When Demonstrations Fail: Representation-Resource Matching in In-Context Adaptation

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning lets a language model acquire a new task from a handful of demonstrations, yet some mapping tasks are learned and others of near-identical surface form are not. We ask a causal question that existing work, which describes the representation in-context learning *forms*, does not: *which representation source makes a demonstration usable?* Grounded in low-resource multilingual translation, we find that correct source–target correspondence is a graded adaptation signal, and that its benefit is gated by representation accessibility: injectively corrupting the source—holding content, demonstrations, and decoding fixed—collapses it, and this gating is selective: a surface bijection survives the corruption that collapses a semantic mapping. In controlled tasks, a computed-intermediate constraint separates via direct-intermediate probes and an explicit-rule control; a held-out-query control shows the numeric task's apparent rule induction is largely inflated by support overlap, while a relational task, whose query pair is held out by construction, reveals a genuine computation failure, and supplying the correct intermediate rescues both. Finally, a cross-model check shows the selective dissociation and the intermediate rescue are positive across five model families (two weaker; Appendix), though their magnitudes do not rank with the tested scale range. Together, these intervention-based results identify a representation-resource constraint on in-context adaptation—an intervention-based behavioral taxonomy of whether the source a task demands is available during adaptation—that scalar accuracy does not capture.

1 INTRODUCTION

The broad puzzle. Why can a language model learn some tasks from a handful of demonstrations, and fail on others that look just as simple? In-context learning (ICL), the model's ability to acquire a new input–output mapping from a few examples in the prompt with no weight update, is the default way we adapt these models to a task, yet its success is strikingly uneven in a way end-task difficulty does not explain. A model that learns an arbitrary token bijection from a few demonstrations will fail an abstract rule of apparently equal surface complexity; the same model, the same number of examples, the same prompt structure, yet opposite outcomes. Something beyond "how hard is the task" is deciding whether a demonstration teaches. That is the question of this paper.

Limitations of existing perspectives. Current work on *why* ICL works has largely asked a different set of questions. It asks **what representation in-context learning forms**: how demonstrations condense into a reusable task vector. It asks **where task information is stored**: which layers or heads recognize the task versus learn the mapping. It asks **how demonstrations encode task structure**: what geometry emerges in the residual stream. These are questions about the *representation the model produces*. We ask a complementary, causal question about the *representation the task requires*:

> *Which representation source does a task require, and is that source sufficiently available during adaptation?*

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.

We study not what representation in-context learning forms, but whether the representation a task demands is available as a usable resource during adaptation. The distinction is not semantic. A task whose output is a function of the input's surface identity can be solved by *accessing* a representation the model already has. A task whose output is a function of a computed intermediate value demands an intermediate representation that may not be effectively available during adaptation; whether it becomes available is a separate, empirically testable constraint, not a given.

Representation-resource matching. This reframes adaptation as a matching problem between what a task *demands* and what a model *has available*. We define a task's **representation demand** as the source its output is a function of: surface identity, pretrained semantics, or a computed intermediate. When that source is accessible, demonstrations suffice; when the required intermediate is not sufficiently available during adaptation, they fail under the tested $K{=}8$ setting. We call this view *representation-resource matching*, established through interventions as a complementary lens, not a completed theory; the distinction is behavioral throughout.

Two discoveries and a cross-model check. We reach this view through interventions, not observation. **First, representation access gates correspondence:** correct correspondence is a graded signal, but its benefit collapses when the source is corrupted, holding content and demonstrations fixed, and on Qwen3-8B this gating is selective (a surface bijection survives the corruption under which semantic-mapping performance collapses). **Second, the intermediate-representation constraint separates into computation and composition:** with direct-intermediate probes, a held-out-query control, and intermediate rescue we show that a computed intermediate is not learned in-context, that the numeric task's apparent rule induction is largely inflated by support overlap while a relational task reveals a genuine computation failure, and that supplying the correct intermediate rescues both. A descriptive cross-model check shows the dissociation and rescue are positive across all five families (two weaker; Appendix).

The field implication. ICL evaluation asks whether a model succeeds; we ask whether it can access or make available the representation resource that makes success possible, a distinction that aggregate scores conflate. Existing capability evaluations implicitly treat failure as insufficient information, reasoning ability, or scale; our results reveal a different failure mode: a model's demonstrations can contain the information, yet adaptation fails because the required representation resource is not available for adaptation. We do not claim to have completed such an evaluation; our evidence shows only that the outcome-only view is insufficient.

2 Related Work

This paper does not propose a new model or a benchmark. Prior work decomposes other aspects of in-context learning; our contribution is the identification of a representation constraint prior work does not measure.

2.1 What makes demonstrations work

A closely related line asks which properties of demonstrations drive adaptation. Ground-truth labels matter in some settings (Yoo et al., 2022), yet for classification tasks Min et al. show they matter surprisingly little: random labels perform nearly as well (Min et al., 2022), and in translation demonstration selection measurably shapes in-context performance (Agrawal et al., 2023). Our correspondence-dose result engages this contrast directly: in low-resource translation the correct pairing is behaviorally load-bearing (Finding 1), which we attribute to the task's demand for a specific source representation rather than to a general property of demonstration labels. Models can learn mappings over arbitrary symbols (Wei et al., 2023; Mirchandani et al., 2023), lexinvariant models test whether lexical identity is used (Huang et al., 2023), and substitution-cipher studies quantify in-context "learning" over novel mappings (Fang et al., 2025). Scaffolding explicit intermediate steps (Nye et al., 2021; Wei et al., 2022) is our supplied-intermediate (rescue) condition: CoT-style prompting makes the intermediate available, bypassing intermediate computation rather than rule identification (measured by the explicit-rule control). Prompt-corruption studies systematically perturb prompt components to show demonstration structure outranks surface meaning (Min

et al., 2022; Shivagunde et al., 2024). These lines ask how demonstrations work; we ask which representation source a task requires.

2.2 Representation formation during in-context learning

What representations does in-context learning produce? Demonstrations condense into a reusable task vector (Hendel et al., 2023), refined into causally-grounded function vectors (Todd et al., 2024). Later work characterizes the in-context representation that forms during adaptation (Park et al., 2025) and traces task-vector emergence and limitations (Dong et al., 2026). This line characterizes the representation the model *forms* and *uses*; we ask which source the task's mapping is a *function of*.

2.3 Mechanistic computation of in-context learning

Where and how does in-context computation occur? In-context learning has been framed as implicit learning: transformers can implement gradient descent in their forward pass (von Oswald et al., 2023) or some identifiable learning algorithm (Akyürek et al., 2023), and ICL learns label relationships rather than conventional learning (Kossen et al., 2024). Recent work decomposes this computation further, separating task *recognition* from task *learning* (Wang et al., 2025; Pan et al., 2023) and localizing the two to specific attention heads (Yang et al., 2026); a double dissociation study likewise separates a task *schema* from a *binding* mechanism via activation patching (Kim, 2025). This line decomposes *how* and *where* computation happens; we decompose *which source* adaptation depends on.

Closest to our decomposition, Tao et al. (2024) separate an ICL *inference* function from a *verbalization* function, validated by intervention; their split presumes the input feature is already fixed, whereas our explicit-rule control was designed to isolate an earlier feature-identification step, an interpretation the held-out control of Finding 3 leads us to retract (support-overlap reuse). Jang et al. (2024) expose a "demonstration shortcut" between pretrained semantic priors and the new input–label relation; we instead corrupt the source and measure a *selective* dissociation. Sun et al. (2025) analyze input–label mappings and their use; our direct probes locate the constraint upstream of the mapping.

2.4 Latent variables, task induction, and instruction induction

Adaptation has also been framed as inferring a latent task concept (an implicit Bayesian posterior over the task (Xie et al., 2022; Chan et al., 2022)) or recovering it as an explicit instruction (Honovich et al., 2023), while hypothesis search improves inductive reasoning by making a latent rule explicit (Wang et al., 2024). Latent-concept disentanglement studies how transformers represent numerical and transitive latent structure in controlled ICL tasks via mechanistic interpretability (Hong et al., 2026), and concept-aware data construction trains models to better exploit latent concepts from demonstrations (Štefánik et al., 2024). These lines ask *what* latent concept ICL infers, or how to train for it; we ask, behaviorally and on frozen models, *whether* the required intermediate is identified, computed, and used during adaptation, an intervention-based decomposition of *why* adaptation fails rather than a description of *what* is represented. The computation leg is not rule induction: the relational-order task fails even when the relation semantics are stated while the hidden order remains implicit. This is a failure to *produce*, rather than identify, the intermediate, a compositional limit of the kind Dziri et al. (2023) documents and a known failure mode of compositional generalization (Garg et al., 2022; Lake & Baroni, 2018; Hupkes et al., 2020). The schema/binding dissociation (Kim, 2025) is mechanistic (activation patching); ours is behavioral and isolates a third leg (intermediate computation) through the explicit-rule control.

2.5 Representation-centric views of model capability

What does a model represent, beyond its task score? A representation-centric line argues that concepts are encoded linearly in model internals (Park et al., 2024), and that models reason semantically rather than symbolically (Tang et al., 2023). Our cross-model result extends this direction: models differ not only in what they represent, but in which representation sources they can access or make available for a given task. This is the central concern of the elicitation and latent-knowledge line,

which recovers knowledge a model possesses but fails to express (Burns et al., 2023; Belrose et al., 2023; Mallen et al., 2023).

2.6 Beyond scalar capability evaluation

What does a scalar capability score miss? Broad benchmark suites summarize capability as an aggregate score over many tasks (Srivastava et al., 2022), and analyses of emergence caution that "emergent" abilities can appear or vanish with the metric chosen (Schaeffer et al., 2023). Our result makes this concrete: across five families the scalar joint accuracy on the numeric task varies (0.40–0.73) even though the dissociation and rescue are positive in every family (two weaker; Appendix), so the score alone does not reveal which representation source is at issue.

2.7 Position of this work

Prior work asks *what representation emerges*, *where ICL is computed*, and *how capability is scored*; we ask *which source a task requires and whether it is available during adaptation*. Our contribution is the intervention-based synthesis: source-targeted interventions (corrupting access, supplying the intermediate) that localize *failure* rather than representation. "Representation resource" is an organizing term for these interventions, not a new mechanistic primitive. We manipulate the source and show adaptation tracks its availability, complementary to, not a correction of, the lines above.

3 Framework: Representation Resources in In-Context Adaptation

3.1 Representation demand

A task's **representation demand** is the representation source on which its correct output is a function. Formally, let r_t denote the source a task demands and $\mathcal{A}_m$ the sources a model makes available during adaptation; success requires $r_t \in \mathcal{A}_m$, and the constraint this paper studies is the mismatch $r_t \notin \mathcal{A}_m$. We study three sources, each paired with the diagnostic intervention that targets it:

Source	Output is a function of	Diagnostic intervention
Surface identity	the input token itself	survives source corruption
Pretrained semantics	the input's meaning	collapses under source corruption
Computed intermediate	an episode-specific value	direct probe + explicit rule + rescue

These routes each host a matching intervention; together they form an intervention-based decomposition of representation-resource constraints. We call the third route a *computed intermediate* rather than a "novel latent": the operations (modulo, order) are familiar, and only the intermediate value is episode-specific. A task's demand is assigned by task construction and revealed behaviorally by which source-targeted intervention changes success; reading it automatically from an arbitrary task is an open problem, not something we solve.

Status of the framework. $\mathcal{A}_m$ is not measured independently of the tasks: it summarizes which interventions change success, so "success implies $r_t \in \mathcal{A}_m$" reads observed behavior, not an independent latent. "Available" names availability for *use*, whereas internal probes measure *existence* (a linear probe finds the intermediate predictable even when behavior fails; Appendix), so the framework is a *behavioral taxonomy*, not a predictive account.

3.2 Access versus intermediate availability

We decompose successful adaptation into behavioral conditions. *Access* is met when the required representation is already usable by the model (surface identity, pretrained semantics). When a task instead requires a computed intermediate, two conditions are separable by intervention: *intermediate computation* (the model produces the intermediate under direct elicitation, with no episode-specific label mapping) and *intermediate composition* (the model uses the supplied intermediate within the

episode-specific mapping). An apparent third distinction, rule induction, is inflated by *support overlap* (Finding 3): the query digit usually recurs among the demonstrations, so an observed digit-to-label correspondence can be reused rather than the rule induced. *Intermediate availability* summarizes whether the required intermediate is usable during adaptation.

3.3 Model profile

For a model M under an elicitation and decoding policy π (prompting, thinking mode, decode budget), a **representation-resource profile** records its measured ability on each route: $R(M, \pi) = (\text{surface access},\ \text{semantic access},\ \text{intermediate availability})$. The surface- and semantic-access coordinates are the normal (K=8) accuracies on the surface-bijection and semantic-category tasks; the intermediate-availability coordinate is the pair of joint and supplied-intermediate accuracies, reporting both the implicit-computation and the mapping-only endpoints. A profile is not three benchmark scores; it is three mechanism-relevant measurements, each tied to a failure mode and a targeting intervention. The profile is *diagnostic* (measured after the fact); a small five-model check shows a clean held-out signal only for the surface route, while the computed-intermediate transfer is confounded by support overlap (Appendix). The profile is conditional on π rather than intrinsic to M: the thinking-enabled control shows the intermediate-availability coordinate shifts under deliberation (Appendix).

3.4 The intervention chain

The paper's argument is a single intervention chain: correct correspondence yields graded adaptation; corrupting the source collapses it; the routes separate (surface robust, semantic fragile, computed intermediate not-learned); and supplying only the correct intermediate restores it. Direct-intermediate probes, an explicit-rule control, and a held-out-query control then characterize the computed-intermediate failure: the numeric task's apparent rule induction is inflated by support overlap, whereas the relational order is barely computed.

4 Experimental Setup

We study in-context adaptation under controlled interventions. No model is fine-tuned; every result is obtained by prompting a frozen decoder-only language model with demonstrations.

Models. We evaluate five decoder-only model families: Qwen3-8B (the primary model, carrying the full intervention analysis), and Llama-3.1-8B-Instruct, Gemma-3-12B-it, GLM-4-9B-chat, and Mistral-7B-Instruct-v0.3 (for the cross-model comparison). All are used with greedy decoding.

In-context mapping tasks. We construct four tasks that differ in the representation source their output is a function of. Each task presents $K = 8$ input–output demonstrations and queries a whole-string-unseen input (its functional feature may recur among the demonstrations; Finding 3); the label mapping is re-randomized per episode so the task cannot be solved from a fixed prior mapping.

- **Surface identity (cipher).** The output is an arbitrary bijection of the input token, a function of token identity alone.
- **Pretrained semantics (category induction).** The output is an arbitrary mapping of the input's semantic class, with the query an unseen member, so the task is a function of pretrained meaning.
- **Numeric residue.** The output is a function of an abstract intermediate (a digit's residue modulo 4), which is neither token identity nor a pretrained class.
- **Relational order.** The output is a function of a hidden total order assembled by transitive closure over oriented pairs, requiring a computed intermediate representation.

Metrics. For the translation-based correspondence analysis we report corpus-level chrF++ (micro-averaged); for the synthetic mapping tasks we report exact-label accuracy. Controlled tasks use N=200 episodes per condition and the translation dose uses N=200 (Uyghur) or 197 (Kazakh,

Models	Qwen3-8B (primary); Llama-3.1-8B, Gemma-3-12B, GLM-4-9B, Mistral-7B (cross-model)
Tasks	surface bijection; semantic category; numeric residue; relational order
Translation metric	chrF++ (corpus, micro-averaged)
Controlled-task metric	exact-label accuracy
Language grounding	Uyghur, Kazakh, Tibetan (traditional Mongolian as naturalistic validation)

Table 1: Models, tasks, and evaluation. All results are obtained by prompting frozen decoder-only models; no model is fine-tuned.

Language	q=0	q=2	q=4	q=6	q=8	**slope** β	**95% CI**
Uyghur	16.64	17.81	18.10	18.46	18.49	$+0.218$	$[0.085, 0.316]$
Kazakh	8.81	9.81	9.58	9.92	9.99	$+0.123$	$[0.057, 0.198]$
Tibetan	5.59	6.36	6.81	7.05	7.46	$+0.221$	$[0.129, 0.318]$
Mongolian (fragmented)	2.33	2.43	2.48	2.42	2.21	$\mathbf{-0.013}$	$[-0.079, 0.039]$

Table 2: Correspondence is a graded signal: chrF++ rises with the number of correctly paired demonstrations (q) in three well-represented languages, but not in traditional Mongolian, whose fragmented script makes the source unreadable (slope ≈ 0).

Tibetan) queries, each with 5 demonstration-order seeds averaged per query; all contrasts are paired bootstrap over the episode/query, with 10,000 resamples for the controlled tasks and 5,000 for the translation dose.

Data and languages. The correspondence and access analyses are grounded in low-resource machine translation from three typologically distinct source languages, Uyghur (Arabic script), Kazakh (Cyrillic script), and Tibetan (Tibetan script), into Chinese, using FLORES-200 parallel data; a ten-language screening extends the correspondence analysis. Traditional Mongolian (vertical script) is introduced as a separate naturalistic accessibility probe: its source side is Mongolian-script text from the MC^2 crawl (FLORES-200 covers Mongolian only in Cyrillic); it is not part of the primary benchmark.

Interventions. Each research question is answered by an intervention with a negative control: **correspondence dose** (vary how many of K demonstrations carry a correct pairing, $q \in \{0, 2, 4, 6, 8\}$); **source corruption** (injectively encode the source into private-use Unicode codepoints, holding content, demonstrations, and decoding fixed); **direct-intermediate probe** (ask the model to emit the task's intermediate with no episode-specific label mapping); **explicit-rule control** (state the latent rule while preserving the episode-specific label mapping); **intermediate rescue** (supply the task's intermediate value); and **specificity control** (supply a wrong or random feature at matched budget).

5 Results

We present three findings, each established by an intervention and stated with its limitation. Together they support the framework of Section 3: adaptation depends on whether the representation a task demands can be accessed or made available. Throughout, these results provide behavioral evidence for representation-resource constraints, not a mechanistic localization of circuits.

5.1 Finding 1: Correspondence provides a graded adaptation signal

Observation. Holding the demonstration set fixed and varying how many demonstrations carry a correct source–target pairing, performance rises with the number of correct pairings: correctly paired demonstrations beat shuffled or unpaired demonstrations of the same volume (raw $\gg$ shuffled $\approx$ unpaired). A same-format control (identical paired layout, only the target permutation changes) reproduces the graded dose (pairing-only benefit $+1.74$ / $+0.89$ / $+1.41$ chrF++), distinct from the raw q=0$\rightarrow$8 deltas of Table 2 ($+1.85$ / $+1.18$ / $+1.87$), which also change the format; the close agreement shows the graded signal is not an unpaired-pool artifact. The dose-response slope is positive in all three source languages (Uyghur, Kazakh, and Tibetan), and an exploratory ten-language screening is positive in all ten languages (8/10 CIs exclude zero; unadjusted; Table 7). Absolute translation quality is low (chrF++ 5.6–16.6, BLEU ≤ 4.5 at q=0), but COMET (a semantic-adequacy metric) is moderate (0.57–0.80), consistent with semantic adequacy despite lexical divergence; the corre-

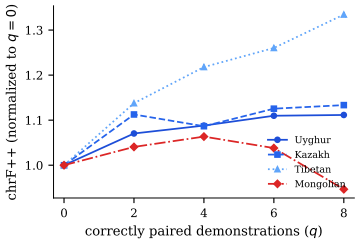

Figure 1: Correspondence is a graded adaptation signal: chrF++ (normalized to q=0) rises with q in three source languages, but not in traditional Mongolian, a tokenizer-fragmented naturalistic control.

Language	Δ_{08} normal	Δ_{08} byte-corrupted
Uyghur	$+1.97$ $[-0.72, 3.12]$	$+\mathbf{0.36}$ $[-0.07, 0.86]$
Kazakh	$+1.21$ $[0.24, 2.00]$	$+\mathbf{0.02}$ $[-0.20, 0.26]$
Tibetan	$+1.33$ $[-0.03, 2.77]$	$+\mathbf{0.31}$ $[-0.18, 0.82]$
Mongolian (naturalistic)	-0.12 $[-0.69, 0.35]$	— (tokenizer-fragmented)
Macro interaction	$+\mathbf{1.27}$ $[0.22, 1.84]$	
Selective dissociation J (Qwen3-8B)	$+0.83$ $[0.72, 0.93]$	
Selective dissociation J (five families)	$+0.26$ to $+0.95$ (positive)	

Table 3: Accessibility gates transfer: corrupting only the source reduces the correspondence effect (Δ_{08}) toward zero in all three languages, significant in aggregate (macro interaction $+1.27$ $[0.22, 1.84]$), with per-language normal-dose intervals excluding zero only for Kazakh, and the gating is selective on Qwen3-8B, and the dissociation is positive across all five families (Appendix). Traditional Mongolian, whose script is tokenizer-fragmented (1.53 tokens/char, the highest of the four languages), shows the same collapse without any intervention. The selective dissociation J is a difference-of-differences defined in the Appendix.

spondence effect is consistent in sign across chrF++, BLEU, and COMET (Appendix). The effect is not specific to translation: a synthetic surface bijection shows the same graded dose.

Intervention. A correspondence dose: vary how many of the K demonstrations carry a correct pairing ($q \in \{0, 2, 4, 6, 8\}$), holding the demonstration set fixed.

Interpretation. Correct correspondence is a graded adaptation signal, not a binary on/off switch, and not a translation artifact.

Limitation. This establishes *that* the signal exists; it does not yet say what makes it usable (that is Finding 2).

5.2 Finding 2: Representation accessibility gates whether demonstrations transfer

Observation. Injectively corrupting the source (holding content, demonstrations, and decoding fixed) reduces the correspondence benefit across languages, with a significant aggregate interaction ($+1.27$ chrF++). The gating is selective: on Qwen3-8B, a surface bijection survives the corruption that sends a semantic mapping to chance (selective dissociation $J = +0.83$), and the same collapse appears naturalistically in traditional Mongolian, whose vertical script the model tokenizes to the point of unreadability (a Mongolian-to-Chinese probe, read as qualitative corroboration rather than a direct replication). The selective dissociation is not an artifact of the surface task's abstract vocabulary: on Qwen3-8B, a surface bijection over natural words survives the same corruption equally ($I_{\text{nat}} = -0.08$ vs. the abstract -0.09 and semantic $+0.74$); across families the surface is not uniformly robust, yet the dissociation stays positive everywhere (Appendix).

Intervention. Source corruption: injectively encode the source into private-use Unicode codepoints, holding content, demonstrations, and decoding fixed. The encoding is information-preserving but also changes tokenization, so we add controls isolating this confound: a surface bijection over natural words and a tokenization-preserving word substitution on the semantic task; the latter drops q=8 accuracy from 0.855 to 0.280 ($\approx$ chance 0.25), matching the byte-corrupted 0.345 (Appendix).

Interpretation. The contrast is differential: on Qwen3-8B a surface bijection survives the same encoding change that collapses a semantic mapping, and this survives replacing the abstract tokens with natural words. The two tasks therefore depend on representation availability in different ways:

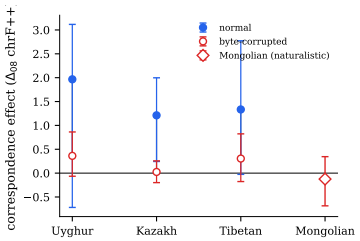

Figure 2: Accessibility gates transfer: the correspondence effect collapses when only the source is corrupted (normal vs. byte-corrupted), holding content, demonstrations, and decoding fixed. Traditional Mongolian, a tokenizer-fragmented script, shows the same collapse without any intervention.

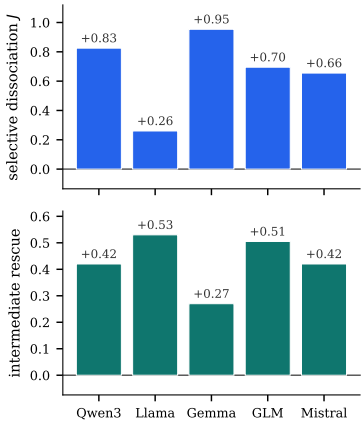

Figure 3: Both intervention effects are positive beyond the primary model: the selective dissociation J (top) and the intermediate rescue (bottom) are positive for every family (a descriptive cross-section; scale and architecture are confounded; full profile in the appendix).

Condition	Numeric	Held-out digit	Relational
No intermediate (joint)	0.58	0.025	0.58
Explicit rule	0.795	0.205	—
Direct intermediate	0.845	0.365	0.555
Correct intermediate	**1.00**	**0.995**	**0.91**
Wrong feature	0.54	—	0.62
Random feature	0.30	—	0.58
Rescue	+0.42	+0.97	+0.33

Table 4: The numeric-residue task's apparent computation is inflated by support overlap: under a held-out query digit (guaranteed never in the demonstrations) joint, explicit-rule, and direct accuracies fall to 0.025 / 0.205 / 0.365, while supplying the residue stays at 0.995; the relational task's query pair is held out by construction, so its failure is a genuine computation deficit. Note: numeric direct = 0.845 under the original 16-token setting; the matched 128-token baseline for the held-out comparison is 0.855. Relational joint = 0.58 at the matched 128-token budget (0.48 at 16 tokens).

correspondence is usable only insofar as the required source representation remains available, rather than corruption having destroyed meaning or merely fragmented tokenization.

Limitation. The tokenization-isolating controls now span all five families (Appendix): word substitution collapses the semantic task to chance everywhere, and the dissociation stays positive under either surface vocabulary.

5.3 Finding 3: A computed intermediate is a real constraint, but support overlap inflates apparent computation

Observation. A task whose output is a function of a computed intermediate is not reliably learned in-context: on Qwen3-8B the numeric-residue task reaches 0.58 (chance 0.25) and the relational-order task 0.58 (chance 0.50); supplying the correct intermediate rescues both to near-ceiling (rescue +0.42 numeric, +0.33 relational). The relational deficit is a genuine computation failure: its query pair is by construction the two endpoints of the hidden order and never appears among the demonstrations (a generator audit finds zero leakage over 5,000 episodes; Appendix), so it cannot be reused and is barely computed (0.555 under direct elicitation; chance 0.50). The numeric deficit, however, is inflated by *support overlap*: the demonstrations contain the query digit in roughly 80% of episodes, so the model may reuse an observed digit-to-label correspondence rather than generalize the modulo rule. A held-out-digit control (in which the query digit is guaranteed never to appear in any demonstration) collapses the numeric task: joint accuracy falls from 0.58 to 0.025 $[0.011, 0.057]$ (far below chance), the explicit-rule condition from 0.795 to 0.205 $[0.155, 0.266]$ (not distinguishable from the 0.25 chance level), and direct residue emission from 0.855 to 0.365 $[0.301, 0.434]$ (above chance but sharply reduced); supplying the residue preserves near-ceiling mapping accuracy (0.995), so the residue-to-label mapping itself is not the bottleneck.

Intervention. Direct-intermediate probes, an explicit-rule control, intermediate rescue, a specificity control, and a held-out-digit control (Appendix).

Interpretation. The numeric task's apparent computation is largely explained by support overlap rather than rule-based generalization: once the query digit is removed from the demonstrations, the joint, explicit-rule, and direct accuracies collapse (Table 4), so the original $0.58 \rightarrow 0.795 \rightarrow 0.845$ progression is consistent with direct instance reuse of observed digits, and we therefore retract the feature-identification reading. The relational task fails differently and robustly: relation inference stays near chance (0.555) even when the relation semantics are stated, and its query pair is held out by construction, so its failure is a genuine failure to produce the relation. Across both tasks the supplied-intermediate condition reaches near-ceiling, so the episode-specific intermediate-to-label *mapping* is readily usable even when the intermediate itself is not computed. We do not determine whether the model internally represents the intermediate.

Limitation. The held-out-digit control is run on Qwen3-8B for the numeric task; the relational task is support-held-out by construction and measured across all five families (Appendix). The controlled tasks use $K{=}8$ and one prompt template; a count sweep ($K{=}4, 8, 12$) and three equivalent instructions leave the joint failure unchanged (Appendix).

6 Discussion

6.1 Representation resources as an adaptation constraint

A common assumption is that in-context learning is a scalar function of quantity and size. Our evidence suggests a different structure: adaptation depends on *task demand* and *representation availability* instead, and "availability" is not monolithic: one task fails because the intermediate is barely computed, and the other's apparent computation is inflated by support overlap. The decomposition, not the lens, is the contribution.

6.2 Implications for evaluation and intervention

If capability is partly a matter of representation resources, outcome-only evaluation measures one projection of it; future ICL evaluation may characterize what a task *requires*, what a model can *access*, and what it can *make available*. The scalar joint accuracy alone (spanning 0.40–0.73 on the numeric task across five families) does not reveal which source is at issue. The decomposition also suggests, as *hypotheses not methods*, where future adaptation might act: representation-aware prompting, scaffolding, and training for intermediate availability.

6.3 Limitations

Controlled tasks and behavioral evidence. The intermediate-availability decomposition is demonstrated on two controlled abstractions on Qwen3-8B; the transportable findings (correspondence dose, access gating) are grounded in low-resource translation, but the intermediate-availability finding is not yet. "Access," "computation," and "composition" are behavioral distinctions established by intervention, not claims about located circuits.

Diagnostic and descriptive. The profile is measured after the fact; held-out transfer is suggestive but small, and the five-family comparison is descriptive ("models differ," not "scale causes X"; Appendix).

7 Conclusion

In-context learning has been studied as how much a model can do; our results argue it should also be studied as a resource-matching problem: whether the representation a task requires is available during adaptation. Through controlled interventions we identified an intermediate-representation constraint, and showed that adaptation fails not only from too little information, but from a demand–availability mismatch current evaluation does not yet measure.

AI USE STATEMENT

The authors used generative AI assistants for manuscript organization, language refinement, presentation and framing refinement, code documentation, and figure prototyping. All experimental designs, data collection, statistical analyses, interpretations, and scientific claims were developed, conducted, and verified by the authors. All AI-assisted outputs were reviewed and revised by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide all prompts, task generators, evaluation scripts, model configurations, and statistical-analysis code in a supplementary code archive accompanying the submission (the experiment scripts, with descriptive names such as `dose_response.py`, `source_corruption.py`, `linear_probe.py`, and `matched_attention.py`, together with prompt templates and evaluation code, as well as the figure-plotting scripts that reproduce each figure). The appendix describes the task definitions, experimental protocol, and failure examples in full.

ETHICS STATEMENT

This work is a scientific study of in-context learning on synthetic and publicly available translation data. It involves no human subjects, releases no new sensitive data, and presents no deployment of a system that raises foreseeable societal harm.

REFERENCES

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations (ICLR)*, 2023.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Puneesh Deora, Bhavya Vasudeva, Tina Behnia, and Christos Thrampoulidis. In-context occam's razor: How transformers prefer simpler hypotheses on the fly. In *Conference on Language Modeling (COLM)*, 2025.

Yuxin Dong, Jiachen Jiang, Zhihui Zhu, and Xia Ning. Understanding task vectors in in-context learning: Emergence, functionality, and limitations. In *International Conference on Learning Representations (ICLR)*, 2026.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhouxiang Fang, Aayush Mishra, Muhan Gao, Anqi Liu, and Daniel Khashabi. Icl ciphers: Quantifying "learning" in in-context learning via substitution ciphers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Shivam Garg, Dimitris Tsipras, Percy S. Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Guan Zhe Hong, Bhavya Vasudeva, Vatsal Sharan, Cyrus Rashtchian, Prabhakar Raghavan, and Rina Panigrahy. Latent concept disentanglement in transformer-based language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Qian Huang, Eric Zelikman, Sarah Li Chen, Yuhuai Wu, Gregory Valiant, and Percy Liang. Lexinvariant language models. *arXiv preprint arXiv:2305.16349*, 2023.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 2020.

Joonwon Jang, Sanghwan Jang, Wonbin Kweon, Minjin Jeon, and Hwanjo Yu. Rectifying demonstration shortcut in in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Chaeha Kim. Task schema and binding: A double dissociation study of in-context learning. *arXiv preprint arXiv:2512.17325*, 2025.

Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*, 2018.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. In *Conference on Robot Learning (CoRL)*, 2023.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.527.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. In *International Conference on Learning Representations (ICLR)*, 2025.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Namrata Shivagunde, Vladislav Lialin, Sherin Muckatira, and Anna Rumshisky. Deconstructing in-context learning: Understanding prompts via corruption. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4509–4529. ELRA and ICCL, 2024.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Michal Štefánik, Marek Kadlčík, and Petr Sojka. Concept-aware data construction improves in-context learning of language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12335–12352. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.733.

Chenghao Sun, Zhen Huang, Yonggang Zhang, Le Lu, Houqiang Li, Xinmei Tian, Xu Shen, and Jieping Ye. Interpret and improve in-context learning via the lens of input-label mappings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*, 2023.

Junyi Tao, Xiaoyin Chen, and Nelson F. Liu. Inference and verbalization functions during in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning (ICML)*, 2023.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis search: Inductive reasoning with language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, and Ji-Rong Wen. Investigating the pre-training dynamics of in-context learning: Task recognition vs. task learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

Haolin Yang, Hakaze Cho, and Naoya Inoue. Localizing task recognition and task learning in in-context learning via attention head analysis. In *International Conference on Learning Representations (ICLR)*, 2026.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwaran Lee, Sang-Woo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

A TASK DEFINITIONS

Each task presents $K = 8$ demonstrations and queries an unseen input; the label mapping is re-randomized per episode so the task cannot be solved from a fixed prior.

Surface bijection (cipher). *Input/output:* a sequence of source tokens ($V = 12$, length 4) maps through a random bijection f to a target-token sequence; a demonstration is $(a, f(a))$. *Required source:* token identity (surface). *Expected behavior:* learnable, because the bijection is preserved under injective encoding.

Semantic category mapping. *Input/output:* a word maps to one of four labels (A/B/C/D); each episode draws 4 categories from a 12-category lexicon, 2 demo words per category, and a query word that is an unseen member of one category. *Required source:* pretrained meaning. *Expected behavior:* learnable, but fragile under corruption, because access to meaning is cut.

Numeric residue. *Input/output:* an object string whose embedded digit is reduced modulo 4 (residues $\{0{:}[0,4,8], 1{:}[1,5,9], 2{:}[2,6], 3{:}[3,7]\}$), mapped to a per-episode label permutation. *Required source:* a computed intermediate (the residue). *Expected behavior:* fails absent the intermediate; rescued by supplying it.

Relational order. *Input/output:* an ordered pair of abstract tokens (`qx, jt, mv, bs, cr`) maps to whether the first precedes the second (A/B); the query is the most-distant pair, requiring transitive closure over the demonstrations. *Required source:* a computed intermediate (a relation). *Expected behavior:* fails absent the relation; rescued by supplying it.

B EXPERIMENTAL PROTOCOL

Demonstration construction. $K{=}8$ demonstrations per task. For translation, demonstrations are source–target sentence pairs; for the controlled tasks, they are input–label pairs with a per-episode label permutation.

Dose schedule. A correspondence dose varies $q \in \{0, 2, 4, 6, 8\}$, the number of correctly paired demonstrations. Two formats are reported: an unpaired format (the remaining $8{-}q$ form separate source and target pools marked "no one-to-one correspondence") and a same-format pairing-only control (all K keep the identical paired layout; only the target permutation varies). Nested subsets keep demonstration identity fixed across q.

Retrieval. For translation, demonstrations are retrieved by LaBSE sentence similarity to the query (top-8).

Decoding. Greedy decoding; the model's chat template with `enable_thinking=False`. GLM-4-9B is run with `use_cache=False` (its `modeling_chatglm.py` is incompatible with the transformers 4.57.6 cache object).

Evaluation metrics. Translation: corpus-level chrF++ (micro-averaged, `word_order=2`). Controlled tasks: exact-label accuracy; the surface-bijection task additionally reports token accuracy and exact match.

Bootstrap. All primary-model contrasts use paired bootstrap with the episode/query as the sampling unit, and per-query seed-averaging where multiple seeds exist; the controlled tasks and direct probes use $10{,}000$ resamples, and the translation dose uses $5{,}000$.

Thinking-enabled control. Decoding above disables the model's thinking mode. We re-ran the numeric joint condition with thinking enabled (greedy, 4096-token budget, answer extracted after the think block). Unconditional joint accuracy rises from 0.58 to 0.78 ($39/50$; 95% CI $[0.66, 0.88]$), and the model is correct on every episode on which it converges ($39/39$); 22% of episodes never exit the think block within the budget. As with the thinking-off numbers, this measurement is subject to the support-overlap confound of Finding 3 (the query digit recurs in roughly 80% of episodes), so it qualifies the magnitude of the joint accuracy, not the interpretation of the deficit. The main-text Finding-3 numbers remain the thinking-off measurements.

C Intervention details

For each intervention we state what changes, what is held fixed, and why that targets the intended factor.

Correspondence dose. *Changed:* how many of the K demonstrations carry a correct source–target pairing. *Held fixed:* demonstration identity, content, decoding. *Why it identifies the factor:* the only thing that varies is pairing correctness, so a graded response targets the value of correct correspondence, not volume or content.

Source corruption (byte). *Changed:* the source's surface form (each UTF-8 byte b mapped to `chr(0xE000+b)`, a private-use codepoint). *Held fixed:* content, demonstrations, decoding. *Why it identifies the factor:* the encoding is injective and content-preserving, providing a controlled probe of whether source-representation accessibility contributes to successful adaptation. The intervention may also alter lower-level encoding statistics, so we interpret it as evidence for an accessibility constraint rather than a complete mechanistic localization.

Direct-intermediate probe. *Changed:* whether the task's output is the intermediate itself (the residue, or the before/after relation) rather than an episode-specific label. *Held fixed:* episodes, K, decoding. *Why it identifies the factor:* it measures whether the model can emit the intermediate when the rule is stated, but it also removes the label mapping, so it does not by itself separate rule identification from composition.

Explicit-rule control. *Changed:* the instruction states the latent rule ("the label depends on the digit modulo 4") while the demonstrations and query keep the arbitrary label mapping. *Held fixed:* episodes, K, decoding, output format. *Why it identifies the factor:* it reveals the rule without supplying the intermediate, and together with the held-out-digit control separates genuine rule generalization from support-overlap instance reuse.

Intermediate rescue + specificity. *Changed:* whether the task's intermediate value is supplied alongside the input (correct, wrong, or random). *Held fixed:* prompt budget and format. *Why it identifies the factor:* supplying the correct intermediate, and only the correct one, restores the task, separating the mapping-only component from both "more information" and the intermediate computation. On the relational task a wrong or random feature gives a modest lift above the joint baseline (0.62/0.58 vs. 0.58), which may reflect the intermediate's format providing structural scaffolding rather than a causal dependence on its correctness.

Held-out-digit control. *Changed:* the query digit is sampled from the two digits that do not appear in that episode's demonstrations. *Held fixed:* episodes, K, decoding, output format. *Why it identifies the factor:* when the query digit never occurs among the demonstrations, an

observed digit-to-label correspondence cannot be reused, so any residual accuracy reflects genuine modulo-4 generalization rather than support-overlap instance reuse.

Support-overlap audit. For the relational task, support overlap is excluded by construction: the query pair is the two endpoints of the hidden order, and the demonstrations are drawn only from adjacent pairs and non-adjacent pairs that exclude the query pair; a generator audit over $5{,}000$ episodes confirms zero direct or reversed leakage of the query pair into the demonstrations. For the held-out numeric comparison we use the 128-token re-measurement of the original direct condition (0.855; the main-text 16-token measurement is 0.845), matching the decode budget of the held-out run.

Below-chance held-out error audit. The held-out joint accuracy (0.025, far below the 0.25 chance level) is systematic rather than random. The held-out query digit is by construction a residue-0 or residue-1 digit, but in $195/200$ wrong joint predictions the model's label corresponds to residue 3 (149) or residue 2 (43), the two classes whose digits are fully present in the demonstrations, and almost never to the held-out residue (0 or 1). Consistent with a failure to generalize the residue to the held-out digit, the model defaults to the label of a fully-exemplified residue class, which is never the correct one; the explicit-rule condition (0.205) breaks this default back to chance, and only supplying the residue (0.995) recovers the mapping.

Per-language gating interaction. The accessibility gate is not carried by a single language. The per-language interaction $I_L = \Delta_{08}(\text{normal}) - \Delta_{08}(\text{byte})$ is $+1.60$ $[-1.13, 2.90]$ (Uyghur), $+1.19$ $[0.26, 2.02]$ (Kazakh), and $+1.03$ $[-0.12, 2.13]$ (Tibetan); all three point estimates are positive and of comparable magnitude, and the byte-corrupted Δ_{08} collapses toward zero in every language ($+0.36$/$+0.02$/$+0.31$). Leaving out any single language keeps the pooled interaction positive ($+1.11$ to $+1.40$), and dropping the largest per-language effect (Uyghur) still yields $+1.11$ $[0.39, 1.74]$, excluding zero. The per-language intervals are wide because the effect is small relative to per-query variance at $N{=}100$ queries per language (the accessibility intervention evaluates a 100-query subset of the full $N{=}200/197$ query pools used for the correspondence dose); the macro-interaction is the appropriately powered pooled estimate.

D CROSS-MODEL MEASUREMENTS

The following are *descriptive* measurements across five model families on identical episodes; they are not a ranking, and scale and architecture are confounded (Table 5).

Model	Surface n/b	Semantic n/b	Joint / Rule / Direct / Supp.	Selective J	Params
Qwen3-8B	0.940 / 0.920	0.855 / 0.345	0.58 / 0.80 / 0.86 / 1.00	$+0.83$	8.0B
Llama-3.1-8B	0.439 / 0.213	0.585 / 0.215	0.40 / 0.50 / 0.99 / 0.93	$+0.26$	8.0B
Gemma-3-12B	0.961 / 0.919	0.985 / 0.275	0.73 / 0.88 / 0.96 / 1.00	$+0.95$	12B
GLM-4-9B	0.825 / 0.664	0.835 / 0.235	0.50 / 0.58 / 1.00 / 1.00	$+0.70$	9.4B
Mistral-7B	0.180 / 0.183	0.685 / 0.215	0.42 / 0.79 / 0.90 / 0.84	$+0.66$	7.25B

Table 5: Five-family descriptive profile on identical episodes (not a ranking; scale and architecture are confounded). Surface and Semantic "n/b" are the normal and byte-corrupted source conditions.

The numeric-decomposition columns are four conditions on the mod-4 residue task: *joint* (construct the residue and map it to a label), *rule* (the mod-4 rule is revealed but the label mapping is preserved), *direct* (emit the mod-4 residue itself), and *supplied* (the residue is provided, so only the label mapping remains). The corrected profile tells a consistent story. The selective dissociation J is positive for every family ($+0.83$, $+0.26$, $+0.95$, $+0.70$, $+0.66$), and the intermediate rescue is positive everywhere ($+0.42$, $+0.53$, $+0.27$, $+0.51$, $+0.42$); J and rescue are computed from unrounded accuracies (e.g., GLM joint 0.495 yields rescue $+0.51$). One caveat: the surface-bijection control that J uses is itself near its chance floor for Mistral (normal $q{=}8$ 0.18, chance 0.083) and weakly learned for Llama (0.44), so for these two families J reflects a barely-learned surface route rather than a robust one; the dissociation is cleanest on Qwen3, Gemma, and GLM, where the surface is both learned and corruption-robust. The numeric-decomposition columns (joint/rule/direct) are subject to the same support-overlap confound identified in Finding 3: they were measured with the query digit recurring among the demonstrations in roughly 80% of episodes, so the apparent rule lift and direct "computation" are largely instance reuse; the held-out control (Qwen3) collapses

joint/rule/direct to $0.025/0.205/0.365$. We therefore do not read these columns as feature identification or computation; only the supplied-intermediate column (residue provided) is a clean measure, and it is near-ceiling for every family (1.00, 0.93, 1.00, 1.00, 0.84), so the residue-to-label *mapping* is not the bottleneck in any family. Neither the gating nor the rescue ranks with parameter count. The relational task is support-held-out *by construction* (the query pair is the two endpoints of the hidden order and never appears among the demonstrations), so a cross-model run there is the clean unconfounded test of the intermediate-computation bottleneck that the support-overlapped numeric columns above cannot provide. Across the five families the direct relation probe (emit "before"/"after") stays near chance (0.485–0.620) while the supplied-intermediate condition (relation provided, map to a label) stays well above chance (0.755–1.000; Table 6):

Model	Direct (compute)	Supplied (map)	Gap
Qwen3-8B	0.555	0.905	$+0.35$
Llama-3.1-8B	0.485	0.755	$+0.27$
Gemma-3-12B	0.600	1.000	$+0.40$
GLM-4-9B	0.620	0.825	$+0.21$
Mistral-7B	0.570	0.810	$+0.24$

Table 6: Five-family relational-order profile on the support-held-out relational task: direct (compute) vs. supplied (map) accuracy and the compute–map gap; the query pair is held out by construction, so these contrasts are unconfounded by support overlap.

Gemma and GLM compute the relation marginally above chance ($0.60/0.62$), whereas Qwen3, Llama, and Mistral are at or below chance ($0.555/0.485/0.57$); supplying the relation lifts every family well above chance (0.755–1.000). The computation, not the mapping, is therefore the main observed limitation in every family, and the unconfounded evidence no longer rests on Qwen3 alone. The joint (compute-and-map) condition is not reported here because it elicits verbose reasoning on this transitive-closure task (unlike the terse direct and supplied probes), so its accuracy is contaminated by truncation; the direct and supplied probes are the clean compute/map endpoints.

Task identifiability of the rule. "Rule identification" presumes the demonstrations fix the intended rule; for the numeric task the rule is identified only relative to a hypothesis class. Under the unrestricted class of arbitrary functions of the object string, no finite demonstration set fixes the rule, so the model must commit to a structured hypothesis (here, that the label is a function of the embedded digit modulo some m) through a prior, the hypothesis preference studied directly by Deora et al. (2025), who find in-context learners favor simpler hypotheses among those compatible with the demonstrations. This is the ordinary ambiguity of induction, not a deficit specific to our setting. We therefore read "rule-identification failure" as a failure to commit to that structured rule from an underdetermined sample, not as a failure to recover a logically-implied rule. This is consistent with the held-out control of Finding 3: once the query digit is removed, the model does not generalize the modulo rule even when it is stated. The $K{=}4, 8, 12$ sweep does not help for the same reason: more demonstrations re-sample the same residue classes without resolving the ambiguity of the underlying hypothesis.

Matched-attention control for the explicit-rule lift. The explicit-rule instruction names both the feature (the digit) and the rule (modulo 4); a feature-only control ("the label depends on a property of its embedded digit") names the digit without stating the rule. On Qwen3-8B the feature-only condition reaches 0.795, matching the explicit-rule condition (0.795; difference $+0.000$ $[-0.030, +0.030]$). This parity is explained by support overlap rather than by feature identification: both conditions leave the query digit present in the demonstrations in roughly 80% of episodes, so both are solved by reusing the observed digit-to-label correspondence, and the held-out control (Finding 3) collapses both to at or below chance.

E HELD-OUT TRANSFER

To test whether the profile *predicts* rather than merely *describes*, we evaluate held-out instantiations of the same representation sources and ask whether a model's profile coordinate predicts its held-out accuracy across the five families (a descriptive $N{=}5$ cross-section, not a powered test). The surface route gives a clean held-out signal: a held-out surface bijection over *natural words* is perfectly rank-correlated with the surface-access coordinate (Spearman $\rho = 1.00$, exact two-sided permutation

$p = 0.017$). The computed-intermediate route is inconclusive: both the mod-4 joint coordinate and the held-out mod-5 target are subject to the support-overlap confound of Finding 3 (the mod-5 task uses K=10, so its query digit also recurs among the demonstrations), so the mod-4→mod-5 correlation (Spearman $\rho = 0.90$, exact two-sided permutation $p = 0.083$, reproducing the ranking up to a single Mistral/GLM inversion) may reflect shared digit-reuse rather than a clean computed-intermediate profile; a scalar baseline (parameter count) does not predict mod-5 ($\rho = 0.36$, $p = 0.55$). We therefore read only the surface route as carrying a held-out signal beyond scalar capability; a support-held-out mod-5 measurement would be needed to test the computed-intermediate route cleanly.

F LINEAR-PROBE DIAGNOSTICS

To test whether the intermediate is *present* even when behavior fails (turning "available for use" from a behavioral label into a representation-level, falsifiable distinction from *existence*) we probe the mod-4 residue linearly from the hidden states in the failed joint condition. For each episode we take the last-token hidden state at every layer and fit a ridge classifier (standardized features, 5-fold stratified cross-validation) predicting the query residue (4-way, chance 0.25). The residue is linearly predictable at near-ceiling for every family (best-layer accuracy Qwen3 0.995, Llama 0.970, GLM 1.000, Mistral 1.000, Gemma 1.000) even though their joint behavior is far lower (0.58/0.40/0.50/0.42/0.73), and the embedding layer is near chance (0.29), so the residue is not trivially readable from the answer token itself; it becomes linearly predictable only after attention routes the query digit's information forward. The per-layer profile is model-specific: Qwen3-8B builds residue predictability across layers and retains it to the final layer (0.995), whereas the other four peak early (Llama/GLM layer 5, Mistral layer 2, Gemma layer 30) and lose it toward the output (0.39/0.73/0.84/0.83). The residue is thus linearly predictable even though the joint behavior fails. The held-out control (Finding 3) clarifies that this joint failure is largely instance reuse under support overlap rather than a feature-location deficit, so the "predictable but not used" reading should not be read as evidence that feature identification is the dominant failure leg. We do not claim the predictability is causally used, nor that the model computed the residue as a task-relevant intermediate: the residue is a simple function of the digit, so linear predictability establishes that the digit's information is routed forward, not a causal mod-4 computation.

G SELECTIVE DISSOCIATION STATISTIC

The selective dissociation J (Table 3) quantifies whether source corruption suppresses the correspondence-dose effect more for the semantic task than for the surface bijection. For each task, let $\Delta_{\mathrm{lv}} = \mathrm{Acc}(\mathrm{lv}, q{=}8) - \mathrm{Acc}(\mathrm{lv}, q{=}0)$ be the correspondence-dose effect under source level $\mathrm{lv} \in \{\mathrm{normal}, \mathrm{byte}\}$, and define the per-task interaction $I = \Delta_{\mathrm{normal}} - \Delta_{\mathrm{byte}}$ (a difference of differences). Then $J = I_{\mathrm{sem}} - I_{\mathrm{sym}}$, the difference between the semantic interaction and the symbolic (surface-bijection, token-accuracy) interaction. J is estimated by paired episode-level bootstrap (10,000 resamples) that pairs the same episode indices across the two tasks. On Qwen3-8B, $I_{\mathrm{sem}} = 0.735$ $[0.63, 0.84]$, $I_{\mathrm{sym}} = -0.09$ $[-0.125, -0.055]$, and $J = 0.825$ $[0.72, 0.93]$; the dissociation is positive across all five families (J from $+0.26$ to $+0.95$).

H TEN-LANGUAGE SCREENING

The correspondence-dose result extends to an exploratory ten-language screening (Qwen3-8B, N=100 queries per language, K=8, three demonstration-order seeds; FLORES dev/devtest). The table reports the positive-correspondence contrast (raw vs. unpaired demonstrations) and the source-script tokenization cost (tokens/char). The benefit is positive in all ten languages and its bootstrap CI excludes zero in eight of ten (no multiplicity adjustment). The benefit does not anti-correlate with tokenization cost (Spearman $\rho = +0.44$, exact one-sided $p = 0.10$), so we find no evidence for a monotonic negative association between the benefit and tokenization cost in this ten-language sample; the traditional-Mongolian collapse is consistent with an unreadability threshold rather than a density trend.

Language	tokens/char	raw − unpaired	95% CI
Amharic	1.55	$+1.17$	$[0.10, 2.21]$
Tigrinya	1.59	$+1.66$	$[0.46, 2.92]$
Burmese	1.60	$+1.68$	$[0.35, 3.04]$
Nepali	1.10	$+0.45$	$[-0.57, 1.58]$
Swahili	0.45	$+1.50$	$[0.67, 2.40]$
Mongolian	0.72	$+0.61$	$[-2.00, 1.68]$
Telugu	1.65	$+1.91$	$[1.09, 2.90]$
Khmer	1.18	$+1.98$	$[0.56, 3.57]$
Sinhala	1.70	$+1.72$	$[0.46, 3.16]$
Georgian	1.12	$+2.88$	$[1.78, 4.07]$

Table 7: Ten-language screening (Qwen3-8B, N=100 queries per language, K=8): the positive-correspondence contrast (raw vs. unpaired demonstrations) and the source-script tokenization cost. The benefit is positive in all ten languages; its bootstrap CI excludes zero in eight of ten. "Mongolian" here is Cyrillic Mongolian; the traditional vertical Mongolian script is a separate naturalistic probe.

I ACCESSIBILITY CONFOUND CONTROLS

Because the byte encoding changes both tokenization and embedding inputs, we separate these factors with two controls run on all five families at a uniform 128-token decode budget (Table 8). A tokenization audit shows the encoding inflates the token count of a source by $\times 6$ for abstract surface tokens, $\times 12$ for natural words, and $\times 6$–$\times 9$ for the Uyghur/Kazakh/Tibetan translation sources, so the surface task's robustness cannot be attributed to tokenization invariance.

Model	word-subst. q=8	I_{nat}	I_{sem}	J_{nat}
Qwen3-8B	0.28	-0.08 $[-0.11, -0.05]$	0.735	$+0.82$
Llama-3.1-8B	0.25	$+0.26$ $[0.18, 0.33]$	0.470	$+0.21$
Gemma-3-12B	0.18	$+0.03$ $[-0.01, 0.07]$	0.945	$+0.92$
GLM-4-9B	0.18	$+0.21$ $[0.15, 0.27]$	0.725	$+0.51$
Mistral-7B	0.22	$+0.03$ $[-0.03, 0.08]$	0.675	$+0.65$

Table 8: Five-family accessibility-confounding controls at a uniform 128-token decode budget: tokenization-preserving word-substitution accuracy on the semantic task (q=8), the natural-word surface interaction I_{nat}, the semantic interaction I_{sem}, and the natural-word selective dissociation J_{nat}.

Tokenization-preserving word substitution. Replacing each source word of the semantic task with a random word from a disjoint single-token vocabulary destroys pretrained semantics while preserving tokenization. The q=8 accuracy collapses to chance (0.25) in every family, so the byte-corruption collapse is not a tokenization artifact in any family.

Natural-word surface bijection. Replacing the abstract $V = 12$ vocabulary with natural words and re-running the surface bijection tests whether its corruption robustness is an abstract-vocabulary artifact. The interaction I_{nat} is near zero on Qwen3 (-0.08) and Gemma ($+0.03$), so the surface is genuinely corruption-robust on those families; it is positive on Llama ($+0.26$) and GLM ($+0.21$), where even the natural-word surface bijection is corruption-sensitive; and Mistral learns the normal surface to only 0.21, so its near-zero interaction is uninformative. Crucially, recomputing the selective dissociation with the natural-word surface ($J_{\text{nat}} = I_{\text{sem}} - I_{\text{nat}}$) remains positive in every family ($+0.82$, $+0.21$, $+0.92$, $+0.51$, $+0.65$), so the claim that semantic mappings collapse far more than surface mappings is robust to the abstract-vs-natural vocabulary choice even though absolute surface robustness is not universal.

J TRANSLATION QUALITY

The primary translation metric is corpus chrF++, which is low in absolute terms. To establish that the correspondence-dose effect reflects genuine translation rather than a format-level artifact, we report three complementary signals at q=0 vs. q=8: chrF++, BLEU, and COMET (wmt22-comet-da), together with qualitative examples.

BLEU is near zero throughout, but COMET (which scores semantic adequacy rather than lexical overlap) is moderate, consistent with the model producing semantically reasonable but lexically divergent translations; the correspondence effect is consistent in sign across all three metrics. Qual-

Language	**chrF** q=0	**chrF** q=8	**BLEU** q=0	**BLEU** q=8	**COMET** q=0	**COMET** q=8
Uyghur	16.64	18.49	0.76	1.55	0.80	0.82
Kazakh	8.81	9.99	4.42	4.47	0.64	0.66
Tibetan	5.59	7.46	0.50	2.29	0.57	0.59

Table 9: Translation-quality metrics at q=0 vs. q=8 for the three primary languages: corpus chrF++, BLEU, and the COMET (wmt22-comet-da) reference-free score.

Research question	**Intervention (control)**	**Observation**	**Conclusion**
When are demonstrations useful?	correspondence dose q (shuffled $\approx$ unpaired)	chrF++ rises with q (3 languages)	correspondence is a graded signal
Does accessibility gate transfer?	source corruption, source-only (surface bijection survives; content held fixed)	Δ_{08} collapses (3 languages)	accessibility gates usefulness
Is a computed intermediate available?	direct probe + explicit rule + rescue + held-out-digit control	held-out digit collapses numeric joint/rule/direct; relational direct $\approx$ chance	support overlap inflates apparent computation; supplied mapping intact

Table 10: Evidence map: one intervention decomposition per row: an intervention with its control, the observation, and the conclusion. The three rows map to the three findings; the interventions, not the experiments' order, are the organizing principle.

itatively, the gain from q=0 to q=8 is a genuine recovery of meaning rather than a format change: for a Kazakh source about drinking alcohol in moderation, q=0 renders a fluent hallucination ("you have many friends; the more friends, the greater your influence") while q=8 recovers the source ("drink in moderation; alcohol harms the body"); for a source about moose posing a danger, q=0 reads "an American product" where q=8 correctly reads "elk are not aggressive".

Two qualifications situate this. First, the correspondence effect, while modest in absolute chrF terms, is reliably positive; the dose-trend slope has a 95% confidence interval excluding zero in each of the three languages (Table 2). Second, as a quality reference, a dedicated translation model (tencent/Hy-MT2-7B, prompted zero-shot) reaches chrF++ 26.8/17.4/15.1 on Uyghur/Kazakh/Tibetan-to-Chinese, versus Qwen3-8B's 18.5/10.0/7.5 at q=8, so the Finding 1 effect operates in a low-quality regime, well below the quality a dedicated MT system achieves.

K Robustness to demonstration count and prompt wording

The intermediate-availability decomposition uses K=8 demonstrations and a single prompt template, which invites two alternative accounts of the joint numeric failure: that it is sample-inefficient rule induction (which more demonstrations would cure) and that it is sensitive to instruction wording. We test both on the numeric-residue task (Qwen3-8B, N=200, seed 0, the same episodes as the main result).

Demonstration count. Holding the instruction fixed and varying the demonstrations per residue class (K=4, 8, 12, all demonstrations correct), accuracy is 0.365 [0.30, 0.43], 0.58 [0.51, 0.65], and 0.525 [0.46, 0.60]. Increasing K from 8 to 12 (a 50% increase) does not raise accuracy (the point estimate is slightly lower, with overlapping intervals), and all three are far below the explicit-rule baseline of 0.795. The failure is therefore not a matter of too few demonstrations.

Prompt wording. Re-running K=8 with three semantically equivalent instructions yields 0.58, 0.615 [0.55, 0.68], and 0.64 [0.58, 0.71], all far below the explicit-rule baseline. The failure is not an artifact of a single wording.

Together these controls show the joint numeric failure at K=8 is not a matter of sample-efficiency or prompt sensitivity; its underlying cause is the support-overlap confound identified in Finding 3. We nonetheless state the intermediate-availability finding for the K=8, single-template setting.

L REPRODUCIBILITY AUDIT

Claim	Section	Table/Figure	Appendix
correspondence graded	Finding 1	Table 2, Fig 1	§Protocol, Dose
accessibility gates	Finding 2	Table 3, Fig 2	§Source corruption
rule ID / computation / joint gap	Finding 3	Table 4	§Direct probe, Explicit-rule, Rescue
model-dependent profile	Finding 3	Fig 3	§Cross-model

Every claim resolves to a main-text section, a table/figure, and the appendix protocol that produced it.

M EXPERIMENT ACCOUNTING

Experiment	Model	N	Seeds	Bootstrap
Correspondence dose (translation)	Qwen3-8B	200 / 197	5	5,000
Accessibility / corruption	Qwen3-8B	100	3	10,000
Controlled tasks (4 abstractions)	Qwen3-8B	200	1	10,000
Direct-intermediate probes	Qwen3-8B	200	1	10,000
Explicit-rule control	Qwen3-8B	200	1	10,000
Cross-model profile	5 families	200	1	—

The controlled tasks, direct probes, and cross-model comparison all share the same seed-0 episode set, so conditions are paired by episode. The translation dose pairs conditions by query and averages 5 demonstration-order seeds per query. The accessibility intervention evaluates a 100-query subset per language (3 demonstration-order seeds for the unpaired $q{=}0$ baseline, 1 for the paired $q{=}8$).

N PROMPT TEMPLATES

Numeric residue, original: "`Map each item to its label. Output only the label.`", then `{obj} -> {label}` per demonstration and `{query} ->`.

Numeric residue, supplied-intermediate (rescue): the same, except each demonstration is written `{obj} (mod4={residue}) -> {label}`.

Relational order: "`The label of a pair depends on which of the two items comes first. Output only the label.`", with `{X} {Y} -> {label}` (original) or `{X} {Y} (before/after) -> {label}` (correct).

Numeric residue, direct: "`For each item, output the value of the digit modulo 4. Output only the number (0, 1, 2, or 3).`", with `{obj} -> {residue}` per demonstration.

Numeric residue, explicit rule: "`The label of each item depends on the value of its digit modulo 4. Infer the mapping from modulo-4 values to labels from the demonstrations. Output only the label.`", with `{obj} -> {label}` per demonstration (the residue is not supplied).

Relational order, direct: "`For each pair, output whether the first item comes before or after the second in the hidden order. Output only 'before' or 'after'.`", with `{X} {Y} -> before/after`.

Semantic category: "`Map each word to its label. Words of the same kind share the same label. Output only the label.`", with `{word} -> {label}`.

Translation (dose): aligned demonstrations `{src_lang}: {src} / Chinese: {tgt}`, unpaired demonstrations listed as two separate pools marked "no one-to-one correspondence".

O REPRESENTATIVE FAILURE CASES

Joint numeric failure, rescued by the intermediate. The query `fx_8` (residue 0, label A) is answered *wrong* under the implicit-rule joint condition but *correct* once the residue is supplied (`fx_8 (mod4=0) ->`). As the held-out control shows, such failures are dominated at the aggregate level by support-overlap instance reuse rather than by a pure composition bottleneck.

Naturalistic access failure. Under traditional Mongolian, the model emits fluent-but-wrong target text (well-formed Chinese that ignores the source), the same signature as injected byte corruption, without any intervention. The fragmentation is not Qwen3-specific: all five model tokenizers split the vertical script at least as much as Qwen3-8B, so the failure is consistent with a tokenizer/script-accessibility limitation rather than a single-model quirk. This situates our naturalistic control within the tokenizer-fairness line: subword tokenization is measurably unfair across languages, fragmenting low-resource scripts and taxing monolingual performance (Ahia et al., 2023; Petrov et al., 2023; Rust et al., 2021); our contribution is the behavioral link from that tokenizer-level disparity to in-context accessibility, not the disparity itself.