

Team champ at SHROOM-Visions: Transferring Fine-Grained Hallucination Detection with Hidden-State Probes and Backbone Adaptation

Aygalic Jara--Mikolajczak^{1,2}, Thomas Lavergne¹, Sophie Rosset¹

¹Laboratoire Interdisciplinaire des Sciences du Numérique, ²SCIAM

Correspondence: jara@lisn.fr, lavergne@lisn.fr, sophie.rosset@lisn.fr

Abstract

We describe a multilingual, character-level hallucination detection system for vision-language models combining multi-layer hidden states, text-to-vision LoRA transfer, multitask learning, and language-specific span emission. Our submission achieved a validation Character AUPRC of 0.5530 and an average test Cor/Cor+lbl of 0.5243/0.4174. A controlled post-submission factorial analysis improved validation Character AUPRC to 0.5595. Team *champ*¹ placed third in French, fourth in Italian and Chinese, and fifth in English.

1 Introduction

Fine-grained hallucination detection requires identifying not only if a vision-language model (VLM) response is unreliable, but also the specific character spans and error types involved. The SHROOM-Vision shared task (Vázquez et al., 2026) evaluates this across four languages, making efficient localization and cross-lingual generalization key challenges. We employ hidden-state probes as a computationally efficient alternative to autoregressive prompting, investigating whether supervision from the text-only Mu-SHROOM task can mitigate the need for target-modality annotations by transferring the probe head, LoRA adapters (Hu et al., 2021), or both. Our system integrates model and representation selection, a multi-layer probe head, text-to-vision adaptation, and language-specific span emission.

Contributions. We (1) evaluate the transferability of fine-grained hallucination probes from text to vision; (2) quantify the resulting reduction in required vision-specific supervision; and (3) provide a controlled factorial analysis of probe-head and adapter initialization and training objectives.

2 Related Work

Task setting. SHROOM-Vision (Mickus et al., 2026) extends the SHROOM series to image-conditioned text. Systems must assign a hallucination probability to every character and categorize each span into one of five types: invention, mischaracterisation, OCR error, miscounting, or other. While modalities and categories are new, the task retains Mu-SHROOM’s character-level soft labels (Vázquez et al., 2025). This shared objective allows us to test whether text-side supervision can reduce the need for image-conditioned annotations while the detector learns vision-specific features and categories.

From verification at generation time to probing. Early reference-free detectors trade efficiency for reliability via web search (Varshney et al., 2023), output agreement sampling (Manakul et al., 2023), or atomic fact checking (Min et al., 2023). Only the latter localizes errors, and all three incur high inference costs. In contrast, internal-representation probes decode signals directly from hidden states (Alain and Bengio, 2018; Azaria and Mitchell, 2023) in a single forward pass, enabling efficient fine-grained localization across languages (Alvi et al., 2026). While CH-Wang et al. (2024) find such probes effective in-domain, they note poor transfer across tasks and domains, motivating our controlled evaluation of text-to-vision transfer.

Detecting hallucinations in vision-language models. HalluShift (Dasgupta et al., 2025) and HalluShift++ (Nath et al., 2025) use internal state features for a trained classifier, a method adapted as the official token-level baseline using a frozen evaluator LVLM (Mickus et al., 2026). Unlike ours, it uses hand-designed feature summaries and learns only from vision annotations. Similarly, HALP (Kogilathota et al., 2026) and HalLoc (Park et al., 2025) employ single-pass probing, with HalLoc

¹French for “mushroom.”

adding a multitask head over a frozen backbone. Conversely, Gunjal et al. (2024) jointly fine-tune the head and InstructBLIP (Dai et al., 2023) backbone for sub-sentence prediction. ZINA (Wada et al., 2026) uses prompting for span localization and classification, requiring costly autoregressive decoding. All these methods rely exclusively on target-modality annotations.

Reusing text-side signal on vision. *TruthProbe* (Choi et al., 2026) uses linear probes on attention heads to derive *Truth Scores* that correlate strongly between base LLMs and their multimodal descendants, though it does not localize spans. Other works explore shared representations, such as applying the logit lens (nostalgebraist, 2020) to VLM images (Jiang et al., 2024) or identifying shared hallucination subspaces across LVLMs (Duan et al., 2025). Finally, text-to-image curricula have been applied to safety, guiding image projection (Xu et al., 2024) or inference-time correction (Liu et al., 2025). Our work extends this logic of cross-modal transfer to fine-grained hallucination localization.

3 System overview

We address fine-grained hallucination detection in SHROOM-Vision (Mickus et al., 2026). Given an image, a user prompt, and a VLM-generated response, the system identifies hallucinated character spans and assigns a hallucination category and confidence to each span. The four evaluation languages are English, French, Italian, and Chinese.

Our detector concatenates token representations from Qwen 3.5 27B layers 8, 32, and 45 and passes them through a two-hidden-layer multilayer perceptron (MLP; 256 units per layer, ReLU activations, 0.3 dropout) with separate scalar hallucination and five-way category output heads. During training, the probe head is jointly optimized with LoRA adapters applied to the backbone, allowing those representations to adapt to the detection task. During post-processing, token probabilities are aligned to response characters, spans are emitted using a language-specific emission-and-abstention policy, and punctuation and whitespace are trimmed from span boundaries.

Code and experiment tracking are available on the project GitLab page.²

²gitlab.lisn.upsaclay.fr/jara/champ

4 Data and preprocessing

We randomly partitioned the labeled SHROOM-Vision data into 60% training, 20% internal development, and 20% validation using seed 42. For data-efficiency experiments, we reshuffled the training partition once with seed 42 to construct nested subsets shared across all training-seed replicates. All model, layer, architecture, checkpoint, and emission-policy selection used all 3,020 examples in the validation split. For final system fitting, we combined the training and internal-development partitions while keeping this validation split untouched. Mu-SHROOM (Vázquez et al., 2025) was split 80/20 for text training and validation.

We used each model family’s native preprocessing pipeline and, for computational efficiency, resized images exceeding 1,003,520 pixels while preserving their aspect ratio. The same configuration was used for training, validation, and test inference. We defined a character or token as positive when at least one annotator assigned it a non-zero hallucination probability.

5 Detector

5.1 Model and layer selection

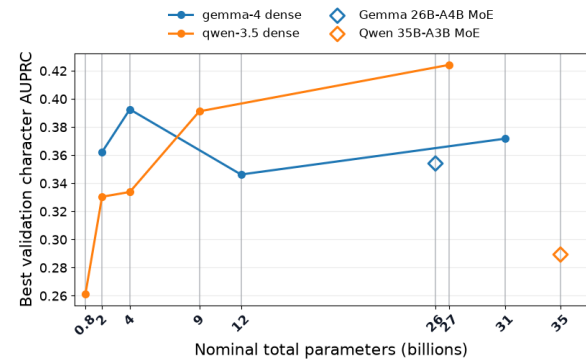


Figure 1: Best validation Character AUPRC for each model, plotted against nominal total parameter count. Diamonds distinguish mixture-of-experts models from dense models.

We trained linear probes on hidden representations from ten approximately evenly spaced decoder blocks of dense and mixture-of-experts Gemma 4 (Team et al., 2026) and Qwen 3.5 (Qwen-Team, 2026) models. We selected the backbone by its best single-layer validation Character AUPRC. Among dense models, this score generally increased with model size, while the tested mixture-of-experts models fell below this trend (Figure 1).

Qwen 3.5 27B performed best at layer 45; we therefore dropped Gemma 4 and used Qwen 3.5 27B for the remainder of the study. Within this backbone, we retained the best early, middle, and late candidate layers (8, 32, and 45) for multi-layer probing³

5.2 Representations and prediction head

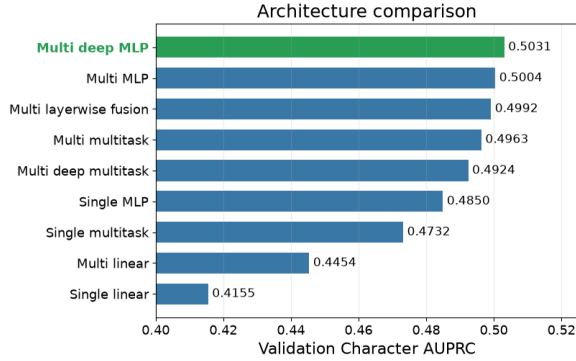


Figure 2: Validation Character AUPRC used to select the probe architecture. Single-layer probes consume one hidden-state layer, whereas multi-layer probes consume three. Layerwise fusion projects each layer separately before fusion; the deep MLP uses a larger shared network after concatenation.

The detector used Qwen 3.5 27B hidden states from layers 8, 32, and 45. Their concatenated representations were passed through the deep shared MLP selected in Figure 2, with two output heads. (1) A scalar detection head predicted hallucination probability using a sigmoid and binary cross-entropy loss. (2) A five-way category head predicted the hallucination type using a softmax and conditional soft cross-entropy loss over hallucinated tokens. The category loss was weighted by 0.3. We fixed this weight based on a preliminary sweep over $\{0, 0.1, 0.3, 1.0\}$ and did not retune it in the reported experiments. The training target had six components: clean probability plus the five hallucination-type probabilities.

The base-model weights remained frozen. We applied rank-32 LoRA adapters to decoder blocks 0–45, since block 45 is the deepest representation consumed by the probe and later blocks cannot affect its predictions. Within these blocks, LoRA targeted the attention projections (q_proj, k_proj, v_proj, and o_proj) and MLP projections (gate_proj, up_proj, and down_proj). The adapters and probe head were jointly optimized during final vision training. All final-system arms

³Detailed results for every tested model–layer pair appear in Figure 6 of Appendix B.

used seed 42 and the same data ordering and optimization budget.

5.3 Submitted curriculum and controlled factorial

An earlier controlled binary study tested whether a probe head, LoRA adapter, or both pretrained on text-only Mu-SHROOM improved training on the 60% vision partition. It selected the Mu-SHROOM head with a fresh adapter at 0.5461 validation Character AUPRC; this detector was not a submission candidate.

The submitted system used a different curriculum: a multitask head pretrained on Mu-SHROOM was first fitted to vision, then paired with an independently Mu-SHROOM-pretrained adapter for joint optimization on the combined 80% vision partition. Thus, the head was already vision-fitted, while the adapter supplied direct text initialization.

After submission, we brought together the factors studied separately in our binary-attribution and submitted-system comparisons. The resulting controlled $2 \times 2 \times 2$ factorial varied head initialization (text or vision), adapter initialization (text or fresh), and objective (binary or multitask). All eight arms were freshly trained with the submitted curriculum, seed, and validation set.

Unlike the submitted vision-head, text-adapter, multitask arm, the post-submission factorial selected the text-head, fresh-adapter, multitask arm by validation Character AUPRC (0.5595). Figure 3 gives the complete comparison.

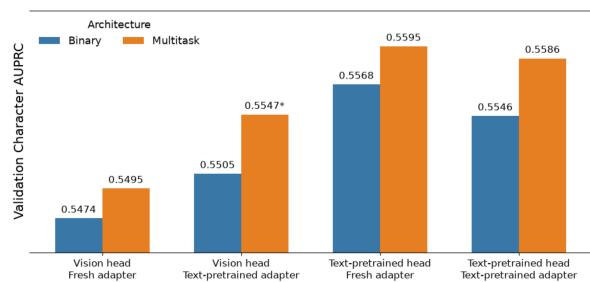


Figure 3: Validation Character AUPRC in the controlled final-fit factorial, grouped by probe-head and adapter initialization and split by training objective. The star marks the curriculum corresponding to the submitted system; every bar represents a fresh post-submission run.

The data-efficiency study carried forward the text-head initialization selected by the controlled transfer comparison. At each vision-training fraction, it compared this initialization with a fresh probe head while jointly training a fresh LoRA

Table 1: Organizer-platform results on the held-out test set for the controlled final-fit factorial. Each detector used an emission policy selected on validation. Cor+lbl is reported only for multitask probes, which predict hallucination categories. Head and adapter initializations are text-pretrained (Text), vision-fitted (Vision), or fresh; objectives are binary (Bin.) or multitask (Multi). Bold marks the best value in each column.

Head	Adapter	Obj.	EN		FR		IT		ZH	
			Cor+lbl	Cor	Cor+lbl	Cor	Cor+lbl	Cor	Cor+lbl	Cor
Text	Fresh	Bin.	–	0.5121	–	0.5358	–	0.5238	–	0.5402
Text	Fresh	Multi	0.3828	0.5018	0.4352	0.5319	0.418	0.5398	0.4713	0.5573
Text	Text	Bin.	–	0.4936	–	0.5391	–	0.5345	–	0.5464
Text	Text	Multi	0.3747	0.4991	0.444	0.5401	0.4212	0.5254	0.4764	0.555
Vision	Fresh	Bin.	–	0.4916	–	0.5199	–	0.5245	–	0.5457
Vision	Fresh	Multi	0.3762	0.4963	0.413	0.5213	0.4026	0.5318	0.4536	0.5438
Vision	Text	Bin.	–	0.5017	–	0.5232	–	0.5291	–	0.5473
Vision	Text	Multi	0.373	0.4972	0.4112	0.5183	0.4201	0.5356	0.4603	0.5511

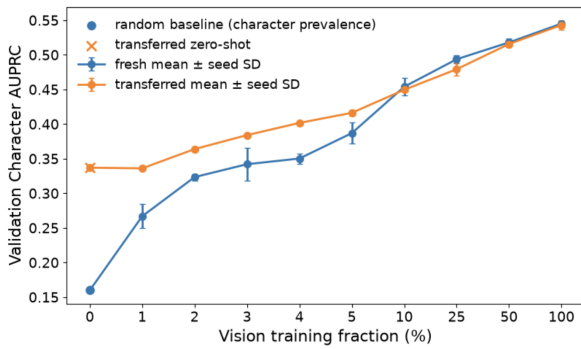


Figure 4: Three-seed mean validation Character AUPRC as the vision-training fraction increases. Both arms jointly train a fresh LoRA adapter and differ only in whether the probe head is freshly initialized or initialized from text pretraining; error bars show the sample standard deviation across seeds.

adapter in both arms. Figure 4 shows a mean Character AUPRC benefit at every 1–5% fraction, but no mean benefit from 10% onward. Thus, text-head initialization acts mainly as a low-data prior under joint adaptation.

The complementary frozen-backbone analysis, which isolates head initialization over ten seeds, is documented in Figure 5 of Appendix A.

6 Span emission

Token probabilities were aligned to response characters. Per language, threshold τ formed contiguous spans, while gate γ suppressed responses whose maximum probability was too low. We selected both by validation Cor with the detector frozen, then trimmed span-edge whitespace and punctuation. Table 2 reports the official platform results on the organizer-held test set.

Lang.	τ	γ	Cor	Cor+lbl	IoU
EN	0.28	0.44	0.4898	0.3721	0.4318
FR	0.20	0.40	0.5167	0.4095	0.4778
IT	0.28	0.44	0.5399	0.4213	0.4612
ZH	0.36	0.52	0.5506	0.4668	0.4892
Average	–	–	0.5243	0.4174	0.4650

Table 2: Selected emission policy and official held-out results. Average is the unweighted mean across languages.

7 Submitted system performance

On the complete validation set, the submitted detector reached Character AUROC/AUPRC 0.8464/0.5530 and Token AUROC/AUPRC 0.8656/0.5912. On the organizer-held test set, the platform returned average Cor/Cor+lbl of 0.4487/0.3181 with the initial global threshold, 0.5139/0.4090 after language-specific emission tuning, and 0.5243/0.4174 after edge trimming; final average IoU was 0.4650. Validation Character AUPRC selected one of four detector arms. Keeping it fixed, we submitted these three variants in sequence; edge trimming scored best on the platform and became final.

8 Post-submission results

On complete validation, this post-submission detector reached Character AUROC/AUPRC 0.8512/0.5595 and Token AUROC/AUPRC 0.8690/0.5981. Table 1 reports the organizers’ platform results on the held-out test set for all eight arms.

9 Conclusion

We presented a multilingual character-level hallucination detection system combining multi-layer

hidden-state probing, text-to-vision transfer, multi-task learning, and language-specific span emission. Our results show that text-side initialization is most useful under limited vision supervision, while the post-submission factorial analysis suggests that the submitted transfer configuration can be improved. Future work should verify these findings across seeds and model families.

Limitations

The earlier binary-transfer and submitted-system comparisons differed in objective, so their winners are not directly comparable. A matched post-submission factorial addressed this discrepancy but selected a different configuration from the submitted system. Its single seed and small margins cannot establish whether this ordering exceeds training noise; most model, layer, and architecture comparisons likewise lack uncertainty estimates.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011016140R1 made by GENCI. This work was funded by the ANRT under Cifre contract No. 2025/0109, with the scientific and financial support of LISN and SCIAM.

References

- Guillaume Alain and Yoshua Bengio. 2018. [Understanding intermediate layers using linear classifier probes](#). *Preprint*, arXiv:1610.01644.
- Riasad Alvi, Nurul Labib Sayeedi, and Md. Faiyaz Abdullah Sayeedi. 2026. [MultiHaluDet: Multilingual Hallucination Detection via LLM Hidden State Probing](#). In *Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026)*, pages 63–74, San Diego, United States. Association for Computational Linguistics.
- Amos Azaria and Tom Mitchell. 2023. [The Internal State of an LLM Knows When It’s Lying](#). *Preprint*, arXiv:2304.13734.
- Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. [Do Androids Know They’re Only Dreaming of Electric Sheep?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4401–4420, Bangkok, Thailand. Association for Computational Linguistics.
- Miso Choi, Seonga Choi, Mincheol Kwon, Woosung Joung, Jinkyu Kim, and Jungbeom Lee. 2026. [The Truth Stays in the Family: Enhancing Contextual](#)
- [Grounding via Inherited Truthful Heads in Model Lineages](#). *Preprint*, arXiv:2606.15821.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning](#). *Preprint*, arXiv:2305.06500.
- Sharanya Dasgupta, Sujoy Nath, Arkaprabha Basu, Pourya Shamsolmoali, and Swagatam Das. 2025. [HalluShift: Measuring Distribution Shifts towards Hallucination Detection in LLMs](#). *Preprint*, arXiv:2504.09482.
- Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. [TruthPrInt: Mitigating Large Vision-Language Models Object Hallucination via Latent Truthful-Guided Pre-Intervention](#). In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7372–7382.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. [Detecting and Preventing Hallucinations in Large Vision Language Models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18135–18143.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-Rank Adaptation of Large Language Models](#). *Preprint*, arXiv:2106.09685.
- Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. 2024. Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations. In *The Thirteenth International Conference on Learning Representations*.
- Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. 2026. [HALP: Detecting Hallucinations in Vision-Language Models without Generating a Single Token](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6067–6085, Rabat, Morocco. Association for Computational Linguistics.
- Qin Liu et al. 2025. [Unraveling and Mitigating Safety Alignment Degradation of Vision-Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3631–3643, Vienna, Austria. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. [SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models](#). *Preprint*, arXiv:2303.08896.
- Timothee Mickus, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Chuyuan Li, Aman Sinha, Lorenzo Vaiani, Jörg Tiedemann, and Raúl Vázquez. 2026. [Can humans dream of electric sheep?](#)

human-written samples for fine-grained vision-and-language hallucination benchmarking. *Preprint*, arXiv:2608.01021.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. **FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation**. *Preprint*, arXiv:2305.14251.

Sujoy Nath, Arkaprabha Basu, Sharanya Dasgupta, and Swagatam Das. 2025. **HalluShift++: Bridging Language and Vision through Internal Representation Shifts for Hierarchical Hallucinations in MLLMs**. *Preprint*, arXiv:2512.07687.

nostalgebraist. 2020. Interpreting GPT: The logit lens.

Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. **HalLoc: Token-level Localization of Hallucinations for Vision Language Models**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29893–29903.

QwenTeam. 2026. Qwen Studio. <https://qwenlm.github.io/blog/qwen3.5/>.

Gemma Team et al. 2026. **Gemma 4 Technical Report**. *Preprint*, arXiv:2607.02770.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jian-shu Chen, and Dong Yu. 2023. **A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation**. *Preprint*, arXiv:2307.03987.

Raúl Vázquez et al. 2025. Welcome to SemEval-2025 Task-3 — Mu-SHROOM, the Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes. <https://helsinki-nlp.github.io/shroom/>.

Raúl Vázquez, Aman Sinha, Chuyuan Li, Artem Shelmanov, Artem Vazhentsev, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Lorenzo Viani, Jörg Tiedemann, and Timothee Mickus. 2026. **Overview of shroom-visions 2026: A shared task on hallucination detection in large vision-language models**. *Preprint*, arXiv:2608.25662.

Yuiga Wada, Kazuki Matsuda, Komei Sugiura, and Graham Neubig. 2026. **ZINA: Multimodal Fine-grained Hallucination Detection and Editing**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32528–32538.

Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. 2024. **Cross-Modal Safety Mechanism Transfer in Large Vision-Language Models**. In *The Thirteenth International Conference on Learning Representations*.

A Head-only transfer

B Layer and model selection

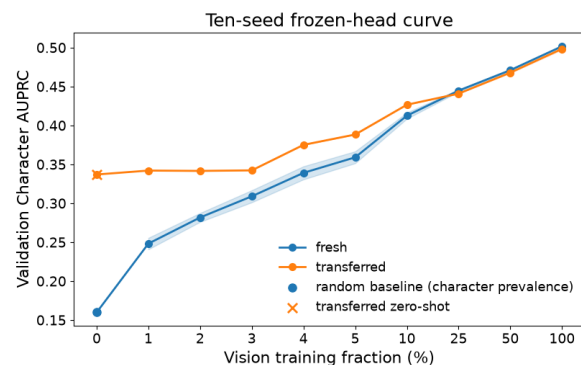


Figure 5: Ten-seed validation Character AUPRC for frozen fresh and text-pretrained probe heads as the vision-training fraction increases. Shaded bands show variation across seeds. Transfer helps through 10% of the vision-training partition; its advantage becomes slightly negative at larger fractions.

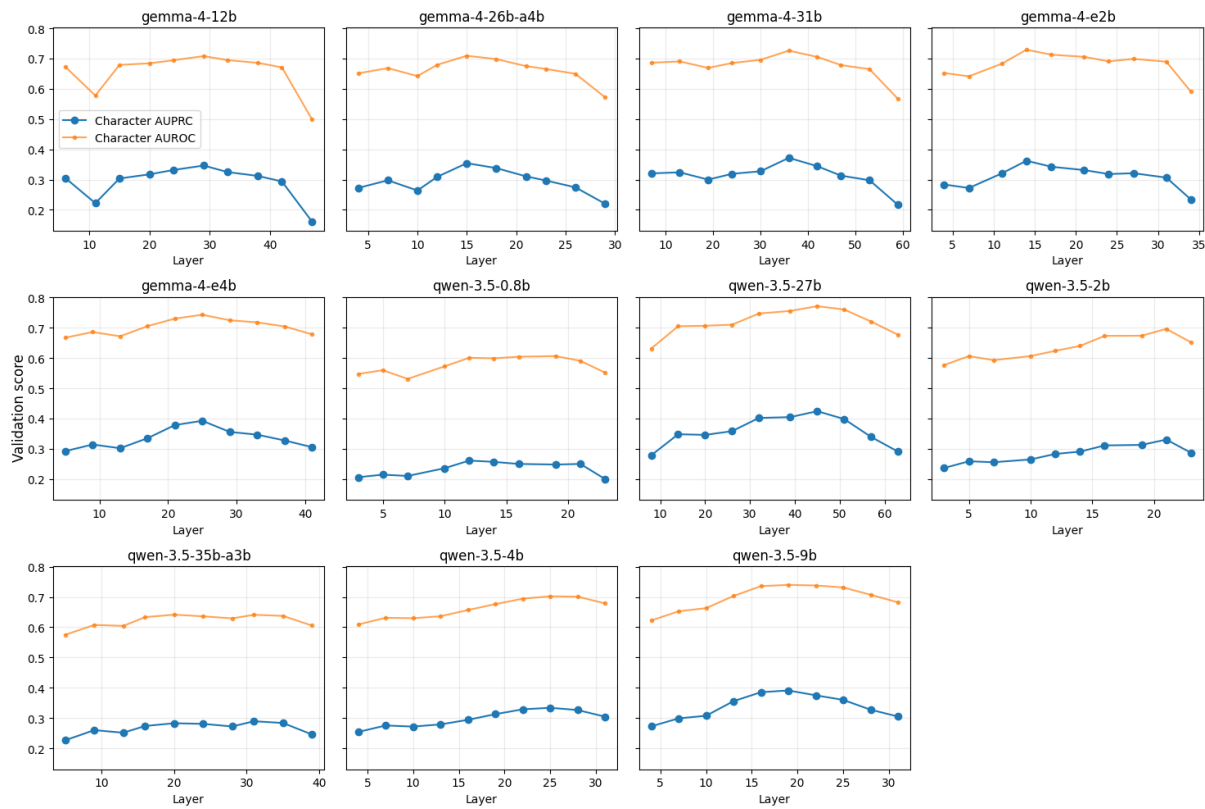


Figure 6: Validation Character AUPRC and AUROC across candidate backbones and decoder-block outputs. Selection used Character AUPRC; Qwen 3.5 27B at layer 45 achieved the highest value.