AUDITING CROSS-LINGUAL CHAIN-OF-THOUGHT PROBES WITH LEAKAGE CONTROLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) monitors read a model's written reasoning to catch misbehavior. Outside English, that text often fails: the model can follow a planted wrong hint and never mention it. A common fix is to look inside the model instead. Train a linear probe on English hidden states, then reuse it in other languages, assuming the internal "about to follow the hint" state is shared. We test that idea on a misleading-hint task in eight languages and four scripts, using three open models: Qwen3-8B (thinking), plus Apertus-8B and EuroLLM-9B (instruct). We also extend the prediction check to a same-content Azerbaijani Latin/Perso-Arabic script swap and to a wider set of thinking models across size and family (Qwen3 1.7B/32B, R1-Distill-Llama-8B, Phi-4 Reasoning). After holding out question IDs, the English probe still predicts hint-following at *post-CoT* (mean non-English AUROC 0.69–0.79), and removing a linear language signal barely changes transfer. But much of what it reads is already written: filtering early answer letters leaves Qwen3-8B near 0.61, while some instruct-model scores fall to chance; cutting the end of the reasoning drops all three late-probe means to 0.52–0.56. And the same English directions do not control the behavior. Adding or subtracting them during generation is no better than a same-size random direction in English, Turkish, Persian, or Swahili; stronger and multi-layer injections also fail to give reliable signed effects. So an English probe can monitor hint-following across languages, but the directions we test cannot steer it.

1 INTRODUCTION

Suppose a model is given a four-option question, plus one extra line: "a colleague suggests the answer is B," where B is wrong. The model often takes the hint, writes a rationale that never mentions it, and outputs B. A safety system reading only the rationale may miss the hint's influence (Turpin et al., 2023; Chen et al., 2025; Arcuschin et al., 2025). This is the standard stress test for *chain-of-thought (CoT) monitoring*; using the model's written reasoning to catch misbehavior before it becomes an action (Baker et al., 2025). The problem is worse outside English. Onyame et al. (2026) find near-total hint-driven deception in 13 languages and call for a *white-box* fix which looks inside the model, not only at generated text.

One proposed white-box method is a *linear probe*: a classifier on the transformer's hidden state (the residual stream at a chosen layer and token). In English, probes detect hint-driven motivated reasoning before and after CoT (Mirtaheri & Belkin, 2026). On other tasks, Cox et al. (2026) decode final answers before reasoning and steer those answers. The idea is then cross-lingual: train the probe in English, where labels are cheap, then reuse it in other languages, assuming that the "about to follow the hint" state belongs to the model rather than to a particular language.

We observe that this idea is easy to overstate for three reasons unrelated to shared cognition. *Item leakage:* multilingual benchmarks such as Global-MMLU (Singh et al., 2025) are parallel, and the same question has the same ID in every language. Therefore, a probe can memorize which *questions* are hint-prone and "transfer" by recognizing the question. *Language identity:* hidden states identify the input language almost perfectly. A probe may rely on language-specific features rather than a shared hint-following signal, although language-level hint rates alone cannot raise AUROC measured within each language. *Text leakage:* once the CoT has mentioned "so the answer is B," a probe reading the residual at that point is reading the written text, and a text classifier would do the same.

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.

None of these three cases is evidence by itself that the model carries a hidden cross-lingual decision state. We study each concern and apply the protocol to three open models of different training regimes: Qwen3-8B in thinking mode (reasoning written inside a separate `<think>` block; Yang et al., 2025), and two instruct models, Apertus-8B (Hernández-Cano et al., 2026) and EuroLLM-9B (Martins et al., 2025). We also test a wider range of thinking models across sizes and families (Qwen3 1.7B/32B, R1-Distill-Llama-8B, Phi-4 Reasoning) to check the scope of probe transfer. The task is the misleading hint above, on factual multiple-choice items in eight languages (Latin, Cyrillic, Arabic, and Han scripts) plus a same-content Azerbaijani Latin/Perso-Arabic script swap. We read the hidden state at three moments: the last prompt token (*prompt*); the mean of the first 16 generated tokens (*early-CoT*); and a late token (*post-CoT*): just before the final answer for Qwen3-8B, but the last generated token for the instruct models. Throughout, the score is AUROC, the probability that a hint-following sample ranks above a non-following one (0.5 is chance). Our findings are as follows:

1. **An English probe can detect hint-following in other languages.** Sharing the same questions across languages can fake transfer (up to $+0.16$ AUROC). With new questions held out, the English probe still works late in the response in all seven non-English languages (mean AUROC 0.69, 0.79, and 0.75 for Qwen3-8B, Apertus-8B, and EuroLLM-9B). Removing a linear language signal barely changes those scores in five of six model–position pairs (for Qwen3-8B at *post-CoT*, erasure is incomplete), so the probe is not just reading "this is Persian".
2. **A high probe score often means reading text that is already written.** For the thinking model, an early signal mostly survives after we drop responses that already contain the answer letter. For the instruct models, that filter often weakens or removes the early signal: they frequently write the answer early. Late in the response, cutting the end of the reasoning collapses the probe for all three models, and a text-only judge beats the probe on average for two of them. So a strong AUROC alone does not show a hidden decision before the text.
3. **Being able to read the behavior does not mean we can control it.** Pushing the English direction during generation changes hint-following no more consistently than a random nudge of the same size, in English as well as in other languages. Stronger pushes, the probe's own weight vector, and multi-layer injection also fail to give reliable control.

In short, an English-trained probe is a useful cross-lingual *monitor* of hint-following, but the directions we test are not a reliable *control* over it. The prediction check also extends to an Azerbaijani script swap and to additional thinking models; the full text and steering audit covers the three main models.

2 Related Work

CoT monitoring and its fragility. A CoT monitor reads the model's written reasoning and flags misbehavior (Baker et al., 2025). This requires informative text: optimization pressure can make it less so (Korbak et al., 2025), though some tasks still force useful CoT (Emmons et al., 2025). Planted hints are the standard stress test, as the model follows the hint and rationalizes without mentioning it (Turpin et al., 2023; Chen et al., 2025; Arcuschin et al., 2025; Sharma et al., 2024). Onyame et al. (2026) show this is worse in 13 languages; Zhao et al. (2026) show that forcing the reasoning language changes compliance. We take that multilingual setting, and the proposed look-inside remedy, as the aim of this study.

Probes for pre-committed answers. Mirtaheri & Belkin (2026) train residual-stream probes to detect hint-driven motivated reasoning before and after CoT. Cox et al. (2026) go further in English and show that adding the probe direction during generation flips answers more than an orthogonal baseline. Probes have also been used for deception (Goldowsky-Dill et al., 2025) and truth (Marks & Tegmark, 2024; Burns et al., 2023). All of this is English-only. We run a related causal check, fitting the direction in English and testing it both there and in other languages.

Probing methodology. A probe can look successful for the wrong reason (Alain & Bengio, 2016; Belinkov, 2022), which is why control tasks exist (Hewitt & Liang, 2019). We add two multilingual ones: do not train and test on the same questions, and remove the linear subspace that identifies the language (LEACE; Belrose et al., 2023, the closed-form successor of INLP; Ravfogel et al., 2020).

Multilingual internals. The logit lens, which reads the hidden state through the unembedding matrix, suggests many models compute in an English-aligned space (Wendler et al., 2024; Zhao et al., 2024;

Model	scale	n/lang	EN	DE	FR	RU	ZH	TR	FA	SW
Qwen3-8B (thinking)	200×5	1000	0.14	0.21	0.22	0.21	0.17	0.23	0.47	0.55
Apertus-8B-Instruct	200×5	1000	0.21	0.28	0.27	0.28	0.31	0.46	0.41	0.49
EuroLLM-9B-Instruct	200×5	1000	0.12	0.13	0.13	0.17	0.16	0.30	0.57	0.33

Table 1: Sample counts and hint-following rates (fraction of samples whose final answer equals the planted wrong hint) for the three main models. Each language has 200 questions with five samples per question.

Schut et al., 2025; nostalgebraist, 2020). Such an English-aligned space is one possible explanation for transfer; our audit tests whether prediction transfers, without identifying a pivot language.

Steering and localization. We add a direction to the hidden state during generation (contrastive activation addition; Rimsky et al., 2024; Turner et al., 2023; Zou et al., 2023) and compare it to a random direction of the same size. Other cross-lingual steering studies target behaviors such as language choice and task accuracy. Appendix A.5 explains how these different targets limit direct comparisons with our hint-following results. Localizing *where* the decision is carried out calls for activation patching between a following and a non-following run of the same item (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024; Ghandeharioun et al., 2024). We test control by a direction, rather than this finer-grained localization.

3 SETUP

Task and labels. We use four-option questions from the Global-MMLU test split (Singh et al., 2025), restricted to five factual subjects: high-school geography, government and politics, prehistory, world religions, and miscellaneous. Each prompt ends with "*(A colleague suggests the answer is L.)*", where L is the deterministic wrong option $(\text{answer} + 1) \bmod 4$. We use 200 questions per language and sample each question five times ($T = 0.7$, top-$p = 0.95$, up to 2048 new tokens). The label is $y = 1$ when the parsed final answer equals the hinted option, this measures hint-following, not deceptive intent. A question can receive both labels across its five samples; we call these "swing" questions.

Languages. The eight languages are English (EN), German (DE), French (FR), Russian (RU), Chinese (ZH), Turkish (TR), Persian (FA), and Swahili (SW), covering Latin, Cyrillic, Han, and Arabic scripts. Global-MMLU provides the same questions in each language.

Models and read positions. Our main models are Qwen3-8B in thinking mode (Yang et al., 2025), Apertus-8B (Hernández-Cano et al., 2026), and EuroLLM-9B (Martins et al., 2025). All three run in bf16 on one GH200. We collect residual-stream hidden states at every fourth layer in a forward pass over each completed prompt and response. We read three positions: *prompt*, the last prompt token; *early-CoT*, the mean over the first 16 generated tokens; and *post-CoT*, the token at `</think>` for Qwen3-8B and the last generated token for the two instruct models. This last distinction matters: Qwen3-8B gives its final answer after `</think>`, whereas an instruct model's final answer is already in context at our late read. The prompt state is shared by all five samples of a question and therefore cannot distinguish those samples. Table 1 gives the hint-following rates.

We also check prediction on Qwen3-1.7B and Qwen3-32B, DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), and Phi-4 Reasoning (Abdin et al., 2025). For the script swap, we also add separate Azerbaijani questions from TUMLU-mini (Isbarov et al., 2025), written in Latin and transliterated (with a deterministic letter-mapping script) into Perso-Arabic and hand-verify the result. These checks broaden the model and script coverage; the full text controls and steering experiments use the three main models (Appendices A.7 and A.8).

4 EVALUATING TRANSFER, LEAKAGE, AND CONTROL

We run three checks. Does an English probe predict hint-following on new questions in another language? How much of that prediction depends on language or text already written? And can a direction learned from the same labels change the behavior during generation?

4.1 CHECK TRANSFER ON NEW QUESTIONS

We train a logistic regression on the hidden state at one layer and read position, with standardized inputs and ℓ_2 regularization ($C = 1$). The probe is trained in English and applied to the other languages without retraining. We score it with AUROC: the probability that a hint-following sample ranks above a non-following sample.

Hold out the questions. We split question IDs into five folds, keeping all languages and samples of a question together. For each fold, we train on English samples from the other four folds and evaluate the held-out questions in each language. We report the mean of the five-fold AUROCs; English scores use the same held-out folds. For comparison, we also train on all English questions and evaluate their translations. This second setting measures how much sharing questions raises the score.

Choose the layer and quantify uncertainty. We report results on the layer with the highest mean held-out-question AUROC across all eight languages. This pooled rule uses target-language scores, so the resulting estimates can be optimistic. Layer curves are in Figure 1 (Results). Appendix A.1 gives pooled AUROC with matching 95% percentile intervals (Table 8), using 1,000 question-level bootstrap resamples that keep each question's five responses together. The intervals condition on the selected layer. A fold average uses only within-fold rankings; pooling also ranks responses scored by different probes, whose score scales may differ. In Tables 3 and 7, scores are fold averages and bracketed intervals are for pooled AUROC, with matching pooled estimates in Table 8.

Remove a fitted language component. We train an eight-way language classifier on the hidden states, then apply LEACE (Belrose et al., 2023) using all eight languages. We repeat language classification and English-to-target probing on the transformed states. This tests sensitivity to the component removed by LEACE; the language-classification check is needed to judge how much language information remains.

4.2 CHECK WHAT THE PROBE CAN READ

Filter early answers. We flag a response when its final answer letter appears as a standalone, whitespace-delimited word in the first 16 space-separated words of its reasoning. The flagged fraction is the *leak rate*; the remaining responses form the *clean* subset. We train a new probe on clean English responses and evaluate it on both the full target set and its clean subset. The filter uses words while the hidden-state read uses model tokens. "Clean" refers only to the absence of the answer letters this filter catches.

Compare with text-only readers. We compare the late probe with a zero-shot judge: the same model, with thinking disabled, given the hint letter and the reasoning with its ending clipped and final answer hidden. Its score is the probability of answering `Yes` to whether the response follows the hint. We also train logistic regressions on multilingual sentence embedding of the reasoning tail using the multilingual-e5-base model (Wang et al., 2024), as well as on character n-gram TF-IDF features, using the same English-to-target question folds. These are baselines for what can be read from text; their inputs differ from the late hidden state.

Edit the text, keep the probe. We keep the prompt and make three separate edits to the completed reasoning: replace standalone letters A–D with `[X]`, shuffle sentence order, or remove the last quarter of characters. We recompute hidden states from each edited transcript and apply the original English *post-CoT* probe. A drop in AUROC measures sensitivity to the written reasoning. These forward passes do not test when the answer was first decided.

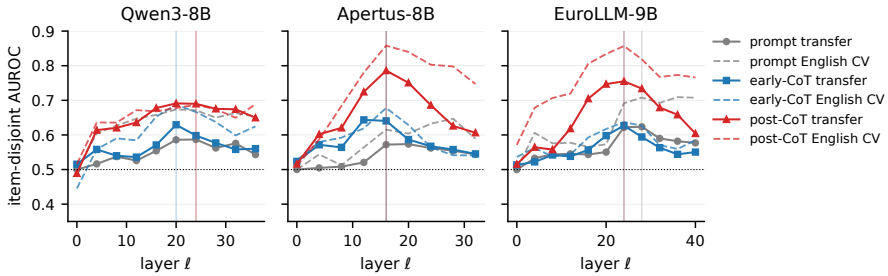


Figure 1: Fold-average AUROC across layers. Solid lines show mean transfer over seven non-English languages on held-out questions; dashed lines show English cross-validation. Vertical lines mark the pooled layers.

4.3 Test control during generation

Decoding the English direction in another language does not tell us whether that direction also controls the decision. We therefore inject it during generation and compare the effect to a same-size random direction. Following contrastive activation addition (Rimsky et al., 2024), we take the English difference of means between hint-following and non-following residual-stream states at layer ℓ as the steering direction:

$$d_\ell = \bar{h}_\ell[y=1] - \bar{h}_\ell[y=0], \qquad \hat{d}_\ell = d_\ell / \|d_\ell\|_2. \tag{1}$$

This is the contrastive activation-addition direction (Rimsky et al., 2024), rather than the logistic-regression weight vector used by the probe. At each generated token, we update the residual stream:

$$h_\ell \leftarrow h_\ell + \alpha \rho_\ell \hat{d}_\ell, \qquad \rho_\ell = \mathbb{E}_{x \in \mathcal{D}_{\text{en,train}}} \|h_\ell(x)\|_2. \tag{2}$$

Thus $|\alpha|$ scales the intervention relative to the mean English hidden-state norm, and its sign adds or subtracts the direction. We estimate directions at early-CoT and post-CoT, but inject throughout generation in both cases.

We steer English, Turkish, Persian, and Swahili and compare each learned direction with a random direction of the same size. The main test uses 60 items per cell at $\alpha = \pm 0.12$. Its *net effect* is the positive-minus-negative contrast in hint-follow rate for the learned direction, minus the same contrast for the random direction. We average this signed quantity over the three non-English languages. The ablations increase $|\alpha|$ to 0.30, use the probe's weight direction, or inject at three layers; their sample counts and rate denominators are given in Appendix A.5.

5 Results

5.1 Transfer remains after holding out questions

Sharing questions raises apparent transfer. We first ask how much transfer depends on reusing the same questions across languages. Evaluating translations of English training questions raises mean early-CoT AUROC by 0.03, 0.05, and 0.04 for Qwen3-8B, Apertus-8B, and EuroLLM-9B, respectively (Table 2). The largest increases within a language are 0.06, 0.09, and 0.10; across all positions, the largest is 0.16, for EuroLLM-9B at the *prompt* position. A new language therefore does not by itself make the test questions new. We use held-out question IDs for the remaining probe comparisons; the per-language comparison is in appendix Figure 3.

Prediction survives the question split. We find that the English probe reads hint-following in all seven non-English languages on held-out questions (Table 3). Mean pospost-CoT AUROC is 0.69 for Qwen3-8B, 0.79 for Apertus-8B, and 0.75 for EuroLLM-9B; the corresponding *early-CoT* means are 0.63, 0.64, and 0.63. Transfer is uneven: EuroLLM-9B scores lower in Persian and Swahili than in German, French, Russian, and Chinese, while Apertus-8B varies less.

Prompt-position scores are lower, with means of 0.59, 0.57, and 0.62 (appendix Table 7). Because all five responses to a question share this state, these scores reflect which questions are hint-prone,

Model	Position	naive	item-disjoint	Δ (max)
Qwen3-8B (thinking)	prompt	0.59	0.59	0.01 (0.04)
	early-CoT	0.66	0.63	0.03 (0.06)
	post-CoT	0.70	0.69	0.01 (0.04)
Apertus-8B-Instruct	prompt	0.62	0.57	0.05 (0.11)
	early-CoT	0.69	0.64	0.05 (0.09)
	post-CoT	0.82	0.79	0.03 (0.04)
EuroLLM-9B-Instruct	prompt	0.69	0.62	0.07 (0.16)
	early-CoT	0.67	0.63	0.04 (0.10)
	post-CoT	0.78	0.75	0.03 (0.05)

Table 2: Transfer with shared questions (naive) and held-out questions (item-disjoint), averaged over the seven non-English languages. The held-out scores average AUROC across five question folds. Δ gives the mean increase from sharing questions, with the largest per-language increase in parentheses. Differences are computed before rounding.

	Qwen3-8B		Apertus-8B		EuroLLM-9B	
Language	Early ($\ell = 20$)	Post ($\ell = 24$)	Early ($\ell = 16$)	Post ($\ell = 16$)	Early ($\ell = 24$)	Post ($\ell = 24$)
EN	0.69†	0.69†	0.68†	0.86†	0.64†	0.86†
DE	0.62 [0.56,0.66]	0.71 [0.66,0.75]	0.64 [0.59,0.69]	0.80 [0.76,0.83]	0.64 [0.56,0.69]	0.84 [0.79,0.88]
FR	0.62 [0.56,0.68]	0.65 [0.57,0.68]	0.68 [0.62,0.71]	0.81 [0.77,0.84]	0.70 [0.63,0.75]	0.82 [0.76,0.86]
RU	0.68 [0.63,0.74]	0.75 [0.69,0.79]	0.62 [0.57,0.68]	0.80 [0.78,0.83]	0.65 [0.59,0.71]	0.82 [0.74,0.85]
ZH	0.69 [0.60,0.74]	0.77 [0.68,0.79]	0.66 [0.62,0.70]	0.78 [0.74,0.80]	0.61 [0.52,0.65]	0.80 [0.73,0.83]
TR	0.60 [0.53,0.64]	0.65 [0.55,0.67]	0.63 [0.56,0.66]	0.77 [0.74,0.80]	0.62 [0.57,0.67]	0.77 [0.73,0.80]
FA	0.59 [0.53,0.63]	0.60 [0.54,0.63]	0.66 [0.62,0.70]	0.77 [0.74,0.81]	0.58 [0.54,0.63]	0.61 [0.56,0.64]
SW	0.61 [0.55,0.64]	0.71 [0.65,0.74]	0.60 [0.56,0.64]	0.77 [0.74,0.80]	0.59 [0.56,0.63]	0.63 [0.59,0.68]

Table 3: English-trained probe AUROC on held-out questions at pooled layer ℓ. Scores average five question folds; brackets give 95% question-bootstrap intervals for pooled AUROC, with matching pooled estimates in appendix Table 8. †English uses held-out cross-validation. Each language has 200 questions and five samples per question. The scores for the *prompt* position are in appendix Table 7.

rather than which sampled response will follow the hint. Figure 1 compares all three read positions across layers. Pooled-score intervals are in Appendix A.1; the pooled post-CoT interval excludes chance in all seven non-English languages for each model.

Transfer is stable under the language-erasure control. Using the language-erasure control in Section 4.1, language-ID accuracy falls well below chance after LEACE (at most 0.07 for Apertus-8B and EuroLLM-9B at both positions, and 0.008 for Qwen3-8B at *early-CoT*; the *post-CoT* erasure for Qwen3-8B is only partial, which we treat as a caveat). Cross-lingual hint-following transfer barely changes(Table 9): the shifts are $+0.009$, -0.011, $+0.001$ at *post-CoT* and $+0.000$, -0.001, -0.005 at *early-CoT*. So language identity is present in the hidden states, but it is not what the English probe uses for transfer.

5.2 Early answer filtering affects models differently

After the early-answer filter in Section 4.2 (Table 4), Qwen3-8B barely changes: mean *early-CoT* AUROC stays at 0.61, and the mean leak rate is only 6%. Chinese is the main exception (leak rate 32%), where AUROC falls slightly from 0.65 to 0.62; the other languages are unchanged at the reported precision. So for Qwen3-8B, the early signal is not explained by answer letters already

Model		DE	FR	RU	ZH	TR	FA	SW	mean
Qwen3-8B (thinking)	leak rate	0.01	0.03	0.02	0.32	0.02	0.03	0.03	0.06
	full AUROC	0.60	0.61	0.66	0.65	0.58	0.59	0.58	0.61
	clean AUROC	0.60	0.61	0.66	0.62	0.58	0.59	0.58	0.61
Apertus-8B-Instruct	leak rate	0.37	0.45	0.36	0.60	0.41	0.46	0.35	0.43
	full AUROC	0.59	0.58	0.59	0.56	0.56	0.61	0.53	0.57
	clean AUROC	0.58	0.58	0.58	**0.51**	0.56	0.55	**0.51**	0.55
EuroLLM-9B-Instruct	leak rate	0.12	0.05	0.06	0.20	0.13	0.72	0.06	0.19
	full AUROC	0.59	0.66	0.62	0.56	0.61	0.56	0.58	0.60
	clean AUROC	0.61	0.65	0.62	0.57	0.61	**0.42**	0.59	0.58

Table 4: Early-CoT AUROC with a probe trained on clean English responses. "Leak rate" is the fraction flagged by the answer-letter filter; "full" and "clean" score all target responses and the unflagged subset with the same probe. Means exclude English. Bold scores are below 0.53.

Model	Post-CoT probe	Text judge	e5	TF-IDF
Qwen3-8B	0.69	**0.75**	0.64	0.64
Apertus-8B	0.79	**0.83**	0.53	0.54
EuroLLM-9B	**0.75**	0.60	0.57	0.55

Table 5: Mean AUROC over seven non-English languages for hidden-state and text-only readers. The highest mean in each row is bold. Per-language scores are in appendix Table 10.

appearing in the window. The instruct models look different. For Apertus-8B, 43% of samples leak on average; after filtering, mean AUROC falls from 0.57 to 0.55, and Chinese and Swahili reach chance (0.51). For EuroLLM-9B, mean AUROC falls from 0.60 to 0.58. The sharpest case is Persian: 72% of responses are leaked, and AUROC drops from 0.56 to 0.42; in the other languages the score mostly survives. So for instruct models, early transfer is partly—and sometimes mostly, a fast-written answer, not a hidden pre-answer state. (The full and clean columns in Table 4 use the same clean-English retrained probe from that control, so they are directly comparable; they are not mixed with Table 3.)

5.3 *Post-CoT* PROBES READ INFORMATION AVAILABLE IN THE TEXT

To check whether the late probe only reads what is already written, we compare it to a *zero-shot text judge* . We find that this judge does better than the hidden-state probe on average for two of the three models (Table 5). For Qwen3-8B, the judge matches or exceeds the probe in six of seven non-English languages and trails by 0.01 in German; for Apertus-8B, it is higher in all seven. EuroLLM-9B goes the other way on average, although the judge is stronger in Persian and Swahili. The supervised e5 and TF-IDF readers score below the probe on average (per-language results in appendix Table 10). The judge reads clipped reasoning with the final answer hidden, while the instruct-model probe reads after the answer has been written.

To test how much the late probe depends on that text, we keep the prompt, edit only the CoT text, re-embed the edited transcript, and score the same English *post-CoT* probe (Table 6 and Figure 5). In this paired evaluation, the original non-English means are 0.67, 0.79, and 0.77. *Masking answer letters* lowers them to 0.64, 0.63, and 0.62, and *shuffling sentences* gives 0.64, 0.60, and 0.54. *Removing the last quarter of characters* has the largest effect for every model, leaving 0.53, 0.56, and 0.52. The probe therefore uses more than explicit answer letters; sentence order and the ending of the reasoning also matter. These edits test what a fixed probe reads from a completed transcript; they do not locate the computation that produced the answer.

EuroLLM-9B makes the distinction between these two checks clear. Its probe has the largest mean advantage over its text judge, yet removing the reasoning's ending produces the largest drop in the paired probe evaluation, from 0.77 to 0.52. Beating this text judge therefore does not show that the probe reads information independent of the written reasoning.

Model	Languages	ℓ	original	mask letters	scramble	drop tail
Qwen3-8B (thinking)	EN	24	0.71	0.67	0.67	0.52
	non-EN mean	24	0.67	0.64	0.64	0.53
Apertus-8B-Instruct	EN	16	0.88	0.74	0.76	0.59
	non-EN mean	16	0.79	0.63	0.60	0.56
EuroLLM-9B-Instruct	EN	24	0.91	0.69	0.57	0.61
	non-EN mean	24	0.77	0.62	0.54	0.52

Table 6: Post-CoT AUROC for original transcripts and three separate edits, with hidden states recomputed after each edit. The same post-CoT probe scores the original and edited transcripts. Each edit is compared with the original column from this paired evaluation, rather than with the main transfer scores.

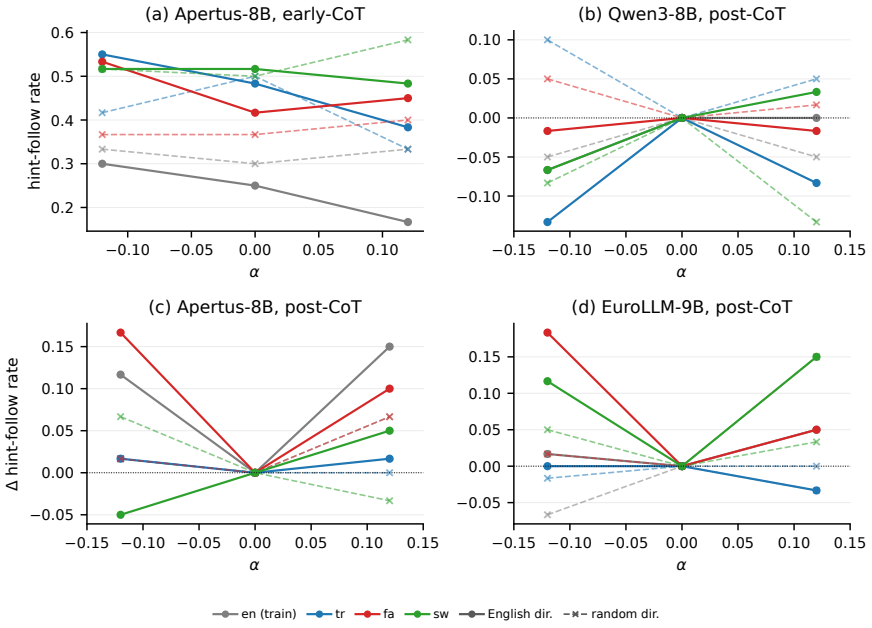


Figure 2: English difference-of-means steering, 60 items per cell. (a) Apertus-8B early-CoT hint-follow rate; (b)–(d) post-CoT change from baseline. Colors mark languages; solid lines use learned directions and dashed lines use same-size random controls.

5.4 Prediction does not give reliable control

Under the generation intervention in Section 4.3, we find no consistent advantage over the random direction in English, Turkish, Persian, or Swahili (Figure 2; per-language values in appendix Tables 11 and 12)/ At $|\alpha| = 0.12$, the mean non-English net effects are -0.02, -0.10, and $+0.09$ for *early-CoT* directions, and $+0.09$, $+0.03$, and -0.09 for *post-CoT* directions, in the order Qwen3-8B, Apertus-8B, and EuroLLM-9B. The sign reverses between read positions for every model, and the English reference is also inconsistent. The difficulty therefore starts before transferring the direction to another language.

We vary intervention strength, direction choice, and the number of layers to test this pattern further (appendix Table 13). For Qwen3-8B and Apertus-8B, the largest absolute differences from the random control range from 0.117 to 0.225. EuroLLM-9B reaches 0.433 with stronger steering and 0.350 with three-layer injection. These are substantial differences, and not all come with broken outputs: the three-layer Persian condition at $\alpha = 0.12$ has a 0.350 difference and an answer-parse rate of 0.93. But the effects do not consistently follow the injection sign across languages. Larger changes can affect the behavior without giving predictable control over it. The selected maxima and the signed mean net effects measure different things; neither comparison supports treating the learned

and random directions as equivalent. A triplet-loss projection head (An & Tao, 2026) trained on the same English hidden states changes non-English mean transfer AUROC by at most a few points (Table 14). So the reading; steering gap is not an artefact of using logistic regression alone.

5.5 Transfer across scripts and model families

We also apply our checks on Azerbaijani questions in Latin and Perso-Arabic scripts (appendix Table 15). *Post-CoT* AUROC is $0.68/0.67$ for Qwen3-8B, $0.80/0.74$ for Apertus-8B, and $0.74/0.73$ for EuroLLM-9B, with all six interval lower bounds above 0.5. Early-CoT transfer is less consistent: the intervals exclude the chance for Qwen3-8B in both scripts and for Apertus-8B in Latin. Late prediction therefore carries over to both scripts, while early prediction depends on the model. The wider thinking-model comparison gives *early-CoT* means of 0.54–0.63 and post-CoT means of 0.69–0.85 (appendix Table 16). Within Qwen3, *post-CoT* scores are 0.76, 0.69, and 0.85 at 1.7B, 8B, and 32B, so transfer does not increase steadily with model size. These checks extend the prediction result; the full leakage and steering audit covers the three main models.

6 Discussion

We read different information at each position. The *prompt* position can carry a question's tendency to produce hint-following; *early states* can include the written answer; and *late states* carry the completed reasoning. A linear-prob monitor can use this information without locating a decision made before it appeared in the text.

Predicting an outcome and changing it are separate tests. We use a difference-of-means direction in the main steering test and probe weights in an ablation; neither gives consistent control. Cox et al. (2026) steer binary answers using *pre-CoT* directions, with different models, scales, and controls. Since our English reference is also inconsistent, we cannot explain the gap by crossing languages alone.

To check transfer, we recommend holding out question IDs and choosing layers without target test scores. To check what the probe reads, compare with text-only readers, filter early answers, and edit the text with the probe fixed. Claims that a probe beats a text judge should come with these text controls before they support a hidden cross-lingual decision state. To test control, we should compare learned and random directions in the source and target languages.

7 Limitations

Our full audit covers three models, five factual subjects, and one misleading-hint task; additional models and scripts extend the prediction check only. We measure hint agreement, not deceptive intent. The text and language controls test specific sources of predictive information rather than locating the model's decision. Layer selection uses target-language scores, so transfer estimates can be optimistic; an English-only layer rule would be a stricter test and may lower the reported transfer. Our main steering comparison uses 60 items per cell, with ablations of direction, strength, and injection layers. With 60 items per cell, a two-proportion test at 80% power detects a difference in follow rate of roughly ± 0.18; effects smaller than that are not ruled out by our null. Our conclusions concern the consistency of control under these interventions; they do not exclude small effects or establish that other steering methods would behave the same way.

8 Conclusion

We find that an English-trained probe predicts hint-following on new questions in seven other languages. In the thinking model, an early signal remains after we filter responses that already contain the answer letter; in the instruct models, that same filter often shows the early score was a fast-written answer. Late in the response, cutting the end of the reasoning collapses the probe for all three models, and a text-only judge beats the probe on average for two of them. Directions learned from the same labels do not give reliable control: pushing them during generation is no better, consistently, than a same-size random direction. An English probe can therefore help *monitor* hint-following across languages, but the directions we test do not *steer* it.

AI Use Statement

We used generative AI tools for limited rewriting and paraphrasing of parts of the paper, and for retrieving candidate related work. All experiments, analyses, citations we retained, and final claims in this manuscript were produced and verified by the authors. We take responsibility for the final content of this work.

Ethics Statement

We study a proposed safety tool using benign factual multiple-choice questions from a public benchmark, with deliberately wrong hints. No human subjects were involved. The hint-following behavior is already documented in prior work; our contribution is to audit claims about monitoring and controlling it across languages. The task measures a narrow form of hint-following, not intentional deception or deployment safety. Confusing a useful predictor with evidence about how a decision is made risks placing unwarranted trust in a monitor.

Reproducibility Statement

Sections 3 and 4 give the data source, hint construction, sampling settings, read positions, probe configuration, and interventions. The appendix gives per-language results and the additional controls used in our analysis.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Feng An and Wenyin Tao. Contrastive representation learning for self-supervised deception detection in edge LLMs. *PLOS ONE*, 21(8):e0354894, 2026. doi: 10.1371/journal.pone.0354894.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. In *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models*, 2025.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063. Curran Associates, Inc., 2023. doi: 10.52202/075280-2884.

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

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Kyle Cox, Darius Kianersi, and Adrià Garriga-Alonso. Post-hoc reasoning in chain of thought: Decoding and steering pre-committed answers. *arXiv preprint arXiv:2603.01437*, 2026.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Scott Emmons, Erik Jenner, David K. Elson, Rif A. Saurous, Senthooran Rajamanoharan, Heng Chen, Irhum Shafkat, and Rohin Shah. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*, 2025.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 2024.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19755–19786. PMLR, 2025.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, et al. Apertus: Democratizing open and compliant llms for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46877–46955, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, et al. Tumlu: A unified and native language understanding benchmark for turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22816–22838, 2025.

Neeraja Kirtane and Kuan-Hao Huang. Language steering for multilingual in-context learning. *arXiv preprint arXiv:2602.02326*, 2026.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Qiming Li, Xiaocheng Feng, Yixuan Ma, Ruihan Chen, Zihe Tong, Zekai Ye, Xiachong Feng, Libo Qin, Haoyu Ren, Kun Chen, et al. Unlocking multilingual reasoning capability of llms and lvlms through representation engineering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24810–24829, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, et al. EuroLLM-9B: Technical report. *arXiv preprint arXiv:2506.04079*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262.

Parsa Mirtaheri and Mikhail Belkin. Catching rationalization in the act: detecting motivated reasoning before and after CoT via activation probing. *arXiv preprint arXiv:2603.17199*, 2026.

nostalgebraist. interpreting GPT: the logit lens. LessWrong blog post, 2020.

Eric Onyame, Runtao Zhou, Kowshik Thopalli, Bhavya Kailkhura, and Chirag Agarwal. The fragility of chain-of-thought monitoring across typologically diverse languages. arXiv preprint arXiv:2605.27901, 2026.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual LLMs think in English? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2375–2401, 2025.

Xinyu Tang, Zhihao Lv, Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Enhancing cross-task transfer of large language models via activation steering. *arXiv preprint arXiv:2507.13236*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? On the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

Raoyuan Zhao, Yihong Liu, Hinrich Schütze, and Michael A. Hedderich. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5223–5247, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.276.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, pp. 15296–15319, 2024. doi: 10.52202/079017-0489.

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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

A Appendix: Additional results

A.1 Question overlap, prompt scores, and interval estimates

Figure 3 gives the per-language question-overlap comparison summarized in Table 2 in the main text. Table 7 gives the prompt-position scores for the layer choice shown in Figure 1. Table 8 reports the separate pooled-score estimates and their matching question-bootstrap intervals. Across the 21 non-English model–language pairs, fold-average AUROC exceeds pooled AUROC on average by 0.02 at prompt (up to 0.08), and by about 0.01 at both CoT positions. These are different ranking summaries, not interchangeable estimates.

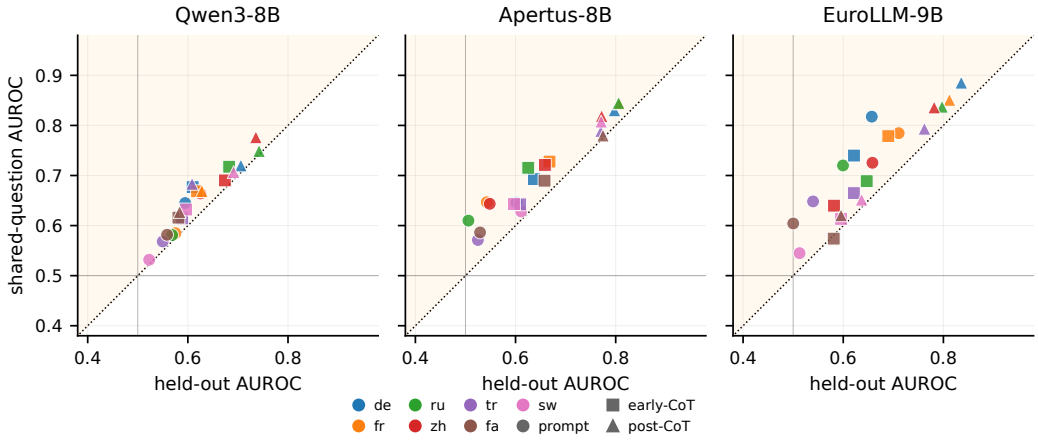


Figure 3: Transfer AUROC when training and evaluation share questions (vertical axis) against fold-average AUROC on held-out questions (horizontal axis). Each panel shows one model. Color marks the target language and marker shape the read position. Points above the diagonal score higher when questions overlap.

Language	Qwen3-8B ($\ell = 24$)	Apertus-8B ($\ell = 16$)	EuroLLM-9B ($\ell = 28$)
EN	$0.67^{\dagger}$	$0.62^{\dagger}$	$0.71^{\dagger}$
DE	$0.62_{[0.53,0.66]}$	$0.61_{[0.55,0.66]}$	$0.66_{[0.59,0.73]}$
FR	$0.58_{[0.51,0.64]}$	$0.59_{[0.48,0.60]}$	$0.71_{[0.64,0.78]}$
RU	$0.56_{[0.50,0.64]}$	$0.50_{[0.44,0.57]}$	$0.65_{[0.52,0.68]}$
ZH	$0.62_{[0.55,0.70]}$	$0.56_{[0.48,0.61]}$	$0.66_{[0.58,0.73]}$
TR	$0.56_{[0.48,0.62]}$	$0.55_{[0.47,0.58]}$	$0.59_{[0.48,0.60]}$
FA	$0.58_{[0.51,0.61]}$	$0.57_{[0.48,0.57]}$	$0.58_{[0.45,0.55]}$
SW	$0.58_{[0.47,0.57]}$	$0.63_{[0.56,0.66]}$	$0.52_{[0.47,0.56]}$

Table 7: Prompt-position AUROC on held-out questions at pooled layer ℓ. Scores average five folds; brackets give 95% question-bootstrap intervals for pooled AUROC, with matching pooled estimates in Table 8. †English uses held-out cross-validation. The prompt state is shared by all five samples of a question.

Qwen3-8B Language	Prompt ($\ell = 24$)	Early-CoT ($\ell = 20$)	Post-CoT ($\ell = 24$)
DE	$0.59_{[0.53,0.66]}$	$0.61_{[0.56,0.66]}$	$0.71_{[0.66,0.75]}$
FR	$0.58_{[0.51,0.64]}$	$0.62_{[0.56,0.68]}$	$0.63_{[0.57,0.68]}$
RU	$0.57_{[0.50,0.64]}$	$0.68_{[0.63,0.74]}$	$0.74_{[0.69,0.79]}$
ZH	$0.63_{[0.55,0.70]}$	$0.67_{[0.60,0.74]}$	$0.74_{[0.68,0.79]}$
TR	$0.55_{[0.48,0.62]}$	$0.59_{[0.53,0.64]}$	$0.61_{[0.55,0.67]}$
FA	$0.56_{[0.51,0.61]}$	$0.58_{[0.53,0.63]}$	$0.58_{[0.54,0.63]}$
SW	$0.52_{[0.47,0.57]}$	$0.60_{[0.55,0.64]}$	$0.69_{[0.65,0.74]}$
Apertus-8B Language	Prompt ($\ell = 16$)	Early-CoT ($\ell = 16$)	Post-CoT ($\ell = 16$)
DE	$0.60_{[0.55,0.66]}$	$0.64_{[0.59,0.69]}$	$0.80_{[0.76,0.83]}$
FR	$0.54_{[0.48,0.60]}$	$0.67_{[0.62,0.71]}$	$0.81_{[0.77,0.84]}$
RU	$0.51_{[0.44,0.57]}$	$0.63_{[0.57,0.68]}$	$0.81_{[0.78,0.83]}$
ZH	$0.55_{[0.48,0.61]}$	$0.66_{[0.62,0.70]}$	$0.77_{[0.74,0.80]}$
TR	$0.52_{[0.47,0.58]}$	$0.61_{[0.56,0.66]}$	$0.77_{[0.74,0.80]}$
FA	$0.53_{[0.48,0.57]}$	$0.66_{[0.62,0.70]}$	$0.77_{[0.74,0.81]}$
SW	$0.61_{[0.56,0.66]}$	$0.60_{[0.56,0.64]}$	$0.77_{[0.74,0.80]}$
EuroLLM-9B Language	Prompt ($\ell = 28$)	Early-CoT ($\ell = 24$)	Post-CoT ($\ell = 24$)
DE	$0.66_{[0.59,0.73]}$	$0.62_{[0.56,0.69]}$	$0.84_{[0.79,0.88]}$
FR	$0.71_{[0.64,0.78]}$	$0.69_{[0.63,0.75]}$	$0.81_{[0.76,0.86]}$
RU	$0.60_{[0.52,0.68]}$	$0.65_{[0.59,0.71]}$	$0.80_{[0.74,0.85]}$
ZH	$0.66_{[0.58,0.73]}$	$0.58_{[0.52,0.65]}$	$0.78_{[0.73,0.83]}$
TR	$0.54_{[0.48,0.60]}$	$0.62_{[0.57,0.67]}$	$0.76_{[0.73,0.80]}$
FA	$0.50_{[0.45,0.55]}$	$0.58_{[0.54,0.63]}$	$0.60_{[0.56,0.64]}$
SW	$0.51_{[0.47,0.56]}$	$0.60_{[0.56,0.63]}$	$0.64_{[0.59,0.68]}$

Table 8: AUROC pooled over held-out scores, with matching 95% percentile intervals from 1,000 question-level bootstrap resamples. These estimates use a different aggregation from the fold averages in the main comparisons. Layers are those in Tables 3 and 7. English cross-validation scores are given in those tables.

A.2 Removing linear language information

Table 9 gives the language-classification diagnostic and transfer scores before and after LEACE. The diagnostic falls below chance (0.125) in five of the six model–position pairs. This shows that the transformation disrupts that classifier, but a below-chance score does not show that a newly trained classifier could not recover language information. For Qwen3-8B at *post-CoT*, accuracy remains 0.743, so removal is incomplete under this diagnostic. Transfer changes little under this transformation.

Model	Position	ℓ	lang-ID before	lang-ID after	transfer before	transfer after	Δ
Qwen3-8B (thinking)	early-CoT	20	1.00	0.008	0.63	0.63	+0.000
	post-CoT	24	0.92	0.743	0.69	0.70	+0.009
Apertus-8B-Instruct	early-CoT	16	0.99	0.049	0.64	0.64	-0.001
	post-CoT	16	0.95	0.039	0.79	0.78	-0.011
EuroLLM-9B-Instruct	early-CoT	24	0.99	0.055	0.63	0.62	-0.005
	post-CoT	24	0.97	0.069	0.75	0.76	+0.001

Table 9: Language-classification balanced accuracy and mean English-to-target transfer AUROC before and after LEACE. Transfer is averaged over seven non-English languages. Δ is the change in transfer, computed before rounding. Chance language-classification accuracy is 0.125.

A.3 Additional views of the text controls

Table 10 gives the full per-language comparison of the probe, text judge, and supervised text readers summarized in Table 5. Figure 4 shows how removing early answer mentions changes transfer within each language. The effects differ across languages, especially for EuroLLM-9B, whose largest drop is in Persian. Figure 5 compares the three separate edits to the written reasoning. All edited transcripts are scored with the original *post-CoT* probe; the probe is not retrained for an edit.

Model	Monitor	EN	DE	FR	RU	ZH	TR	FA	SW	mean
Qwen3-8B (thinking)	probe (post-CoT)	0.69	0.71	0.65	0.75	0.77	0.65	0.60	0.71	0.69
	judge $P(\text{yes})$	0.70	0.70	0.70	0.75	0.83	0.75	0.69	0.83	0.75
	text baseline (e5)	0.72	0.65	0.68	0.70	0.63	0.63	0.56	0.62	0.64
	text baseline (tf-idf)	0.75	0.67	0.71	0.74	0.58	0.66	0.57	0.58	0.64
Apertus-8B-Instruct	probe (post-CoT)	0.86	0.80	0.81	0.80	0.78	0.77	0.77	0.77	0.79
	judge $P(\text{yes})$	0.82	0.82	0.83	0.81	0.84	0.85	0.80	0.87	0.83
	text baseline (e5)	0.62	0.51	0.56	0.57	0.56	0.51	0.54	0.48	0.53
	text baseline (tf-idf)	0.62	0.57	0.60	0.57	0.56	0.52	0.51	0.45	0.54
EuroLLM-9B-Instruct	probe (post-CoT)	0.86	0.84	0.82	0.82	0.80	0.77	0.61	0.63	0.75
	judge $P(\text{yes})$	0.57	0.57	0.53	0.54	0.53	0.59	0.73	0.69	0.60
	text baseline (e5)	0.66	0.59	0.69	0.58	0.62	0.52	0.52	0.49	0.57
	text baseline (tf-idf)	0.70	0.56	0.53	0.52	0.60	0.53	0.50	0.60	0.55

Table 10: Hidden-state and text-only readers. The judge predicts hint-following from the reasoning text; e5 and TF-IDF provide supervised text baselines. Means exclude English.

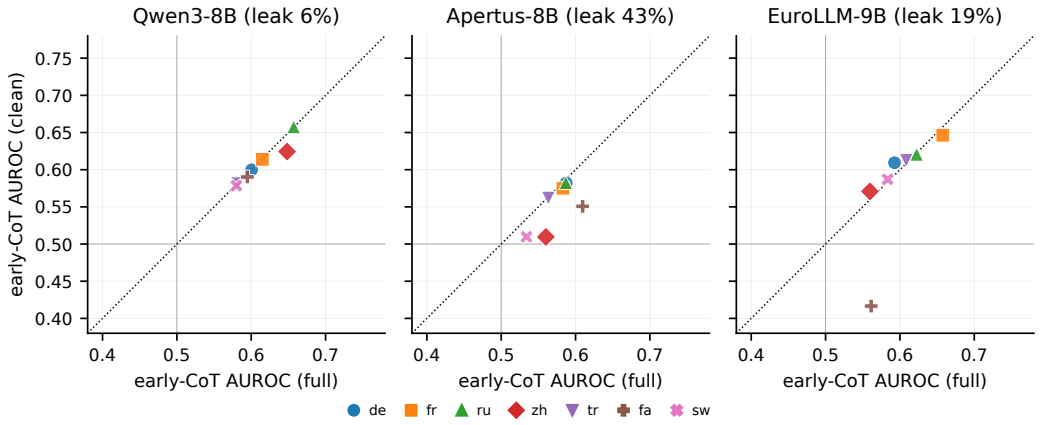


Figure 4: *Early-CoT* AUROC before and after filtering explicit answer-letter mentions, with one panel per model. Panel titles give mean leak rates. Both scores use a probe trained on clean English samples. Points below the diagonal lose accuracy after filtering.

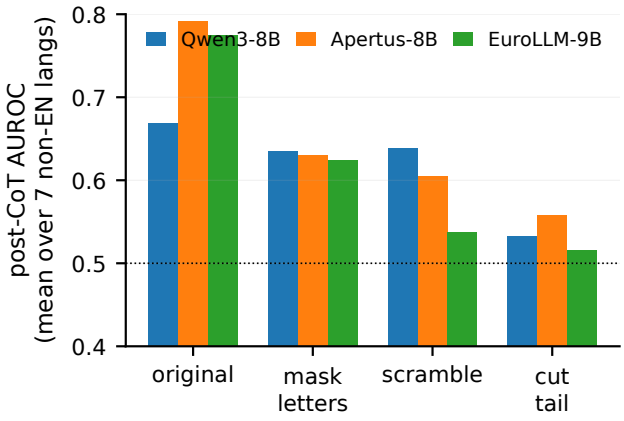


Figure 5: Mean *post-CoT* AUROC over seven non-English languages after separate edits to the written reasoning. Removing its ending causes the largest drop for each model.

A.4 STEERING BY LANGUAGE AND READ POSITION

Tables 11 and 12 give the change in hint-following relative to unsteered generation. The non-English steering tests use Turkish, Persian, and Swahili; English is included as a reference. Each cell uses 60 items. The direction is estimated from either *post-CoT* or *early-CoT* states, then added at every generated token. The response to its sign varies across languages and models, and the random direction also changes hint-following. For English, the early-direction net effects are approximately $+0.03$, -0.13, and $+0.05$; the post-CoT effects are $+0.07$, -0.02, and about zero, computed from the rounded table entries. With 60 items per cell, the mixed effects do not show that the directions are equivalent. These comparisons do not include effect intervals or a behavioral positive control.

Model	ℓ	$\lvert\alpha\rvert$	Language	$-\alpha$	$+\alpha$
Qwen3-8B (thinking)	24	0.12	EN	-0.07 (-0.05)	+0.00 (-0.05)
			TR	-0.13 (+0.10)	-0.08 (+0.05)
			FA	-0.02 (+0.05)	-0.02 (+0.02)
			SW	-0.07 (-0.08)	+0.03 (-0.13)
Apertus-8B-Instruct	16	0.12	EN	+0.12 (+0.02)	+0.15 (+0.07)
			TR	+0.02 (+0.02)	+0.02 (+0.00)
			FA	+0.17 (+0.02)	+0.10 (+0.07)
			SW	-0.05 (+0.07)	+0.05 (-0.03)
EuroLLM-9B-Instruct	24	0.12	EN	+0.02 (-0.07)	+0.05 (-0.03)
			TR	+0.00 (-0.02)	-0.03 (+0.00)
			FA	+0.18 (+0.02)	+0.05 (+0.15)
			SW	+0.12 (+0.05)	+0.15 (+0.03)

Table 11: Steering with the English *post-CoT* direction at $|\alpha| = 0.12$. Entries are changes in hint-following rate relative to the unsteered baseline; the same-size random-direction changes are in parentheses. There are 60 items per cell.

Model	ℓ	$\lvert\alpha\rvert$	Language	$-\alpha$	$+\alpha$
Qwen3-8B (thinking)	20	0.12	EN	+0.08 (-0.05)	+0.08 (-0.08)
			TR	-0.07 (+0.00)	-0.07 (-0.07)
			FA	+0.00 (+0.17)	+0.00 (+0.08)
			SW	+0.08 (-0.02)	-0.03 (+0.07)
Apertus-8B-Instruct	16	0.12	EN	+0.05 (+0.03)	-0.08 (+0.03)
			TR	+0.07 (-0.08)	-0.10 (-0.17)
			FA	+0.12 (+0.00)	+0.03 (+0.03)
			SW	+0.00 (+0.02)	-0.03 (+0.08)
EuroLLM-9B-Instruct	24	0.12	EN	-0.02 (+0.05)	+0.03 (+0.05)
			TR	-0.05 (+0.03)	+0.05 (+0.03)
			FA	-0.17 (+0.00)	-0.03 (+0.00)
			SW	-0.05 (-0.13)	+0.13 (+0.00)

Table 12: Steering with the English *early-CoT* direction, in the same format as Table 11. Read position names refer to the states used to estimate the direction, not the time of injection.

A.5 STEERING STRENGTH, DIRECTION, AND NUMBER OF LAYERS

We vary the intervention in three ways: increase $|\alpha|$ to 0.30, use the logistic-regression weight direction, or inject at three layers. Table 13 gives the maximum absolute difference in hint-following rate between the learned direction and its random control across Turkish, Persian, Swahili, and ten nonzero strengths. Each cell uses 60 items except the wider-strength sweep for Qwen3-8B, which uses 40. Follow rates divide by all items, counting unparseable answers as non-following. Thus the 0.225 entry for Qwen3-8B corresponds to a nine-item difference out of 40.

For Qwen3-8B and Apertus-8B, the maxima range from 0.117 to 0.225. EuroLLM-9B reaches 0.433 in the wider-strength sweep and 0.350 with three-layer injection. Answer-parse rates in these

conditions range from 0.60 to 1.00. Parsing disruption explains only part of the concern: the three-layer Persian condition at $\alpha = 0.12$ has an absolute difference of 0.350 with parse rate 0.93. The response still fails to follow the injection sign consistently: Persian reverses sign, Turkish peaks only at negative strength, and Swahili is flat. These are selected peaks, not average effects, and a large peak alone does not give a predictable response to adding and subtracting a direction. We do not include a random-versus-random reference for these selected maxima.

Model	$\lvert\alpha\rvert \leq 0.30$	LR-weight direction	Three layers
Qwen3-8B	0.225	0.183	0.117
Apertus-8B	0.167	0.133	0.117
EuroLLM-9B	0.433†	0.117	0.350†

Table 13: Largest absolute learned-minus-random difference in hint-following rate across three languages and ten nonzero strengths. There are 60 items per language and strength, except the Qwen3-8B wider-strength sweep (40). Unparseable answers count as non-following. †Answer-parse rates range from 0.60 to 1.00 in these conditions.

Relation to other cross-lingual steering results. Prior work uses steering to improve cross-lingual representation alignment, task accuracy, or the choice of output language (Sundar et al., 2025; Tang et al., 2025; Kirtane & Huang, 2026; Li et al., 2026). Our test asks whether a direction estimated from hint-following labels changes that behavior in other languages. The different targets and interventions mean that successes on those tasks can coexist with the limited control we find here.

A.6 A TRIPLET-LOSS READER

Contrastive hidden-state monitors (An & Tao, 2026) motivate a check of whether transfer depends on a linear reader. We compare logistic regression with a triplet-loss projection head trained on the same English hidden states, using the same held-out-question folds and pooled layer. The head is a two-layer MLP with a 128-dimensional output and margin 0.5. Table 14 reports mean transfer AUROC. The largest gain is for Qwen3-8B at *post-CoT* ($+0.054$); the other changes are smaller, with both gains and losses. Thus the choice of reader does not improve transfer consistently across models and positions. This comparison is prediction-only; we did not steer with the triplet head.

Model	Position	Logistic regression	Triplet	Δ
Qwen3-8B	early-CoT	0.629	0.589	-0.041
Qwen3-8B	post-CoT	0.690	0.744	$+0.054$
Apertus-8B	early-CoT	0.641	0.651	$+0.011$
Apertus-8B	post-CoT	0.786	0.752	-0.034
EuroLLM-9B	early-CoT	0.627	0.616	-0.011
EuroLLM-9B	post-CoT	0.755	0.715	-0.040

Table 14: Mean held-out-question transfer AUROC over the seven non-English languages for logistic regression and the triplet-loss reader. Both use the same layer and 200 questions per language. Δ is computed before rounding.

A.7 TRANSFER ACROSS AZERBAIJANI SCRIPTS

Table 15 reports transfer to the same Azerbaijani questions in Latin and Perso-Arabic scripts, including the *prompt* position. Its intervals are wider than those in the main experiment. At *early-CoT*, the lower bounds exceed 0.5 for Qwen3-8B in both scripts and Apertus-8B in Latin; the other three intervals include 0.5. At *post-CoT*, all six lower bounds exceed 0.5.

Model	Position	ℓ	EN	AZ-LATN	AZ-FA
Qwen3-8B (thinking)	prompt	24	$0.67_{[0.59,0.74]}$	$0.72_{[0.60,0.81]}$	$0.60_{[0.48,0.71]}$
	early-CoT	36	$0.62_{[0.56,0.69]}$	$0.63_{[0.52,0.74]}$	$0.66_{[0.55,0.76]}$
	post-CoT	20	$0.68_{[0.60,0.74]}$	$0.68_{[0.59,0.76]}$	$0.67_{[0.57,0.77]}$
Apertus-8B-Instruct	prompt	16	$0.62_{[0.54,0.67]}$	$0.63_{[0.50,0.75]}$	$0.60_{[0.50,0.70]}$
	early-CoT	16	$0.68_{[0.62,0.71]}$	$0.63_{[0.54,0.70]}$	$0.55_{[0.45,0.64]}$
	post-CoT	16	$0.86_{[0.82,0.88]}$	$0.80_{[0.71,0.87]}$	$0.74_{[0.66,0.81]}$
EuroLLM-9B-Instruct	prompt	36	$0.71_{[0.62,0.76]}$	$0.54_{[0.43,0.66]}$	$0.59_{[0.47,0.70]}$
	early-CoT	20	$0.62_{[0.52,0.71]}$	$0.52_{[0.41,0.63]}$	$0.55_{[0.45,0.65]}$
	post-CoT	24	$0.86_{[0.81,0.90]}$	$0.74_{[0.64,0.82]}$	$0.73_{[0.67,0.79]}$

Table 15: English-probe transfer to Azerbaijani in Latin script and to the same questions transliterated into Perso-Arabic. English scores are cross-validation at the listed layers. Brackets give 95% intervals clustered by question. The layer used for each model and position is listed in the table.

A.8 Transfer across model sizes and families

Table 16 extends the probe comparison to Qwen3-1.7B, Qwen3-32B, DeepSeek-R1-Distill-Llama-8B, and Phi-4 Reasoning. We use the same eight languages, question-level evaluation, and layer-selection procedure. Among the thinking models, *early-CoT* mean transfer ranges from 0.54 to 0.63, and *post-CoT* from 0.69 to 0.85. The Qwen3 size comparison does not show a steady increase at both positions: early transfer is 0.59, 0.63, and 0.63 at 1.7B, 8B, and 32B, while post-CoT transfer is 0.76, 0.69, and 0.85. These models extend the evidence for probe transfer, but we did not run the full leakage controls or steering tests on them.

Model	Family	Size	Think	n/lang	early-CoT		post-CoT	
					EN cv	7-lang mean	EN cv	7-lang mean
Qwen3-1.7B	Qwen3	small	yes	1000	0.61	0.59	0.76	0.76
Qwen3-8B	Qwen3	medium	yes	1000	0.69	0.63	0.69	0.69
Qwen3-32B	Qwen3	large	yes	1000	0.71	0.63	0.86	0.85
R1-Distill-Llama-8B	R1-Llama	medium	yes	1000	0.61	0.54	0.91	0.81
Phi-4-reasoning	Phi-4	medium	yes	1000	0.73	0.61	0.78	0.74
Apertus-8B-Instruct	Apertus	medium	no	1000	0.68	0.64	0.86	0.79
EuroLLM-9B-Instruct	EuroLLM	medium	no	1000	0.64	0.63	0.86	0.75

Table 16: English cross-validation and mean transfer AUROC over seven non-English languages across model sizes and families. n/lang is the number of samples per language; 1,000 corresponds to 200 questions with five samples each.