Sense Representations Are Inducible Interfaces

Anonymous ACL submission

Abstract

Sense representations (explicit, per-token meaning decompositions) are useful for disambiguation, steering, and cross-lingual alignment, but existing approaches require models to be pretrained with sense structure baked in. We introduce ACROS, which *induces* an explicit sense pathway into a frozen pretrained decoder LM through a gated residual addition. On SmolLM2-360M, ACROS preserves base LM quality while supporting three uses of the same induced variables: zero-shot word-sense disambiguation (64.95 F1 on Raganato ALL, competitive with the WordNet first-sense heuristic), low-KL lexical steering across 5,161 CoInCo cases where a simple non-oracle proxy recovers about 90% of positive shifts, and SENSIA cross-lingual adaptation to four languages (mean R@1 0.988, target FLORES PPL 7.94). ACROS makes sense representations an inducible interface for ordinary pretrained LMs.

1 Introduction

Words carry multiple meanings, and a model that can explicitly represent, manipulate, and align those meanings gains capabilities that dense, opaque hidden states do not easily support. Sense-aware systems already exist for individual tasks: disambiguation (Blevins and Zettlemoyer, 2020; Loureiro et al., 2022), representation steering (Turner et al., 2024; Rimsky et al., 2024), and cross-lingual alignment (Cruz et al., 2026). But each is tied to its own architecture or training procedure. No single system exposes the same sense objects for measurement, intervention, and alignment.

Pretrained LMs are the natural unifying framework, as they already encode rich lexical structure (Petroni et al., 2019), but keep it locked inside dense hidden states. Backpack Language Models (Hewitt et al., 2023) come closest to exposing it, giving each token multiple explicit sense vectors. The catch is that Backpacks must be pretrained as

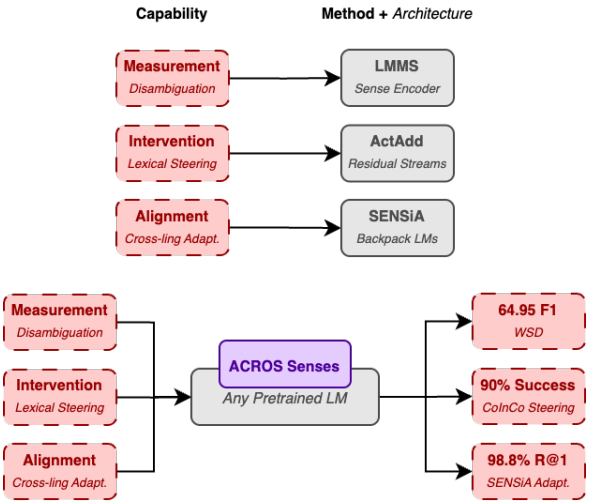

Figure 1: **ACROS unifies measurement, intervention, and alignment in one residual sense interface designed for pretrained decoder LMs.** Existing sense-based capabilities were previously tied to different architectures, with no single interface that supports all three. ACROS inserts a residual sense pathway so the same induced variables can be read, steered, and aligned.

Backpacks. Most strong checkpoints are ordinary decoders, and as we show in Section 2, converting one into a Backpack creates a representational bottleneck that destroys prediction quality.

In this paper, we propose **Architecture-agnostic Residual Induction of Sense Representations** (ACROS), a method that induces an explicit sense pathway into a frozen decoder LM through a gated residual addition, without modifying the original prediction path. Our contributions are:

1. A residual sense-induction method that adds a reusable sense interface to a pretrained decoder LM while preserving its prediction quality.

2. A unified evaluation showing that the *same* induced variables support word-sense disambiguation, low-KL lexical steering, and SENSIA cross-lingual adaptation: three roles that

previously required separate systems.

3. Ablations examining architecture portability across three decoder families, the role of the explicit sense pathway relative to dense hidden-state readout, and interface capacity and scale.

2 Background and Motivation

Sense representations serve three roles in NLP, each supported by separate systems. **Measurement** systems disambiguate words in context through gloss-informed embeddings (Blevins and Zettlemoyer, 2020; Loureiro et al., 2022) or multi-sense distillation (Wang et al., 2025, 2026). **Intervention** methods steer model behavior by modifying internal representations (Keskar et al., 2019; Dathathri et al., 2020; Yang and Klein, 2021; Turner et al., 2024; Rimsky et al., 2024; Zou et al., 2025; Arad et al., 2025; Chalnev et al., 2024), though recent evaluations find this can be brittle (Wu et al., 2025; Da Silva et al., 2025). **Alignment** methods use sense structure to anchor cross-lingual adaptation (Cruz et al., 2026; Hewitt et al., 2023). However, a WSD model cannot steer generation; a steering method cannot be evaluated as WSD; neither provides sense anchors for alignment. For senses to become a *general-purpose interface*, all three roles must be supported by the same representations inside the same model.

2.1 The Backpack Bottleneck

Backpack Language Models (Hewitt et al., 2023) come closest to this goal. Each token has K explicit sense vectors with learned contextualization weights, and the same machinery can be measured, edited (Hewitt et al., 2024), and used for adaptation. However, Backpacks must be pretrained in that form. Most strong LM checkpoints are ordinary decoders, so the practical question is whether an existing LM can be *converted* into a Backpack.

We test this directly by converting SmolLM2-360M (Allal et al., 2025)[1], an English-only Llama-style decoder, into converted Backpack models through continued training, distillation, and frozen-backbone controls. Table 1 tells a consistent story: every conversion variant learns non-collapsed sense vectors, but sharply degrades ordinary LM behavior. FLORES perplexity (Goyal et al., 2022) jumps from 25.1 to between 196 and 357; LAMA factual recall (Petroni et al., 2019) falls well below half of the original. The frozen-backbone control is especially telling: even when the original backbone and LM head are kept fixed and only the sense-side modules are trained, prediction quality still collapses. The failure is not in the training procedure; it is in the output architecture. Appendix A gives extended diagnostics across additional conversion variants.

Variant	PPL	LAMA@1	Sense Sep.
Original	25.1	0.3155	–
BP + CPT	356.9	0.0695	-0.0833
BP + Distil.	196.3	0.0420	-0.1275
BP + Distil. + Freeze	236.9	0.1245	-0.0308

Table 1: **Converting a pretrained LM to a Backpack model creates sense representations, but the resulting model loses the original LM's predictive ability.** PPL is English FLORES perplexity; LAMA@1 is filtered factual cloze accuracy. Sense separation (Sense Sep.) is the mean off-diagonal cosine between sense vectors, where lower or negative values indicate that slots are not collapsing onto the same direction.

Why conversion fails. The bottleneck has a geometric explanation. In a converted Backpack, the pre-logit representation is a convex mixture of K sense vectors: $h_q^{\mathrm{BP}} = \sum_k \alpha_{q,k} v_{q,k}$, with $\alpha_{q,k} \geq 0$ and $\sum_k \alpha_{q,k} = 1$. This forces all prediction-relevant information through a space spanned by at most K vectors. Figure 2 shows the problem concretely: we collect 10,000 final-layer hidden states from the pretrained SmolLM2 backbone on its own pretraining corpus, FineWeb, and measure how much variance the top k singular components capture. At the canonical sense count (K=32), only 84.9% of the variance is retained; reaching 95% requires rank 287, and 99% requires rank 697. A converted Backpack with practical K simply cannot represent the full space that the pretrained LM uses for prediction.

This is not a formal impossibility result (Backpack sense vectors are input-dependent and could in principle span a larger space), but it explains why conversion consistently fails even under favorable conditions. The residual design of ACROS avoids this bottleneck entirely: instead of replacing the hidden state with a sense mixture, we *add* a sense mixture to the existing hidden state.

[1] SmolLM2 was chosen because it is text-only (which eliminates potential visually-learned confounds), English-only (which eliminates multilingual confounds relevant to later evaluation via SENSIA), and has a fully-open pretraining corpus (which prevents test-set leakage concerns).

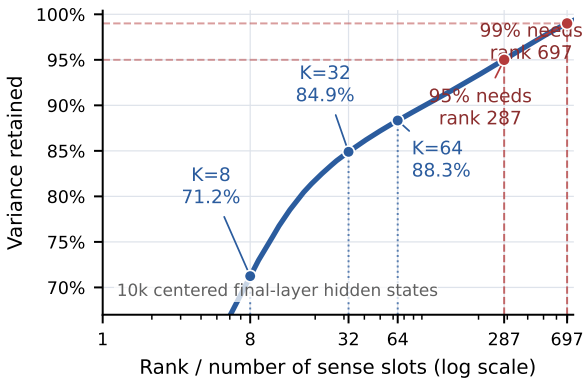

Figure 2: **Conversion fails because practical sense counts cannot span the hidden-state space the LM already uses.** On SmolLM2-360M final-layer states, practical sense counts (K=8, 32, 64) retain far less variance than the ranks needed for 95% or 99% coverage, explaining why Backpack conversion harms prediction quality.

3 ACROS

ACROS adds an explicit sense pathway to a pretrained decoder LM while leaving the original prediction path intact. The design principle is simple: rather than forcing all prediction through a sense mixture (which creates the bottleneck shown in Section 2), we add a sense mixture *alongside* the frozen base hidden state and let a learned gate control how much influence it has.

Architecture. The sense MLP maps token embeddings to K source-token sense vectors[2] $E_{k,j}$. A contextualization layer maps frozen backbone states to per-sense query/key attention weights $C_{k,q,j}$, selecting which source-token senses enter each position's residual update. For base hidden state B_q, scalar gate g, and output matrix W:

$$M_q = \sum_k \sum_j C_{k,q,j} E_{k,j}, \tag{1}$$

$$H_q = B_q + g M_q, \tag{2}$$

$$\text{logits}_q = H_q W^\top. \tag{3}$$

The gate is initialized to zero, so ACROS initially matches the base LM exactly and learns only whether to open the residual sense path. A diagram of the architecture is given in Figure 3. Full implementation details, including the contextualization layer parameterization, parameter counts, and latency measurements, are given in Appendix C.

[2]We use *sense* operationally to mean a continuous residual direction that captures a distinct lexical aspect of a token. These vectors are induced from the language-modeling objective and are not discrete lexicographic entries; the WSD evaluations and qualitative neighbor analyses show that they nevertheless align with human sense distinctions.

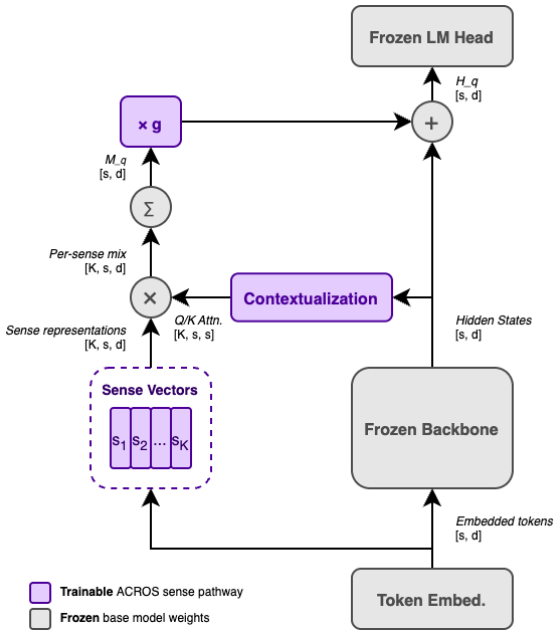

Figure 3: **ACROS keeps the original LM intact and learns a gated residual sense path beside it.** Token embeddings feed both the frozen backbone (producing base hidden state B_q) and a trainable sense MLP (producing K sense vectors per token). A contextualization layer computes per-sense Q/K attention weights from the backbone's hidden states, which are multiplied with the sense vectors and summed to form the sense mixture M_q. A scalar gate g, initialized to zero, controls the residual addition $H_q = B_q + g \cdot M_q$ before the frozen LM head. Gray = frozen; purple = trainable.

Induction. The induced model starts from the same checkpoint, freezes the original backbone and LM head, and trains only the sense MLP, contextualization layer, and scalar gate against the frozen base LM teacher.

For sequence labels y_t, binary token mask m_t, teacher logits z_t^T, ACROS logits z_t^S, and normalization constant $M = \sum_t m_t$, the full induction loss is

$$\mathcal{L} = \alpha \mathcal{L}_{\text{CLM}} + (1-\alpha)\tau^2 \mathcal{L}_{\text{KD}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}. \tag{4}$$

The knowledge-distillation term keeps ACROS close to the pretrained decoder's output distribution. With $p_{t,\tau}^T = \text{softmax}(z_t^T/\tau)$ and $p_{t,\tau}^S = \text{softmax}(z_t^S/\tau)$:

$$\mathcal{L}_{\text{KD}} = \frac{1}{M} \sum_t m_t \text{KL}\big(p_{t,\tau}^T \| p_{t,\tau}^S\big). \tag{5}$$

The causal language modeling (CLM) term keeps the induced model grounded in ordinary next-token

Metric	Original	Frozen BP	ACROS
FLORES PPL ↓	25.1	236.9	**25.09**
LAMA@1 ↑	0.3155	0.1245	**0.3175**
Sense separation	–	-0.0308	-0.0105
WS-353 max ↑	–	0.139	**0.487**

Table 2: **This residual path preserves LM quality while learning useful sense representations.** Unlike Frozen BP, a Backpack LM distilled with a frozen SmolLM backbone, ACROS preserves base-level FLORES PPL and LAMA@1 while improving lexical/sense diagnostics, showing that residual induction adds an interface without damaging the prediction path.

prediction:

$$\mathcal{L}_{\text{CLM}} = -\frac{1}{M}\sum_{t} m_t \log \text{softmax}(z_t^S)_{y_t}. \quad (6)$$

The diversity term discourages the K same-token sense vectors from collapsing into one shared direction:

$$\mathcal{L}_{\text{div}} = \frac{1}{M}\sum_{t} m_t \frac{1}{K(K-1)} \sum_{k\neq\ell} \left(\hat{E}_{k,t}^{\top}\hat{E}_{\ell,t}\right)^2, \quad (7)$$

where $\hat{E}_{k,t} = E_{k,t}/\|E_{k,t}\|_2$. We fix $\lambda_{\text{div}} = 0.005$ throughout; ablating this coefficient is left to future work. Since the base path is frozen and the gate starts at zero, induction only learns a residual sense decomposition that can help the existing prediction path.

Intrinsic Results. We apply ACROS to SmolLM2-360M with 10.5B tokens of FineWeb (Allal et al., 2025). Table 2 shows the payoff of the residual design. ACROS preserves the base model's perplexity and factual recall (LAMA@1 even slightly exceeds the original at 0.3175 vs. 0.3155) while exposing a sense interface with strong lexical content (WS-353 (Finkelstein et al., 2001) Spearman correlation 0.487). The converted Backpack, by contrast, retains some sense structure but at the cost of a $9.5\times$ increase in perplexity.

Qualitative Trace. Table 3 traces a single word — `article` — through all three roles that the induced interface supports. For measurement, cosine similarity between ACROS activation vectors matches the word in context to the correct WordNet gloss. For intervention, boosting sense vector 14 by $1.2\times$ shifts probability mass toward human-judged substitutes; Section 4.2 tests this systematically over 5,161 cases. For alignment, the same sense interface anchors SENSIA adaptation, and the originally English-only model generates coherent Indonesian text; Section 4.3 evaluates this across four languages. Table 4 gives a complementary view: the top sense slots for `article` cluster with distinct lexical neighborhoods (written-work, journalistic, legal, topical, and content-oriented readings), showing that the K slots capture different facets of the word.

Field	Example
Target word	`article`.
Task	Disambiguate the word.
Sentence	"…read the `article` by Caleb Nelson …"
Method	Cosine sim. of target and gloss ACROS activations.
Matched gloss	`article%1:10:00::` "nonfictional prose forming an independent part of a publication."
Task	Increase next-token odds of related substitutes.
Sentence	"I remembered an `article` in The Medusa …"
Method	Boost sense vector 14 by $1.2\times$.
Mass change	$\Delta = +4.36\text{e}{-7}$ (essay, piece, …).
Task	Convert eng model to `ind`.
Method	SENSIA.
Prompt	"Artikel itu menjelaskan bahwa"
Generation	"…pemerintah AS telah menyatakan bahwa mereka tidak akan menerima pembayaran dari pemerintah Suriah untuk senjata yang dibeli dari Rusia."

Table 3: **The resulting induced sense interface supports measurement, intervention, and cross-lingual adaptation.** (Top) The active sense vector successfully matches words in context to the correct WSD gloss. (Middle) Interventions on the selected vector increase likelihood for related substitutes. (Bottom) The induced senses act as anchors for cross-lingual adaptation via SENSIA, generating coherent Indonesian text.

4 Experiments

We evaluate ACROS through the three roles that motivate it: can the induced sense pathway be read as a semantic measurement, intervened on as a local causal handle, and reused as an alignment substrate for cross-lingual adaptation?

4.1 WSD: Senses as Semantic Measurements

Research Question. Do induced ACROS sense activations recover lexical sense distinctions without WordNet or SemCor supervision during induction?

Sense	Nearest words	Reading
14	article, articles, published	Published written piece. "Article" as a written work.
6	news, newspaper, press	News article. "Article" in a journalistic context.
21	ruling, prosecute, judiciary	Legal/institutional. "Article" in regulatory or legal language.
9	topic, discussion, interests	Topic or matter. "Article" as a subject of discussion.
25	substance, content, contribution	Content or substance. "Article" by what it conveys.

Table 4: **Different sense slots represent different facets of `article`.** Each slot retrieves a distinct filtered single-token neighborhood by cosine to the source-token sense vector, showing that ACROS decomposes the word across written-work, journalistic, legal, topical, and content-oriented readings.

Target context: . . . the bank approved the loan . . .
Candidate gloss prompt: `bank: a financial institution`
Score: cosine between target activation and gloss activation

Figure 4: **For measurement, ACROS matches contextual target activations to WordNet gloss activations.** Each candidate WordNet sense is encoded as a `lemma: gloss` prompt and compared against the contextual target activation in the same interface that later supports steering and alignment.

Methodology. We evaluate on Raganato ALL (Raganato et al., 2017), using WordNet only at evaluation time. ACROS scores each candidate through its induced interface: target occurrences and `lemma: gloss` probes are represented by L2-normalized K-dimensional sense-activation vectors (Figure 4), and we predict

$$\hat{s} = \arg\max_{s} \cos(a_{\text{ctx}}, a_{\text{gloss}}(s)).$$

All 7,253 instances are answered without back-off or coverage failures. The same-family control, Base SmolLM2, scores the same glosses by length-normalized conditional likelihood, separating ordinary gloss preference from induced sense matching.

Results. Table 5 shows ACROS performing competitive zero-shot WSD without sense supervision. ACROS reaches 64.95 F1 without any WordNet or SemCor supervision during induction, competitive with the WordNet first-sense heuristic (65.2) and 8.71 points above the same-family gloss-likelihood control (McNemar $p = 1.31\text{e}{-38}$; bootstrap 95% CI $[+7.39, +10.01]$). Supervised systems score higher as expected. Section 5 compares ACROS senses against dense hidden states.

Method	Supervision	F1
BEM	SemCor + WordNet	79.0
LMMS2048	SemCor	75.4
WordNet S1	WordNet heuristic	65.2
ACROS	zero-shot	64.95
Base SmolLM2	zero-shot gloss LM	56.24

Table 5: **ACROS learns strong sense representations that achieve competitive zero-shot WSD performance without WordNet or SemCor supervision.** WordNet S1 is a knowledge-based heuristic; BEM (Blevins and Zettlemoyer, 2020) and LMMS2048 (Loureiro et al., 2022) use sense-annotated resources and are included as supervised context. Base SmolLM2 is a same-family gloss-likelihood control.

4.2 CoInCo: Senses as Intervention Handles

Research Question. Can a selected ACROS sense contribution causally move a target word's substitute distribution while keeping the next-token distribution local?

Methodology. CoInCo (Kremer et al., 2014) provides lexical-substitution judgments for target words in context. After filtering for unambiguous anchors and one-token substitutes, we score 5,161 cases as next-token prompts (Figure 5). We measure the change in log weighted substitute mass, $m(z, Y) = \log \sum_{y \in Y} w_y \, p_z(y)$, where weights are proportional to annotator counts. Success means $\Delta = m(z_{\text{int}}, Y) - m(z_{\text{base}}, Y) > 0$. We use this rather than GAP or P@1 because those evaluate candidate rankings, while our experiment measures a distributional shift.

Intervention. The intervention is target-token-local. For target word at position j and answer position q, we boost one sense contribution analytically:

$$\Delta h_k = g(b-1)\, C_{k,q,j}\, E_{k,j},$$
$$z_{\text{int}} = z_{\text{base}} + \Delta h_k \, W^{\top}.$$

We compare five selectors: target-best (oracle with gold substitute access), self top-k (non-oracle proxy selecting from the model's own one-token top predictions, no gold access), and three non-target-aware controls (contribution-norm, random, norm). We report $D_{\text{KL}}(p_{\text{base}} \| p_{\text{int}})$.

Results. Table 6 shows that boosting the right ACROS sense increases human substitute mass. Target-best moves positive substitute mass in every case. The non-oracle self top-k proxy recovers substantial signal (Δ=+0.001109, 89.8% success) without access to gold annotations, while

Context: The reporter filed the `article` before noon.
Prompt: A context-appropriate substitute for `article` is
Human substitute multiset: story ×3, piece ×2, report, item
Intervention: boost one target-token sense at the answer position

Figure 5: **For intervention, ACROS tests whether a sense edit raises human substitute mass while preserving the base distribution.** We score interventions by how much they increase frequency-weighted probability mass on human CoInCo substitutes while keeping KL to the original next-token distribution low.

Selector	Delta	Succ. (%)	KL
target-best†	+1.64e−3	100.0	5.69e−7
self top-k	+1.11e−3	89.8	7.12e−7
contrib.-norm	+5.00e−6	49.6	8.24e−7
random	−8.00e−6	50.6	1.08e−7
norm	−3.70e−5	48.4	2.46e−7

Table 6: **Sense interventions successfully steer CoInCo substitutes, with self top-k recovering much of the oracle effect.** Delta is mean log substitute-mass change at boost 1.2 over 5,161 cases. †Oracle with gold access.

contribution-norm, random, and norm controls remain near chance, confirming that semantic alignment, not intervention magnitude, determines selector quality. We compare against dense hidden-state controls in Section 5.

4.3 SENSIA: Senses as Anchors for Cross-Lingual Alignment

Research Question. Can the induced sense pathway serve as an alignment anchor for cross-lingual adaptation?

Methodology. SENSIA (Cruz et al., 2026) tests whether the ACROS interface can serve as adaptation infrastructure. Starting from English ACROS, we adapt to Indonesian, Estonian, Swahili, and Turkish with the same English–target parallel pairs used in Cruz et al. (2026), comparing unadapted ACROS and Frozen BP repros. Intrinsic metrics are bidirectional contextual retrieval R@1, sense retrieval R@1 using SENSIA sense pooling, and target-language FLORES PPL.

We also run likelihood-based downstream tasks (XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2022), Belebele (Bandarkar et al., 2024)) and XL-Sum generation (Hasan et al., 2021) on full test sets, with Gemma 3 270M (Gemma Team et al., 2025) and Qwen2 0.5B (Yang et al., 2024) as dense multilingual baselines in the same parameter regime. Appendix D gives extra practical details.

Model	ctx@1	sns@1	PPL
Unadapted	.009 [.006,.011]	.022 [.018,.026]	107.7 [105.6,110.0]
Frozen BP	.693 [.680,.705]	.558 [.544,.571]	79.3 [77.7,80.9]
ACROS	.988 [.985,.991]	.987 [.984,.990]	7.94 [7.84,8.05]

Table 7: **For alignment, SENSIA uses the same induced sense path as a cross-lingual anchor.** After SENSIA adaptation from English to four languages, context- and sense-embedding based retrieval reach near-perfect R@1 on parallel FLORES pairs while target-language FLORES PPL drops sharply; intervals are language-balanced 95% bootstraps over sentence pairs.

Results. Table 7 shows strong SENSIA alignment from ACROS senses. Before adaptation, ACROS has essentially no cross-lingual signal (ctx@1 = 0.009). After adaptation with English–target parallel pairs from Cruz et al. (2026), retrieval jumps to 0.988 contextual and 0.987 sense R@1, while target-language perplexity drops to 7.94. The Frozen BP control improves over the unadapted baseline but its degraded LM substrate limits both retrieval and target-language modeling.

Table 8 shows downstream gains after SENSIA adaptation. ACROS improves over its base model on macro mean and on XCOPA and XStoryCloze, and is competitive with Gemma and Qwen despite starting as English-only. In generation, Gemma and Qwen score higher on lexical overlap but copy far more from the source (Copy-4g 0.784 and 0.411 versus 0.125 for ACROS). Appendices E, G, and H give per-language intervals, sample prompt continuations, and XL-Sum generations.

5 Ablations

5.1 Sense Interface vs. Dense Hidden States

Research Question. Does the explicit sense interface add value beyond the semantic information already present in the base model's dense hidden states?

Methodology. We use two paired controls. For WSD, dense hidden states from the same SmolLM2 backbone replace ACROS sense activations in the same context–gloss matching protocol. For steering, a generous dense target-best control searches for a hidden coordinate/sign that maximizes CoInCo substitute mass under a norm budget matched to the selected sense intervention.

	Likelihood acc.				XL-Sum gen.				
Model	XCOPA	XStory	Bele.	Mean	ROUGE-L	chrF++	Empty	Rep-3g	Copy-4g
Gemma 270M	0.548	0.527	0.255	0.426	**0.139**	**25.73**	0.050	**0.097**	0.784
Qwen2 0.5B	0.532	0.519	**0.299**	**0.436**	0.109	20.23	0.046	0.274	0.411
Base SmolLM2	0.520	0.492	0.239	0.402	0.088	15.30	0.008	0.359	0.368
Frozen BP	0.530	0.485	0.242	0.405	0.013	2.81	**0.000**	0.454	**0.000**
ACROS	**0.558**	**0.542**	0.231	0.424	0.110	18.30	**0.000**	0.283	0.125

Table 8: **ACROS sense representations support effective adaptation of an English LM to other languages via SENSIA.** Macro-averaged downstream and generation results. Higher is better except Empty, Rep-3g, Copy-4g. Best scores in bold.

Article: Misa penahbisan Paus Fransiskus digelar

Reference: Misa penahbisan Paus Fransiskus diadakan di Lapangan Santo Petrus, Roma, di hadapan ratusan ribu orang, termasuk pemimpin dunia dan pemuka agama.

Base SmolLM2:
`1.000.000 / Pesan: 1.000.000 / Pesan: 1.000.000 / Pesan: 1.000.000 / Pesan: ...`

R-L: 0.000 Copy-4g: 0.000 Rep-3g: 0.778

ACROS + SENSIA:
Paus Fransiskus menyambut warga dan memuliakan Santo Petrus. Paus Fransiskus menyambut para anak-anak di Lapangan Santo Petrus. Paus Fransiskus menyambut para pemimpin negara dan tokoh politik dari berbagai negara.

R-L: 0.308 Copy-4g: 0.107 Rep-3g: 0.172

Gemma 3 270M:
Paus Fransiskus menjalani serangkaian prosesi dalam penahbisannya. Paus meninggalkan kediaman sementaranya di Casa Santa Marta sebelum pukul 0900 waktu setempat. Sebelum menghadiri misa, Paus Fransiskus menyambut massa yang bersorak-sorai dan melambaikan bendera.

R-L: 0.133 Copy-4g: 1.000 Rep-3g: 0.000

Figure 6: **ACROS-induced senses provide strong alignment anchors, allowing target-language summarization after SENSIA adaptation.** Base SmolLM2 fails to generate coherent summaries before adaptation. ACROS+SENSIA produces a coherent target-language summary with low source copying (Copy-4g 0.107) after SENSIA adaptation. This contrasts with Gemma, which is pretrained with a multilingual mix but copies almost entirely from the source (Copy-4g 1.000).

Probe	Dense/base	ACROS sense
WSD F1	63.75	**64.95**
CoInCo delta	**+0.010240**	+0.001642
CoInCo KL	1.77e-5	**5.69e-7**

Table 9: **Dense hidden states already contain semantic signal, but ACROS makes that signal cleaner to access and reuse.** Dense target-best can push CoInCo substitute mass more strongly, but ACROS sense target-best achieves much lower KL and is semantically indexed, supporting the claim that ACROS changes access to meaning rather than inventing it from scratch.

Handle	Delta	KL	Largest content-word increases
ACROS sense 19	+0.003057	6.54e-7	expense +0.0028 pp; burden +0.0013 pp; cost +0.0008 pp
Dense coord. 87	+0.021610	2.92e-5	expense +0.0211 pp; burden +0.0118 pp; deficit +0.0058 pp

Table 10: **Dense hidden states can steer strongly, but ACROS senses steer more surgically with much lower KL.** On `coinco:7535` (target word `charge` with human substitutes `expenditure, fee, cost, expense, amount, fine, obligation`), dense target-best moves substitute mass at about $45\times$ the KL with larger collateral shifts than ACROS. "pp" denotes probability points.

Results. Pretrained LMs already encode substantial semantic structure, and our ablation confirms this: dense hidden-state gloss matching reaches 63.75 F1, only 1.20 below ACROS but reliably different (McNemar $p = 1.31\mathrm{e}{-9}$). What ACROS adds is *not* the semantic information itself, but *a cleaner way to access it*. On CoInCo, dense target-best pushes substitute mass about $6.2\times$ harder, but at about $31\times$ the KL. The single-row view in Table 10 makes the tradeoff vivid: for `charge`, dense steering moves the right words but also moves many wrong ones, while ACROS sense 19 produces a smaller, more surgical shift. The explicit sense interface provides lower-dimensional, semantically indexed handles; crucially, these same handles also support WSD and cross-lingual alignment, which raw hidden-state coordinates do not. A residual addition without sense decomposition could add capacity, but would not produce individually addressable sense axes: the per-axis structure is what makes gloss matching, single-sense steering, and SENSIA sense-level pooling possible through the same set of variables.

Appendix F reports additional ablations explor-

ing sense-slot capacity and backbone scale.

5.2 Architecture Portability

Research Question. Does ACROS induce a functional sense interface on decoder architectures beyond the Llama family?

Methodology. We apply the same induction objective and hyperparameters to two additional backbones that differ from SmolLM2 in normalization, activation, position encoding, and block structure: Pythia-410M (parallel attention + FFN, LayerNorm, RoPE, GELU) and OPT-350M (sequential blocks, LayerNorm, learned positional embeddings, ReLU). All three models are induced on the same FineWeb English data and evaluated on FLORES PPL preservation, tokenizer-specific LAMA@1, WSD, and CoInCo steering under both oracle and non-oracle selectors.

Results. Table 11 shows that ACROS induces functional sense interfaces on all three architectures. WSD is the cleanest result: Pythia essentially matches SmolLM2 (64.99 vs. 64.95 F1), and OPT is only half a point lower (64.46). Oracle steering reaches 100% success ($\Delta > 0$) on every backbone. The non-oracle self top-k proxy achieves 81–90% success rate ($\Delta > 0$), while contribution-norm, random, and norm controls remain near chance on all three. LM behavior is preserved throughout, with small directional gains in both PPL and LAMA@1 within every backbone. Steering magnitudes are smaller on Pythia and OPT than on SmolLM2, so we frame this as portability of the interface rather than identical strength across architectures, but the core finding is that the same induction objective, applied without per-backbone tuning, produces a working sense interface on three architecturally distinct decoders. Additional ablations for larger backbone sizes and diagnostic runs are reported in Appendix F.

Metric	SmolLM2	Pythia	OPT
FLORES PPL Δ	-0.05	-0.35	-0.62
LAMA@1 Δ	+.002	+.010	+.003
WSD F1	64.95	64.99	64.46
CoInCo oracle succ.	100%	100%	100%
CoInCo self top-k succ.	89.8%	81.3%	85.4%

Table 11: **ACROS induces functional sense interfaces across architecturally distinct decoders.** SmolLM2 (Llama-style), Pythia (parallel blocks, GELU), and OPT (learned positions, ReLU) all receive the same induction objective and hyperparameters. Δ rows show ACROS minus the original backbone (lower PPL and higher LAMA@1 are better). PPL and LAMA use each backbone's own tokenizer and are comparable within rows but not across them. Success is the fraction of $\sim$5,160 scored cases where $\Delta > 0$.

6 Related Work

Several recent methods add semantic structure to decoder models. Multi-Sense Embeddings (Wang et al., 2025), DSKD (Wang et al., 2026), CoCoMix (Tack et al., 2026), and codebook features (Tamkin et al., 2024) all build or inject explicit semantic variables during training, but none are designed as a single inference-time sense interface that supports measurement, steering, and alignment through the same variables. Probing work has long shown that pretrained LMs encode word-sense information internally; our Section 5 ablation confirms this and shows that ACROS surfaces it as a cleaner, lower-KL interface rather than creating it from scratch. Post-hoc steering via activation additions (Turner et al., 2024), representation engineering (Zou et al., 2025), or SAE features (Rimsky et al., 2024; Arad et al., 2025; Chalnev et al., 2024) can modify behavior but is often brittle (Wu et al., 2025; Da Silva et al., 2025). ACROS provides trained-in, token-local sense slots that give reliable low-KL handles for word-level steering.

7 Conclusion

We have shown that sense representations do not need to be baked into a model's architecture. ACROS induces an explicit, gated sense pathway into an ordinary pretrained decoder LM, and the same set of induced variables supports three capabilities that previously required separate systems: zero-shot word-sense disambiguation competitive with the WordNet first-sense heuristic, low-KL lexical steering where a simple non-oracle semantic proxy recovers about 90% of positive substitute-mass shifts across 5,161 cases, and SENSIA cross-lingual alignment that reaches near-perfect retrieval while reducing target-language perplexity by an order of magnitude over Backpack conversion. The residual design preserves the base model's prediction quality throughout; ACROS adds structure without taking anything away. These results establish that sense-level measurement, intervention, and alignment are *inducible* properties: they can be added to an existing LM rather than requiring a purpose-built architecture.

Limitations

Lexical steering scope. Steering is evaluated via a deliberate local intervention: a $1.2\times$ boost on a single sense at a single position, prioritizing low KL over large probability shifts. An oracle selector using gold substitute access establishes that useful causal directions exist; a non-oracle semantic proxy (self top-k) recovers 81–90% of oracle success without gold access, while magnitude-only controls remain near chance. A learned selector that replaces the proxy is left for future work.

WSD scope. WSD is evaluated zero-shot using the same induced variables that also support steering and alignment; no sense supervision is used during induction. ACROS reaches 64.95 F1 on Raganato ALL, competitive with the WordNet first-sense heuristic (65.2). Supervised systems score higher, as expected, but operate on dedicated architectures that do not simultaneously support steering and cross-lingual alignment.

Cross-lingual evaluation scope. Cross-lingual evaluation uses SENSIA (Cruz et al., 2026), a method architecturally compatible with ACROS, which benefits intrinsic retrieval. Downstream results show that sense-based alignment transfers effectively to lexical and short-context tasks (XCOPA, XStoryCloze); passage-level reasoning (Belebele) transfers less well, and natively multilingual pretraining remains better suited for that setting. Broader evaluation with other adaptation methods and languages is left for future work.

Architecture and scale. Three decoder families are tested (Llama, GPT-NeoX, OPT); encoder-decoders and models above 2B remain future work. Additional ablations for larger backbone sizes are available in Appendix F, characterizing how the current induction recipe behaves at scale.

English-only induction. Cross-lingual ability is scoped to post-hoc SENSIA adaptation from an English-only induced interface; multilingual induction is out of scope and left for future work.

Automatic evaluation only. All evaluations use automatic metrics; no formal human judgments are reported.

Ethical Considerations

This work uses public benchmark and training resources and does not introduce new human-subject data collection. The main ethical risk is inherited from the pretrained base models and datasets: ACROS can expose and steer lexical associations that may encode social biases, stereotypes, or uneven coverage across languages and communities.

The steering results should be read as diagnostic evidence about local lexical control, not as a deployment-ready editing system. Because low-KL interventions can subtly change next-token preferences, similar mechanisms could be misused for lexical framing or manipulation if paired with an improved selector. The self top-k proxy demonstrates that partial non-oracle selection is feasible, which makes the misuse risk more concrete than a purely oracle result; we note this as a reason for caution in downstream applications.

The cross-lingual adaptation experiments cover Indonesian, Estonian, Swahili, and Turkish, but the evaluations are automatic and limited. We do not claim that the adapted models are robust, culturally appropriate, or production-ready for speakers of those languages. Any release or downstream use should include language-community evaluation, bias testing, and clear disclosure that ACROS is a research interface-induction method.

AI use disclaimer. Anthropic's Claude was used to assist with proofreading and editing the final draft of this paper.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son Nguyen, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, and 3 others. 2025. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. 2025. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10241–10259, Suzhou, China. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *Preprint*, arXiv:2411.02193.

Jan Christian Blaise Cruz, David Ifeoluwa Adelani, and Alham Fikri Aji. 2026. Multilinguality as sense adaptation. *Preprint*, arXiv:2601.10310.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19856–19882, Vienna, Austria. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on the World Wide Web*, pages 406–414. ACM.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

John Hewitt, Sarah Chen, Lanruo Lora Xie, Edward Adams, Percy Liang, and Christopher D. Manning. 2024. Model editing with canonical examples. *Preprint*, arXiv:2402.06155.

John Hewitt, John Thickstun, Christopher Manning, and Percy Liang. 2023. Backpack language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9103–9125, Toronto, Canada. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *Preprint*, arXiv:1909.05858.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Loureiro, Alípio Mário Jorge, and Jose Camacho-Collados. 2022. Lmms reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artificial Intelligence*, 305:103661.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation:

A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason E Weston, and Xian Li. 2026. LLM pretraining with continuous concepts. In *The Fourteenth International Conference on Learning Representations*.

Alex Tamkin, Mohammad Taufeeque, and Noah Goodman. 2024. Codebook features: Sparse and discrete interpretability for neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47535–47563. PMLR.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Qitong Wang, Mohammed J Zaki, Georgios Kollias, and Vasileios Kalantzis. 2025. Multi-sense embeddings for language models and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13353–13369, Vienna, Austria. Association for Computational Linguistics.

Qitong Wang, Mohammed J. Zaki, Georgios Kollias, and Vasileios Kalantzis. 2026. Decoder-based sense knowledge distillation. *Preprint*, arXiv:2602.22351.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 67035–67080. PMLR.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Backpack Conversion Diagnostics

We evaluate whether an existing pretrained decoder can be converted into a Backpack-style LM without losing ordinary language-model quality. The source model is `HuggingFaceTB/SmolLM2-360M`, an English-only Llama-style decoder. We compare the original model against three Backpack conversion families:

- **Continued LM training:** initialize a Backpack-Llama model from the pretrained checkpoint and continue causal-LM training in the conversion-output architecture.
- **Distillation:** train the converted Backpack student with the pretrained SmolLM2 model as a frozen teacher, combining next-token training with teacher-student alignment.
- **Frozen-backbone distillation:** freeze the SmolLM2 backbone and LM head inside the Backpack student, and train only the sense-side modules. This tests whether the failure is ordinary backbone drift or the conversion-output family itself.

We use sentence-level FLORES English devtest perplexity and filtered LAMA factual cloze accuracy as compact LM-health diagnostics. Fixed generation smoke tests were also used to check whether fluent prediction behavior survived conversion.

The frozen-backbone controls are the key separation. If conversion failed only because the pretrained weights drifted during training, freezing the backbone and LM head should preserve base-model behavior. However, as the results show, it does not. Even with the original backbone and head frozen, the converted Backpack output remains much worse than the base LM. This supports the interpretation that the output family itself is the problem: the sense mixture is being asked to replace too much of the pretrained hidden state.

Variant	FLORES PPL	LAMA@1	LAMA@5	MRR
Original	25.1	0.3155	0.525	0.4161
Backpack + CPT	356.86	0.0695	0.146	0.1224
Backpack + Distil.	196.30	0.0420	0.135	0.0926
Backpack + Frozen Distil. K8	825.98	0.0520	0.140	0.1034
Backpack + Distil. + Freeze	236.94	0.1245	0.266	0.2004
ACROS	**25.09**	**0.3175**	**0.5245**	**0.4171**

Table 12: Conversion diagnostics. Backpack conversion substantially degrades ordinary LM behavior, including under frozen-backbone controls. The residual ACROS path preserves base-level LM health.

Quantity	Value
Tokens analyzed	10,000
Hidden dimension	960
K=8 cumulative variance	71.2%
Canonical-K cumulative variance	84.9%
K=64 cumulative variance	88.3%
Rank for 95% variance	287
Rank for 99% variance	697

Table 13: SVD diagnostic for centered SmolLM2-360M final-layer hidden states.

B SVD Bottleneck Diagnostic

The converted Backpack architecture routes prediction through a normalized sense mixture:

$$h_q^{\mathrm{BP}} = \sum_{k=1}^{K} \alpha_{q,k} v_{q,k}, \quad \alpha_{q,k} \geq 0, \quad \sum_k \alpha_{q,k} = 1. \tag{8}$$

For each position, this representation lies in the convex hull of the active sense vectors before the LM head. This is useful when the model is pretrained in that form, but it is restrictive when the mixture is asked to replace the full hidden state of an existing LM.

We estimate the size of the base hidden-state space by collecting 10,000 final-layer hidden states from the pretrained SmolLM2-360M model on English FineWeb tokens. We center the hidden-state matrix, compute its singular values, and measure cumulative variance explained by the top ranks. The result is not a formal impossibility proof, because Backpack sense vectors are input-dependent. It is a geometric warning: practical K values are far smaller than the rank needed to preserve most of the base hidden-state variance.

The practical implication is that a converted Backpack with small K must compress structure the base model uses for prediction. The canonical ACROS sense count captures only about 84.9% of the measured variance, while 95% variance requires rank 287 and 99% requires rank 697. This explains why conversion can expose a sense path but still fail as a language model, and motivates a residual design that augments the pretrained output path instead of replacing it.

C ACROS Architecture Details

This appendix expands the implementation details behind the residual architecture in Figure 3. The reported SmolLM2 ACROS checkpoint wraps the original `HuggingFaceTB/SmolLM2-360M` decoder without replacing its prediction path: the pretrained backbone and tied LM head remain frozen, and ACROS trains only the residual sense pathway and a scalar gate.

Sense vectors. For each token embedding, the sense network produces K vectors in the model hidden dimension. In the SmolLM2 K=32 run, token embeddings of dimension d=960 pass through a small residual MLP block, a LayerNorm, and a final MLP whose output dimension is Kd. The result is reshaped into source-token sense vectors

$$E \in \mathbf{R}^{B \times K \times T \times d}, \tag{9}$$

where B is batch size and T is sequence length. These vectors are not vocabulary entries or external dictionary senses; they are token-local residual directions learned from the induction objective.

Contextualization layer. The contextualization layer is a lightweight causal query/key attention module, not a full Transformer block. It takes the frozen backbone hidden states $B \in \mathbf{R}^{B \times T \times d}$ and applies one linear map to produce per-sense queries and keys:

$$Q, K \in \mathbf{R}^{B \times T \times K \times d/K}. \tag{10}$$

For each sense slot k, it computes causal attention weights over source positions:

$$C_{k,q,j} = \mathrm{softmax}_{j \leq q} \left(\frac{Q_{q,k}^{\top} K_{j,k}}{\sqrt{d/K}} \right). \tag{11}$$

Setting	Value
Base model	SmolLM2-360M
Sense slots	$K = 32$
Hidden size	960
Sense MLP scale	4
Frozen modules	Backbone and LM head
Trainable modules	Sense network, contextualization, gate
Induction data	FineWeb English
Sequence length	2048
GPUs	4 A100
Batch / GPU	4
Gradient accumulation	8
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	0.1
Warmup ratio	0.02
Max steps	40,000
Precision	bfloat16
α for CLM/KD mixture	0.5
KD temperature	2.0
λ_{div}	0.005
Gate initialization	0.0
Gate learning rate	main learning rate

Table 14: Main ACROS induction hyperparameters for the SmolLM2-360M K=32 checkpoint.

There is no learned value projection in this layer. The values are the sense vectors $E_{k,j}$ produced from the input embeddings. The residual mixture at query position q is therefore:

$$M_q = \sum_{k=1}^{K} \sum_{j \le q} C_{k,q,j} E_{k,j}. \tag{12}$$

Residual gate. The current model uses one learned global scalar gate g, shared across all layers, positions, tokens, and sense slots:

$$H_q = B_q + gM_q, \qquad z_q = H_q W^{\top}. \tag{13}$$

The gate is initialized to 0, so the initialized ACROS checkpoint exactly matches the base model before induction. The experiments in this paper keep this simple scalar-gate design throughout. Position-wise, token-wise, layer-wise, or sense-wise gates are natural future extensions, but they are not used in the reported results.

The main SmolLM2 K=32 induction run therefore sees approximately $40{,}000\times2048\times4\times4\times8 \approx 10.49$B training tokens, assuming fully packed sequences. No hyperparameter search was performed; all values in Table 14 are the fixed recipe used for the reported induction runs.

Latency. ACROS adds three operations to each full forward pass: the sense MLP, the contextualization attention, and the residual addition before the frozen LM head. The implementation used here prioritizes faithful access to the full sense contextualization tensor; generation currently disables key-value caching, so the numbers below should be read as full-sequence scoring latency rather than optimized autoregressive decoding latency. In a local CPU-only fp32 benchmark with batch size 1, ACROS adds about $1.3\times$–$1.5\times$ full-forward latency over the base SmolLM2 model; on CUDA bf16 the overhead drops to $1.03\times$–$1.06\times$, reflecting the sense path's low arithmetic intensity relative to the frozen backbone on GPU (Table 17).

Run	Wall h	A100 GPU-h
SmolLM2-360M K=32	32.5	129.9
SmolLM2-360M K=8	28.3	113.3
SmolLM2-1.7B K=32	81.2	325.0
Pythia-410M K=32	30.4	121.8
OPT-350M K=32	24.4	97.5
SENSIA adapters (4)	68.5	274.0
Reported-run subtotal	265.3	1061.5
Frozen BP K=32 control	31.8	127.1

Table 15: Approximate training compute for reported checkpoints, computed from W&B runtime artifacts multiplied by the allocated four-A100 recipes. Evaluation-only jobs and local qualitative generation are excluded.

Component	Parameters	Status
Base SmolLM2	361.8M	frozen
Sense network	129.1M	trained
Contextualization layer	1.8M	trained
Global gate	1	trained
Auxiliary projection	0.9M	frozen / unused
ACROS checkpoint total	493.7M	–
Active added path	130.9M	trained

Table 16: Parameter counts for the SmolLM2-360M K=32 ACROS checkpoint. The auxiliary projection is present in the implementation for optional reconstruction warmup, but the reported runs set the warmup to zero, so it is not active in training or inference.

D SENSIA Protocol Details

The SENSIA adaptation experiment follows the same intrinsic alignment logic as the original SENSIA setup (Cruz et al., 2026). We adapt from English into Indonesian, Estonian, Swahili, and Turkish, then average contextual retrieval, sense retrieval, and target-language FLORES PPL across languages.

Adaptation recipe. SENSIA adaptation starts from the English ACROS checkpoint and fine-tunes on English–target parallel data using a three-phase

Backend	Seq. len.	Base ms	ACROS ms	Ratio
CPU fp32	64	66.7	102.2	1.53×
CPU fp32	128	95.0	123.7	1.30×
CPU fp32	256	157.7	201.5	1.28×
CUDA bf16	64	23.8	24.7	1.04×
CUDA bf16	128	23.1	23.8	1.03×
CUDA bf16	256	22.7	24.0	1.06×

Table 17: Full-forward latency for SmolLM2-360M and ACROS K=32 at batch size 1. CPU fp32 timings are means over repeated forward passes on a local machine. CUDA bf16 timings use an NVIDIA RTX A5000 with 20 warmup and 100 timed forwards using CUDA event timing. Neither backend uses KV-cache; numbers reflect full-sequence scoring latency, not optimized autoregressive decoding.

Model	ROUGE-L	chrF++	Copy-4g
Base SmolLM2	0.088	15.30	0.368
ACROS	0.110	18.30	0.125
Frozen BP	0.013	2.81	0.000
Gemma 270M	0.139	25.73	0.784
Qwen2 0.5B	0.109	20.23	0.411

Table 18: XL-Sum full-test generation over available languages. ACROS improves over the pure English Base SmolLM2 starting point and Frozen BP, while copying much less than modern dense baselines.

scheduled objective combining context InfoNCE, sense InfoNCE, and target LM losses with label smoothing (ϵ=0.05). The alignment phase (first 20% of steps) emphasizes retrieval objectives; the middle phase (30%) balances retrieval and LM; the polish phase (final 50%) emphasizes target LM while freezing the sense network and sense weight net. All runs use AdamW with learning rate 5×10^{-5}, linear warmup over 2% of steps, global batch size 64 (16 per GPU, 4×A100), sequence length 256, and 150,000 total steps in bfloat16 mixed precision. Frozen BP uses the same data, objective schedule, step count, and batch recipe, but starts from legacy Backpack checkpoints and differs in parameter trainability and precision defaults; it is included as a degraded-substrate control, not as an identical ACROS variant.

Parallel data. We use the same English–target parallel pairs filtered from OPUS and made available by Cruz et al. (2026). Pairs are filtered by length, language identification, LaBSE similarity, and deduplication. Corpora drawn on include CCMatrix, CCAligned, ParaCrawl, JW300, OpenSubtitles, WikiMatrix, TED2020, GlobalVoices, and language-specific collections including EUbookshop and DGT for Estonian, SETimes for Turkish, and JW300 and Tanzil for Swahili.

Intrinsic retrieval and confidence intervals. Target-language PPL and retrieval both use the FLORES-200 `devtest` split (1,012 examples per language). Retrieval is bidirectional cosine nearest-neighbor R@1 with no re-ranking: source-to-target and target-to-source R@1 are averaged. Context embeddings are the last non-pad hidden state; sense embeddings use SENSIA sense pooling at temperature 0.7. Bootstrap CIs use 10,000 resamples (seed 20260514) over sentence pairs, language-balanced across the four target languages.

XL-Sum generation. Summaries are generated with greedy decoding, `max_input_tokens=512`, and `max_new_tokens=80`, using language-specific completion prompts (no chat template). Estonian is unavailable in XL-Sum; results cover Indonesian (4,780 test examples), Swahili (987), and Turkish (3,397). ACROS and Frozen BP decode without KV-cache due to implementation constraints; dense baselines use KV-cache when supported. ROUGE-L is `rouge-score 0.1.2` F1; chrF++ is `sacrebleu 2.6.0` with word order 2. Copy-4g is 4-gram precision of the generation against the source article; Rep-3g is the repeated 3-gram rate; Empty is the fraction of empty generations.

Downstream likelihood evaluation. XCOPA, XStoryCloze, and Belebele are evaluated zero-shot with the Eleuther LM Evaluation Harness in likelihood-ranking mode (not generation). XCOPA and XStoryCloze have 2 choices (50% chance); Belebele has 4 choices (25% chance). XStoryCloze covers Indonesian and Swahili only; Estonian and Turkish are unavailable. Per-cell intervals in Table 19 are $\mathrm{acc} \pm 1.96\times\mathrm{stderr}$, unpaired.

Baselines. Gemma 3 270M uses `google/gemma-3-270m` (base, not instruct) in bfloat16 with completion prompts. Qwen2 0.5B uses `Qwen/Qwen2-0.5B` (base, not instruct) with completion prompts. Both use KV-cache for generation.

E Downstream Standard-Error Intervals

Table 19 gives approximate 95% intervals ($\mathrm{acc} \pm 1.96\times\mathrm{stderr}$) from the Eleuther evaluation harness for each task–language cell. These are not paired bootstrap tests and should not be used for strong significance claims; they are included so that readers can judge the overlap pattern referenced in the

Task	Lang	Base	Frozen BP	ACROS	Gemma	Qwen
XCOPA	ind	.520 ±.044	.514 ±.044	.594 ±.043	.580 ±.043	.558 ±.044
XCOPA	est	.488 ±.044	.518 ±.044	.546 ±.044	.496 ±.044	.486 ±.044
XCOPA	swh	.534 ±.044	.558 ±.044	.550 ±.044	.544 ±.044	.544 ±.044
XCOPA	tur	.538 ±.044	.530 ±.044	.540 ±.044	.570 ±.043	.538 ±.044
Belebele	ind	.281 ±.029	.230 ±.028	.223 ±.027	.259 ±.029	.324 ±.031
Belebele	est	.220 ±.027	.256 ±.029	.229 ±.027	.259 ±.029	.287 ±.030
Belebele	swh	.220 ±.027	.271 ±.029	.237 ±.028	.229 ±.027	.274 ±.029
Belebele	tur	.236 ±.028	.209 ±.027	.234 ±.028	.272 ±.029	.309 ±.030
XStory	ind	.496 ±.025	.484 ±.025	.554 ±.025	.547 ±.025	.543 ±.025
XStory	swh	.488 ±.025	.485 ±.025	.529 ±.025	.507 ±.025	.494 ±.025

Table 19: Downstream full-dataset results with approximate 95% intervals. Intervals are unpaired standard-error estimates for each task–language cell, included to show uncertainty and overlap across ACROS and baseline models.

Metric	$K = 8$	$K = 32$
WSD F1	64.91	64.95
CoInCo delta	**+0.002018**	+0.001642
CoInCo KL	1.13e-6	**5.69e-7**

Table 20: K capacity ablation. Current probes do not establish a monotonic "more senses is better" trend.

Metric	360M	1.7B
FLORES PPL	25.09	**20.41**
LAMA @1/@5	0.3175/0.5245	**0.3550/0.5770**
Gate value	-0.05090	-0.01813
Contrib. ratio	1.745%	1.548%
WS-353 mean/max	**0.449/0.487**	0.270/0.328
WSD F1	**64.95**	64.18
CoInCo delta	**+0.001642**	+0.000893

Table 21: Scale and diversity ablation. The 1.7B model improves ordinary LM health but not the current sense-interface metrics; a tenfold increase in λ_{div} raises slot orthogonality but does not recover WS-353 or WSD, ruling out sense collapse as the bottleneck.

main text.

F Further Architectural Ablations

The following ablations characterize boundary conditions of the ACROS induction recipe along two dimensions: sense-slot capacity and backbone scale. They are not central to the main claims of the paper. All checkpoints are single runs under the fixed induction recipe; repeated-trial variance estimates are not reported for these rows.

F.1 Sense-Slot Capacity

Research Question. Does increasing the number of sense slots K monotonically improve sense-interface quality under fixed induction?

Methodology. We train matched single-run 360M ACROS checkpoints with $K = 8$ and $K = 32$ and evaluate with the same WSD and CoInCo protocols (Table 20).

Results. WSD is effectively tied across both settings (64.91 vs. 64.95 F1). CoInCo shows a trade-off: K=8 yields larger substitute-mass delta while K=32 yields lower KL. Increasing K changes the granularity of the interface rather than monotonically improving it under the current probes.

F.2 Backbone Scale and Induction Pressure

Research Question. Does the residual ACROS recipe improve when the backbone is scaled from SmolLM2-360M to SmolLM2-1.7B? And does increasing diversity pressure explain any observed degradation?

Methodology. We compare single-run 360M and 1.7B K=32 ACROS checkpoints on LM health (English FLORES PPL, LAMA) and sense-interface quality (gate, contribution ratio, WS-353, WSD, and CoInCo target-best steering). To diagnose the source of any degradation, we additionally train a matched 1.7B checkpoint with λ_{div} increased tenfold (0.005→0.05), holding all other hyperparameters fixed.

Results. The 1.7B backbone is a better language model (FLORES PPL drops from 25.09 to 20.41; LAMA top-1/top-5 rise to 0.3550/0.5770), but the sense interface does not proportionally improve: WS-353, WSD, and CoInCo all weaken relative to 360M. To diagnose the cause, we ran a matched 1.7B checkpoint with λ_{div} increased tenfold (0.005→0.05). Slot orthogonality and contribution ratio both increase under stronger diversity pressure, confirming that the sense vectors are not collapsing. However, WS-353 max and WSD do not recover, which rules out sense collapse as the primary bottleneck. The limiting factor appears to be induction signal strength: a more capable backbone satisfies the objective with less residual contribution, leaving the gate less open regardless

of slot structure. This suggests a concrete design direction: objectives that specifically reward gate activation, such as sense-reconstruction losses or gate-magnitude terms, are likely to extend the current recipe to larger backbones. Systematic exploration of such objectives is out of scope for the current paper and is left for future work.

G Prompt Continuation Samples

Tables 22–23 show illustrative continuations from the English ACROS checkpoint and the four SENSIA-adapted checkpoints. Prompts are translated or lightly adapted across languages. Outputs were generated with greedy decoding and top-p sampling, then lightly trimmed to remove terminal fragments and repeated punctuation.

H XL-Sum Qualitative Summaries

Table 24 shows XL-Sum examples where ACROS copies less source text than Gemma and Qwen. The examples are illustrative; the aggregate copy-rate comparison is reported in Table 8.

I Artifact Provenance and Licensing

Table 25 summarizes the external artifacts used in the experiments. License entries follow the public model cards, dataset cards, or project pages available at the time of writing; for collections such as OPUS, the applicable terms vary by component corpus. We use these resources for research evaluation and adaptation only, and we do not redistribute the raw corpora with this paper.

Software. XL-Sum generation metrics use `rouge-score` 0.1.2 for ROUGE-L and `sacrebleu` 2.6.0 with chrF word order 2 for chrF++. XCOPA, XStoryCloze, and Belebele likelihood evaluations use the Eleuther LM Evaluation Harness 0.4.13.dev0. Bootstrap aggregation and summary statistics use NumPy 2.4.4 and SciPy 1.17.1; the McNemar exact test used for WSD comparisons is implemented directly in the local evaluator. Local reproduction used PyTorch 2.11.0, Transformers 5.5.3, and Datasets 2.21.0.

Data sensitivity. Several artifacts contain web or news text and may include naturally occurring names, public events, and demographic references. We did not collect new human-subject data, attempt re-identification, or augment the benchmark corpora with private information. All artifacts were used for research purposes only, consistent with their access conditions, and no derived artifacts are used outside research contexts.

Prompt	Greedy	Sampled
The president of the country announced that	The president of the country announced that the country will be free from the debt crisis.	The president of the country announced that there will be a referendum on the constitution. If it does not happen by March 19, the country will hold a national election on April 19.
The hospital reported that the number of patients	The hospital reported that the number of patients with the disease has increased by 100% in the past year.	The hospital reported that the number of patients with hepatitis C has decreased by 60% in the past two years due to the drug, but 20% of the people who receive hepatitis C treatments develop drug resistance.
Researchers at the university discovered a new	Researchers at the university discovered a new way to make a material that can be used to make a variety of different devices, including a new type of solar cell.	Researchers at the university discovered a new technique that can convert water to hydrogen gas. This breakthrough, if perfected, will make it much easier to store and transport hydrogen fuel.

Table 22: English ACROS continuation samples.

Lang.	Prompt	Greedy	Sampled
est	Riigi president teatas, et *(EN: The president of the country announced that)*	Riigi president teatas, et ta ei ole valmis presidendiks saamiseks. *(EN: The president of the country announced that they are not ready to become president.)*	Riigi president teatas, et tahab 2020. aastal kinnitada seadusi, mis reguleerivad poliitikat. *(EN: The president of the country announced that they want to approve laws in 2020 that regulate policy.)*
est	Haigla teatas, et patsientide arv *(EN: The hospital reported that the number of patients)*	Haigla teatas, et patsientide arv on 1000. *(EN: The hospital reported that the number of patients is 1000.)*	Haigla teatas, et patsientide arv oli umbes 3000, kellest enamus olid pärit Euroopa Liidu osariikidest. *(EN: The hospital reported that the number of patients was about 3000, most from European Union member states.)*
est	Ülikooli teadlased avastasid uue *(EN: Researchers at the university discovered a new)*	Ülikooli teadlased avastasid uue koloonia, mis on kõige suurem ja kõige väiksem kogu maailmas. *(EN: Researchers at the university discovered a new colony that is the largest and smallest in the world.)*	Ülikooli teadlased avastasid uue tüüpi muutujad, mis mõjutavad looduslikku soojuskiirgust. *(EN: Researchers at the university discovered new types of variables that affect natural thermal radiation.)*
ind	Presiden negara itu mengumumkan bahwa *(EN: The president of the country announced that)*	Presiden negara itu mengumumkan bahwa ia akan menjadi presiden pertama yang menjabat dari 2019 hingga 2024. *(EN: The president of the country announced that he would be the first president to serve from 2019 to 2024.)*	Presiden negara itu mengumumkan bahwa keputusan itu telah dirumuskan oleh Komite Eropa untuk Kepentingan Urusan Timur Tengah. *(EN: The president announced that the decision had been formulated by the European Committee for Middle Eastern Affairs.)*
ind	Rumah sakit melaporkan bahwa jumlah pasien *(EN: The hospital reported that the number of patients)*	Rumah sakit melaporkan bahwa jumlah pasien yang terinfeksi meningkat dari 100 orang pada 2015 menjadi 1.000 orang. *(EN: The hospital reported that infected patients increased from 100 people in 2015 to 1,000.)*	Rumah sakit melaporkan bahwa jumlah pasien meningkat 15% selama periode 10 hari sebelum dan sesudah penyelidikan. *(EN: The hospital reported that the number of patients increased by 15% during the 10-day period before and after the investigation.)*
ind	Para peneliti di universitas menemukan sebuah *(EN: Researchers at the university discovered a)*	Para peneliti di universitas menemukan sebuah komunitas di mana 100% dari mereka yang terinfeksi virus itu mengalami komplikasi. *(EN: Researchers found a community where 100% of those infected with the virus experienced complications.)*	Para peneliti di universitas menemukan sebuah sistem kekebalan tubuh yang kuat yang dapat melawan virus, bakteri, dan parasit. *(EN: Researchers discovered a strong immune system that can fight viruses, bacteria, and parasites.)*
swh	Rais wa nchi alitangaza kwamba *(EN: The president of the country announced that)*	Rais wa nchi alitangaza kwamba, katika kipindi cha miezi 12 ijayo, nchi itaendelea kuwa na uchumi wa kutosha. *(EN: The president announced that, over the next 12 months, the country would continue to have a sufficient economy.)*	Rais wa nchi alitangaza kwamba amewafukuza viongozi wa makundi yenye silaha wakiwa katika kambi za jeshi. *(EN: The president announced that he had expelled leaders of armed groups while they were in army camps.)*
swh	Hospitali iliripoti kwamba idadi ya wagonjwa *(EN: The hospital reported that the number of patients)*	Hospitali iliripoti kwamba idadi ya wagonjwa wa Ebola nchini DRC imepungua kwa asilimia 90 tangu mwezi Aprili mwaka huu. *(EN: The hospital reported that the number of Ebola patients in the DRC has decreased by 90 percent since April this year.)*	Hospitali iliripoti kwamba idadi ya wagonjwa wanaopata ujauzito ni sawa na ilivyokuwa wakati wa mwaka 2016. *(EN: The hospital reported that the number of patients becoming pregnant is the same as it was in 2016.)*
swh	Watafiti katika chuo kikuu waligundua aina mpya ya *(EN: Researchers at the university discovered a new type of)*	Watafiti katika chuo kikuu waligundua aina mpya ya virusi vya corona, ambavyo vilikuwa vimeenea kwa kasi katika miji mingi. *(EN: Researchers discovered a new type of coronavirus that had spread rapidly in many cities.)*	Watafiti katika chuo kikuu waligundua aina mpya ya virusi vya homa ya nguruwe inayoitwa canine parainfluenza. *(EN: Researchers discovered a new type of swine flu virus called canine parainfluenza.)*
tur	Ülkenin cumhurbaşkanı şunu duyurdu: *(EN: The president of the country announced:)*	Ülkenin cumhurbaşkanı şunu duyurdu: "Bizim için, ülkemizin kültürü, ekonomisi ve toplumu. . . " *(EN: The country's president announced: "For us, our country's culture, economy, and society. . . ")*	Ülkenin cumhurbaşkanı şunu duyurdu: "Yeni bir anayasa kabul edildi, eğer iktidarı yeniden kurmaya çalışırsak, 2013'te başlayacağız." *(EN: The country's president announced: "A new constitution has been adopted; if we try to reestablish power, we will start in 2013.")*
tur	Hastane, hasta sayısının *(EN: The hospital, the number of patients)*	Hastane, hasta sayısının 1000'e ulaştığını gösterdi. Ve hastanede 1000 kişi var. *(EN: The hospital showed that the number of patients reached 1000. And there are 1000 people in the hospital.)*	Hastane, hasta sayısının sürekli artmasıyla, 12 hasta tespit edildi. *(EN: As the hospital's number of patients continued to increase, 12 patients were detected.)*
tur	Üniversitedeki araştırmacılar yeni bir keşif yaptı: *(EN: Researchers at the university made a new discovery:)*	Üniversitedeki araştırmacılar yeni bir keşif yaptı: Dünya'nın yüzeyinde bulunan 100 milyon yıllık bir kaya oluşumu. *(EN: Researchers made a new discovery: a 100-million-year-old rock formation found on Earth's surface.)*	Üniversitedeki araştırmacılar yeni bir keşif yaptı: Bu nesnelerin insan yapımı olduğunu keşfettiler. *(EN: Researchers made a new discovery: they discovered that these objects were human-made.)*

Table 23: SENSIA-adapted ACROS continuation samples for the four downstream languages.

Example	Reference	Model output	Metrics
Indonesian: Kuala Lumpur grenade blast	Satu orang tewas dan 12 orang lainnya terluka dalam ledakan di distrik hiburan populer di ibukota Malaysia, Kuala Lumpur. *(EN: One person was killed and 12 others were injured in an explosion in a popular entertainment district in Malaysia's capital, Kuala Lumpur.)*	**ACROS:** Penyelidikan polisi menunjukkan bahwa insiden terjadi di kawasan Bukit Bintang, Malaysia. Selain itu, polisi juga menemukan bahan peledak di mobil yang sedang diparkir di luar sebuah toko di luar kawasan Bukit Bintang. *(EN: Police say the incident occurred in Bukit Bintang and that explosives were found in a parked car nearby.)*	R-L=.078; chrF=22.3; Copy=.069; Rep=.033
		Gemma: Polisi Malaysia melaporkan bahwa sebuah granat diduga dilemparkan ke jalan di luar bar di kawasan Bukit Bintang pada Kamis pagi. Satu bahan peledak lain ditemukan dibawah sebuah mobil yang sedang diparkir tetapi tidak meledak. *(EN: Malaysian police report that a grenade was thrown outside a bar and another explosive was found under a parked car.)*	R-L=.108; chrF=21.6; Copy=.904; Rep=.000
		Qwen: Ledakan terjadi pada Kamis pagi di Bukit Bintang, Malaysia; granat diduga dilemparkan ke jalan di luar bar, dan satu bahan peledak lain ditemukan di mobil yang sedang diparkir. *(EN: An explosion occurred in Bukit Bintang; a grenade was thrown outside a bar and another explosive was found in a parked car.)*	R-L=.143; chrF=24.6; Copy=.559; Rep=.057
Indonesian: Cambridge written tests	Para calon mahasiswa Universitas Cambridge, Inggris, diwajibkan mengikuti ujian tertulis, yang merupakan komponen baru dari proses seleksi, kata direktur pendaftaran perguruan tinggi tersebut, Dr Sam Lucy. *(EN: Cambridge applicants in England must take a written exam as a new component of admissions, said admissions director Dr Sam Lucy.)*	**ACROS:** Universitas Cambridge mengatakan tes tertulis akan dilaksanakan bersamaan dengan wawancara atau sebelumnya. Lucy mengatakan ujian tertulis akan disesuaikan dengan setiap jurusan dan akan dilaksanakan pada November tahun ini. *(EN: Cambridge says written tests will be held with or before interviews and tailored to each subject.)*	R-L=.148; chrF=28.5; Copy=.320; Rep=.000
		Gemma: Universitas Cambridge mengatakan tes tertulis dilaksanakan bersamaan dengan wawancara atau sebelumnya. Rincian tentang ujian tertulis ini dijelaskan Dr Lucy dalam surat yang ia kirim kepada semua sekolah menengah. *(EN: Cambridge says written tests will be held with or before interviews, with details sent by Dr Lucy to secondary schools.)*	R-L=.171; chrF=34.6; Copy=.981; Rep=.000
		Qwen: Universitas Cambridge mengatakan tes tertulis dilaksanakan bersamaan dengan wawancara atau sebelumnya. Ujian tertulis akan disesuaikan untuk tiap jurusan dan akan dilaksanakan ketika calon mahasiswa mengikuti wawancara. *(EN: Cambridge says written tests will be customized by subject and held when applicants attend interviews or before them.)*	R-L=.197; chrF=35.2; Copy=.875; Rep=.000
Swahili: Somalia and al-Shabab	Rais wa Somalia Mohamed Abdullahi "Farmajo" ametangaza vita dhidi ya kundi la wanamgambo la al-Shabab. *(EN: Somali President Mohamed Abdullahi "Farmajo" declared war against the al-Shabab militant group.)*	**ACROS:** Rais wa Somalia Mohamed Abdullahi "Farmajo" ametangaza vita dhidi ya kundi la al-Shabab. Ametoa msamaha wa siku 60 kwa wanamgambo hao, akiwaapisha kwa kuwa ni msamaha wa kimataifa. Rais ametangaza pia *(EN: Farmajo declared war against al-Shabab and gave militants a 60-day amnesty; the output then trails off.)*	R-L=.583; chrF=78.2; Copy=.448; Rep=.000
		Gemma: Rais wa Somalia Mohamed Abdullahi "Farmajo" ametangaza vita dhidi ya kundi wa wanamgambo la al-Shabab. Ametoa pia msamaha wa siku 60 kwa wanamgambo hao, akiwataka kujisalimisha, ili wapate mafunzo, ajira na elimu. Hi ni baada ya shambulizi la *(EN: Farmajo declared war against al-Shabab, then closely follows the article's amnesty and surrender details before cutting off.)*	R-L=.545; chrF=74.6; Copy=1.000; Rep=.000
		Qwen: "Makala: Rais wa Somalia Mohamed Abdullahi "Farmajo" ametangaza vita dhidi ya kundi wa wanamgambo la al-Shabab. Ametoa pia msamaha wa siku 60 kwa wanamgambo hao, akiwataka kujisalimisha, ili wapate mafunzo, ajira *(EN: The output echoes the article prompt, then repeats the article's opening al-Shabab and amnesty details.)*	R-L=.625; chrF=77.6; Copy=.966; Rep=.000
Turkish: Syria Geneva talks	ABD Dışişleri Bakanı John Kerry, İsviçre'nin Cenevre kentinde 22 Ocak'ta yapılacak barış görüşmelerinin Suriye'de bir geçiş hükümeti kurulması için "en iyi fırsat" olduğunu söyledi. *(EN: U.S. Secretary of State John Kerry said the January 22 Geneva peace talks were the best opportunity to establish a transitional government in Syria.)*	**ACROS:** Suriye krizinin Çözüm yolu engillerle dolu olduğunu söylüyor, ABD Dışişleri Bakanı Kerry, Cenevre görüşmelerinin Suriye krizinin Çözümüne gözlerini aç *(EN: Kerry says the path toward resolving the Syria crisis is full of obstacles and connects the Geneva talks to a possible solution.)*	R-L=.310; chrF=35.2; Copy=.375; Rep=.000
		Gemma: Suriye'de yaşanan acılara son vermek için bu fırsatı kaçırmak affedilmez olur. ABD Dışişleri Bakanı, Cenevre görüşmelerinin Suriye krizinin çözümüne yardımcı olabileceğini söyledi. ABD Dışişleri Bakanı, Cenevre görüşmelerinin Suriye krizinin çözüm *(EN: The output says the chance to end Syria's suffering should not be missed, then repeats that Geneva could help solve the crisis.)*	R-L=.250; chrF=31.7; Copy=.455; Rep=.206
		Qwen: Birleşmiş Milletler, Suriye hükümeti ile muhalefetin ilk kez birlikte katılacağını duyurmasından hemen sonra yaptı. Birleşmiş Milletler, ABD ve Rusya, iki tarafın soruna siyasi bir çözüm bulması için aylardır çaba harcıyor. Fakat, İsvi *(EN: The output copies the setup about the U.N., Syria's government and opposition, and U.S.-Russian diplomatic efforts before cutting off.)*	R-L=.140; chrF=23.9; Copy=.897; Rep=.000

Table 24: Illustrative XL-Sum full-test examples. References are the original XL-Sum summaries; italic *EN* lines are author translations for readability. Output excerpts are lightly shortened and normalized for display. Metrics are ROUGE-L F1, chrF++, source-copy 4-gram precision, and repeated 3-gram rate for each full generated summary.

Artifact	Type	License / access terms	Use in this paper
FineWeb	Web text corpus	ODC-By	English induction data for ACROS; also used for the hidden-state SVD sample.
SmolLM2-360M	Base LM	Apache-2.0	Main frozen backbone and teacher for ACROS induction; Base SmolLM2 comparison row.
SmolLM2-1.7B	Base LM	Apache-2.0	Scale ablation backbone.
Pythia-410M	Base LM	Apache-2.0	Architecture-portability backbone.
OPT-350M	Base LM	OPT license / model-card terms	Architecture-portability backbone.
Raganato ALL	WSD benchmark	Component Senseval/SemEval and WordNet access terms	Zero-shot word-sense disambiguation evaluation.
CoInCo	Lexical-substitution corpus	CC-BY-3.0-US	Lexical steering evaluation over filtered public test cases.
FLORES-101/200	Multilingual sentence benchmark	CC-BY-SA 4.0	English perplexity preservation; SENSIA retrieval and target-language perplexity.
OPUS parallel data	Parallel corpus collection	Mixed per-corpus licenses	English–target parallel pairs for SENSIA adaptation.
XCOPA	Commonsense benchmark	CC-BY 4.0	Downstream likelihood evaluation.
XStoryCloze	Commonsense benchmark	CC-BY-SA 4.0	Downstream likelihood evaluation.
Belebele	Reading-comprehension benchmark	CC-BY-SA 4.0	Downstream likelihood evaluation.
XL-Sum	Summarization benchmark	CC-BY-NC-SA 4.0, research-only	Full-test generation evaluation and qualitative summary examples.

Table 25: Artifact provenance for external datasets and pretrained models. “License / access terms” is stated at the artifact level; collection resources may inherit additional restrictions from their component sources.