

**Prompt 19. Image answer judge: user message**

```
Question: {question}
Answer A: {a}
Answer B: {b}

Answers are equivalent if they state the same fact. Allow synonyms, paraphrase,
different wording granularity, spelled-out vs digit numbers, and extra
non-contradicting detail in one answer. They are NOT equivalent if they state
different facts, different counts, or contradict each other.

Reply with strict JSON only: {"equivalent": true} or {"equivalent": false}
```

**Gate instructions to Qwen3-VL-8B.** Closed book, no image:

**Prompt 20. Image gate: closed book**

```
Answer this question about a photograph. You cannot see the photograph, so give
your single best guess based on what such a photograph most likely shows.

Question: {question}

Reply with a short answer only - a single word, a number, or a few words. No full
sentence, no explanation. Always commit to a guess; never reply that you do not
know or cannot tell.
```

Description only, no image:

**Prompt 21. Image gate: description only**

```
Here is a written description of a photograph:

{sentence}

Answer this question about that photograph using only the description above. You
cannot see the photograph itself.

Question: {question}

Reply with a short answer only - a single word, a number, or a few words. No full
sentence, no explanation. If the description does not say, reply exactly: unknown.
```

Open book, with the image:

**Prompt 22. Image gate: open book**

```
Look at the image and answer the question about it.

Question: {question}

Reply with a short answer only - a single word, a number, or a few words. No full
sentence, no explanation, no restating the question.
```

**Written description for image T2T.** The sender writes with greedy decoding under a cap of N words. T2T uses N = 360, a cap no description reaches, so the description ends at its natural length, and N = 90 and 180 serve only the cost comparison of Appendix H.

**Prompt 23. Image T2T description**

```
Describe this image. Answer in plain prose without lists or markdown, in at most
N words.
```

## H COST MEASUREMENT

Every latency and FLOPs figure in the paper comes from one measurement protocol. For each set we take one passage of median length and one of its questions, and run the routes on a single A100 in bfloat16 with HF Transformers at batch size one. Five phases are timed separately under CUDA synchronization. The sender prefills the passage, then decodes the note (nothing for LLM-SemanticLAN), then the codec runs (LLM-SemanticLAN only). The receiver prefills what it is given, that is the question after a latent representation or the note and the question for T2T, and finally decodes 24 answer tokens. Latency to the first answer token is the sum of the first four phases, and total latency adds the fifth, each reported as the median over repeated runs. FLOPs are counted per operator with `FlopCounterMode` during the same runs, with a custom formula for grouped-query attention, so attention terms and the codec's own compute are included rather than estimated from parameter counts. The counter does not see the custom kernels of the recurrent layers of Qwen3.5-9B, so FLOPs on its side are undercounted. Table 11 gives every phase for the six directions among the initial models on the Web passage. Table 12 gives the ratios behind Table 1, T2T over LLM-SemanticLAN per direction and set, averaged over the five sets and reported as the range over the directions of a row.

Table 11: Phase-level cost of the six directions among the initial models on the median Web passage (820 tokens), one question, 24 answer tokens: latency in ms and FLOPs in TFLOPs on separate rows for each route. The T2T note is written at natural length.

| Direction | Route, unit | Sender prefill | Sender decoding | Codec | Receiver prefill | Receiver decoding | Total |
|---|---|---|---|---|---|---|---|
| Qwen3-4B → Qwen3-8B | LLM-SemanticLAN, ms | 55 | – | 19 | 39 | 783 | 896 |
| | LLM-SemanticLAN, TFLOPs | 6.99 | – | 0.24 | 0.22 | 0.36 | 7.82 |
| (note of 892 tokens) | T2T, ms | 54 | 31,332 | – | 89 | 832 | 32,306 |
| | T2T, TFLOPs | 6.81 | 7.86 | – | 13.07 | 0.36 | 28.10 |
| Qwen3-4B → Qwen3-14B | LLM-SemanticLAN, ms | 55 | – | 20 | 43 | 870 | 988 |
| | LLM-SemanticLAN, TFLOPs | 6.99 | – | 0.26 | 0.41 | 0.66 | 8.32 |
| (note of 892 tokens) | T2T, ms | 54 | 30,626 | – | 148 | 890 | 31,718 |
| | T2T, TFLOPs | 6.81 | 7.86 | – | 24.61 | 0.66 | 39.95 |
| Qwen3-8B → Qwen3-4B | LLM-SemanticLAN, ms | 89 | – | 19 | 40 | 799 | 947 |
| | LLM-SemanticLAN, TFLOPs | 12.81 | – | 0.24 | 0.12 | 0.20 | 13.37 |
| (note of 832 tokens) | T2T, ms | 88 | 28,617 | – | 53 | 756 | 29,513 |
| | T2T, TFLOPs | 12.67 | 13.21 | – | 6.57 | 0.20 | 32.65 |
| Qwen3-8B → Qwen3-14B | LLM-SemanticLAN, ms | 89 | – | 20 | 44 | 873 | 1,026 |
| | LLM-SemanticLAN, TFLOPs | 12.81 | – | 0.26 | 0.41 | 0.66 | 14.13 |
| (note of 832 tokens) | T2T, ms | 88 | 28,620 | – | 134 | 866 | 29,708 |
| | T2T, TFLOPs | 12.67 | 13.21 | – | 22.94 | 0.66 | 49.48 |
| Qwen3-14B → Qwen3-4B | LLM-SemanticLAN, ms | 136 | – | 20 | 38 | 772 | 966 |
| | LLM-SemanticLAN, TFLOPs | 23.49 | – | 0.24 | 0.12 | 0.20 | 24.05 |
| (note of 806 tokens) | T2T, ms | 136 | 31,080 | – | 52 | 880 | 32,148 |
| | T2T, TFLOPs | 23.86 | 23.37 | – | 6.36 | 0.20 | 53.78 |
| Qwen3-14B → Qwen3-8B | LLM-SemanticLAN, ms | 136 | – | 20 | 38 | 773 | 966 |
| | LLM-SemanticLAN, TFLOPs | 23.49 | – | 0.24 | 0.22 | 0.36 | 24.32 |
| (note of 806 tokens) | T2T, ms | 135 | 30,552 | – | 85 | 772 | 31,544 |
| | T2T, TFLOPs | 23.86 | 23.37 | – | 11.79 | 0.36 | 59.38 |

The visual sender is measured the same way on 100 images per set, three questions per image on the held-out set and one on the benchmarks. Table 13 lists the phases for the held-out set. Both routes pay for the vision-language model looking at the image, T2T then decodes a description whose length sets its cost, and LLM-SemanticLAN writes once and lets the reader prefill only the question.

## I VISION: HELD-OUT AND EXTERNAL BENCHMARK RESULTS

Table 14 splits the held-out COCO set for every receiver by whether the answer can be recovered from the pre-written description (covered) or not (uncovered). Reading the pre-written description

Table 12: T2T over LLM-SemanticLAN, total latency and FLOPs, for every measured direction and set of Table 1, with the per-direction mean over the five sets. The range of these means over the directions of a row is the cost cell of Table 1.

| Direction | Latency ratio | | | | | | FLOPs ratio | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Web | Synthetic | NewsQA | HotpotQA | MRQA | Mean | Web | Synthetic | NewsQA | HotpotQA | MRQA | Mean |
| Qwen3-4B → Qwen3-8B | 36.1 | 32.1 | 28.6 | 13.1 | 17.4 | 25.5 | 3.6 | 3.4 | 3.5 | 3.0 | 3.5 | 3.40 |
| Qwen3-4B → Qwen3-14B | 32.1 | 29.3 | 27.0 | 11.9 | 15.4 | 23.1 | 4.8 | 4.3 | 4.7 | 3.5 | 4.4 | 4.32 |
| Qwen3-4B → Llama-3.1-8B | 47.9 | 44.7 | 39.6 | 17.6 | 24.7 | 34.9 | 3.6 | 3.3 | 3.5 | 3.0 | 3.5 | 3.38 |
| Qwen3-4B → II-Medical-8B | 32.1 | 27.7 | 29.0 | 12.6 | 17.1 | 23.7 | 3.9 | 3.6 | 3.8 | 3.2 | 3.8 | 3.68 |
| Qwen3-4B → Qwen3.5-9B | 29.0 | 26.3 | 23.9 | 11.2 | 13.1 | 20.7 | 3.9 | 3.6 | 3.8 | 3.2 | 3.8 | 3.67 |
| Qwen3-8B → Qwen3-4B | 31.2 | 31.7 | 22.2 | 13.9 | 17.8 | 23.3 | 2.4 | 2.4 | 2.1 | 2.6 | 2.7 | 2.46 |
| Qwen3-8B → Qwen3-14B | 28.9 | 27.7 | 20.0 | 12.6 | 16.2 | 21.1 | 3.5 | 3.4 | 3.0 | 3.2 | 3.7 | 3.35 |
| Qwen3-8B → Llama-3.1-8B | 43.5 | 41.3 | 29.6 | 18.4 | 25.1 | 31.6 | 2.8 | 2.7 | 2.4 | 2.8 | 3.1 | 2.77 |
| Qwen3-8B → II-Medical-8B | 29.6 | 28.7 | 21.0 | 13.9 | 9.1 | 20.5 | 3.0 | 3.0 | 2.6 | 3.0 | 3.3 | 3.00 |
| Qwen3-8B → Qwen3.5-9B | 26.5 | 25.4 | 18.4 | 11.8 | 14.8 | 19.4 | 3.0 | 3.0 | 2.6 | 3.0 | 3.3 | 2.99 |
| Qwen3-14B → Qwen3-4B | 33.3 | 30.0 | 23.0 | 14.4 | 15.9 | 23.3 | 2.2 | 2.2 | 2.0 | 2.4 | 2.2 | 2.21 |
| Qwen3-14B → Qwen3-8B | 32.7 | 29.7 | 22.3 | 14.5 | 15.6 | 23.0 | 2.4 | 2.4 | 2.2 | 2.5 | 2.4 | 2.39 |
| Qwen3-14B → Llama-3.1-8B | 43.1 | 40.1 | 32.7 | 18.7 | 21.7 | 31.3 | 2.4 | 2.4 | 2.2 | 2.5 | 2.4 | 2.38 |
| Qwen3-14B → II-Medical-8B | 31.2 | 28.0 | 23.8 | 13.9 | 10.9 | 21.6 | 2.6 | 2.5 | 2.3 | 2.7 | 2.5 | 2.52 |
| Qwen3-14B → Qwen3.5-9B | 27.0 | 24.9 | 17.8 | 12.6 | 13.3 | 19.1 | 2.6 | 2.5 | 2.3 | 2.7 | 2.5 | 2.52 |
| Llama-3.1-8B → Qwen3-4B | 12.0 | 19.9 | 11.2 | 8.6 | 12.7 | 12.9 | 1.8 | 2.3 | 2.0 | 2.5 | 3.0 | 2.33 |
| Llama-3.1-8B → Qwen3-8B | 12.1 | 16.0 | 13.1 | 8.6 | 12.8 | 12.5 | 2.0 | 2.7 | 2.3 | 2.7 | 3.4 | 2.62 |
| Llama-3.1-8B → Qwen3-14B | 11.0 | 15.7 | 12.4 | 7.5 | 11.9 | 11.7 | 2.3 | 3.2 | 2.8 | 3.1 | 4.1 | 3.11 |
| II-Medical-8B → Qwen3-4B | 21.3 | 26.5 | 17.2 | 13.5 | 39.3 | 23.6 | 2.2 | 2.5 | 2.1 | 2.8 | 5.5 | 3.01 |
| II-Medical-8B → Qwen3-8B | 21.9 | 28.0 | 17.1 | 13.0 | 37.4 | 23.5 | 2.5 | 2.8 | 2.4 | 3.0 | 6.4 | 3.42 |
| II-Medical-8B → Qwen3-14B | 20.4 | 24.7 | 14.7 | 12.6 | 18.5 | 18.2 | 3.0 | 3.3 | 2.8 | 3.4 | 7.9 | 4.11 |
| II-Medical-8B → Qwen3.5-9B | 18.7 | 22.4 | 15.1 | 11.5 | 32.9 | 20.1 | 2.5 | 2.8 | 2.4 | 3.0 | 6.3 | 3.41 |
| Qwen3.5-9B → Qwen3-4B | 38.0 | 41.8 | 34.7 | 14.7 | 17.5 | 29.3 | 2.7 | 3.1 | 2.8 | 2.7 | 2.8 | 2.82 |
| Qwen3.5-9B → Qwen3-8B | 37.5 | 42.6 | 32.4 | 14.8 | 18.6 | 29.2 | 3.1 | 3.5 | 3.2 | 2.9 | 3.1 | 3.18 |
| Qwen3.5-9B → Qwen3-14B | 33.5 | 38.8 | 30.3 | 13.7 | 17.3 | 26.7 | 3.8 | 4.3 | 3.9 | 3.2 | 3.7 | 3.79 |
| Qwen3.5-9B → II-Medical-8B | 36.6 | 42.8 | 33.2 | 13.7 | 18.7 | 29.0 | 3.1 | 3.5 | 3.2 | 2.9 | 3.1 | 3.18 |

Table 13: Phase-level cost of the visual sender on the held-out COCO set, Qwen3-14B reader, median over 100 images: latency (ms) and FLOPs (TFLOPs) per phase for T2T at three description lengths and for LLM-SemanticLAN with 256 latent positions. The receiver decodes 24 tokens in every route.

| Phase | T2T 90 words | T2T 180 words | T2T 360 words | LLM-SemanticLAN |
|---|---|---|---|---|
| Sender looks at the image (ms / TFLOPs) | 82 / 5.27 | 82 / 5.27 | 82 / 5.27 | 82 / 5.27 |
| Sender decodes the description (ms; tokens) | 3,584; 85 | 6,224; 151 | 12,317; 302 | – |
| Visual encoder and decoder (ms / TFLOPs) | – | – | – | 20 / 0.17 |
| Receiver prefill (tokens) | 101 | 166 | 318 | 16 |
| Receiver to first answer token (ms) | 94 | 99 | 111 | 106 |
| End to end to first answer token (ms) | 3,677 | 6,323 | 12,428 | 198 |
| FLOPs per question, 24 answer tokens (ratio) | 1.5× | 2.0× | 3.0× | 1 |
| Latency ratio to first answer token | 18× | 32× | 62× | 1 |

itself is the ceiling of a text route on covered questions and lies near the floor on uncovered questions by construction. Qwen3-8B and Qwen3-4B never took part in training the visual encoder. Table 15 reports the external image benchmarks by task for the Qwen3-14B receiver.

## J SPECIALIST HAND-OFF

MedQA (Jin et al., 2021) is used only in this appendix, with dataset details in Appendix G.3.

**Hand-off.** Specialist reasoning is a case in which the sender decodes before hand-off. This models a general-purpose agent consulting a specialist and continuing from its proposed solution. Here the sender sees the question and generates a solution, including its predicted answer, without access to the gold label. The same question-and-solution transcript is transferred either as text or, after a sender-side prefill, as a latent representation. II-Medical-8B joins the latent space against Qwen3-8B with the same codec architecture. Under a shared final-answer instruction and decoding settings, the standalone Medical and Qwen3-8B baselines reach 78.8% and 69.6%, respectively.

Table 14: Held-out COCO set by receiver and by whether the answer can be recovered from the pre-written description, judge accuracy (%). T2T uses the 360-word description written by Qwen3-VL-8B. †: receiver never trained with the visual encoder.

| Reader | Condition | Covered (3,626) | Uncovered (2,310) | All (5,936) |
|---|---|---|---|---|
| Qwen3-14B | Reads human description | 89.4 | 17.6 | 61.5 |
| | T2T | 79.3 | 46.5 | 66.5 |
| | LLM-SemanticLAN | 89.7 | 71.0 | 82.4 |
| | No context | 16.7 | 13.2 | 15.3 |
| Qwen3-8B† | Reads human description | 80.0 | 19.5 | 56.4 |
| | T2T | 74.5 | 43.4 | 62.4 |
| | LLM-SemanticLAN | 66.0 | 46.5 | 58.4 |
| | No context | 13.6 | 15.6 | 14.4 |
| Qwen3-4B† | Reads human description | 75.2 | 13.6 | 51.2 |
| | T2T | 69.1 | 37.1 | 56.6 |
| | LLM-SemanticLAN | 71.9 | 50.2 | 63.4 |
| | No context | 19.1 | 14.8 | 17.4 |

Table 15: External image benchmarks by task, Qwen3-14B receiver, absolute accuracy (%). Multiple-choice answers are forced to an option letter under every condition. Depth and distance (*) never occur among the training questions of the visual encoder.

| Task | $n$ | No context | T2T | LLM-SemanticLAN | Qwen3-VL-8B answers from image | Mismatched image |
|---|---|---|---|---|---|---|
| RealWorldQA direct | 337 | 35.3 | 52.2 | 58.2 | 66.2 | 39.2 |
| RealWorldQA choice | 428 | 32.0 | 58.4 | 65.4 | 72.4 | 40.2 |
| CV-Bench count | 788 | 30.6 | 57.5 | 69.2 | 71.4 | 37.2 |
| CV-Bench relation | 650 | 51.2 | 57.7 | 77.8 | 91.4 | 55.1 |
| CV-Bench depth* | 600 | 61.3 | 75.2 | 80.0 | 95.3 | 63.3 |
| CV-Bench distance* | 600 | 56.3 | 55.5 | 66.0 | 90.0 | 59.8 |

Table 16: MedQA transferred without the question, first token restricted to the option letters. LLM-SemanticLAN accuracy relative to T2T (%) with the rows of Table 1, then reference accuracy.

| Property | Models | Directions | MedQA |
|---|---|---|---|
| Scale | Qwen3-4B/8B/14B | 6 trained | 74.4 |
| Domain | II-Medical-8B | 2 trained | 82.5 |
| | | 4 untrained | 81.3 |
| Family | Llama-3.1-8B | 2 trained | 75.4 |
| | | 4 untrained | 73.8 |
| Attention | Qwen3.5-9B | 2 trained | 68.7 |
| | | 4 untrained | 71.2 |
| Among the added models | | 6 untrained | 74.4 |
| No context | | – | 40.9 |
| T2T | natural-length note | 6 | 61.2 |

With the codec of II-Medical-8B at step 12,000, Qwen3-8B answers from the translated state at 81.8%, 12.2 percentage points above answering alone, without receiving the specialist's weights or its generated solution as text (Table 17). Reading the same transcript as text reaches 86.2%, so LLM-SemanticLAN retains 94.9% of T2T accuracy.

**Without the question.** When the vignette is transferred without the question, MedQA keeps the structure of the other sets but not their level. Table 16 gives MedQA under the protocol of Table 1, with the same row structure. Under that separate option-choice protocol, LLM-SemanticLAN accuracy is 60.9–84.0% of T2T accuracy across the 30 directions, while T2T averages 61.2% over the

Table 17: Specialist hand-off on 500 MedQA items, judge accuracy (%).

| Condition | Acc. (%) |
|---|---|
| Qwen3-8B native | 69.6 |
| II-Medical-8B native | 78.8 |
| Medical → LLM-SemanticLAN → Qwen3-8B | 81.8 |
| Medical → T2T → Qwen3-8B | 86.2 |

six founding-model directions. The note itself averages 61.2 and the No context floor 40.9, because a short vignette has to be reasoned over rather than looked up.