ON THE EMERGENCE AND EFFECTIVENESS OF USER PROFILING IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As an increasing portion of humans spend an increasing portion of their time conversing with and directing language models, they will be more subject to the implicit profiling language models are capable of. Understanding how these user profiling capabilities emerge both in the models themselves and during training is essential for privacy and mitigating bias, but existing studies are limited in scope and in realism. We show that language models represent the gender and age of the person writing, that this representation strengthens with scale and forms early in training. Training linear probes on 162k human written blog posts across 29 base models spanning 14M to 32B parameters shows consistent trend lines of about 3 (gender) and 4.4 (age) point increases in accuracy per tenfold increase in size, with a common slope across model families. In OLMo 2, most of the signal appears within the first 3% of pretraining and is mostly unchanged by midtraining and posttraining. Probes trained only on blogs transfer to real user conversations with AI assistants, and transfer improves with model size.

1 INTRODUCTION

Language models are trained on vast amounts of human-written text to predict the next token. Predicting what a person will write next is easier if you know something about who they are, so next-token prediction should reward models that represent the people producing the text (Andreas, 2022). As people increasingly interact with language models, we should expect the same to apply to them. This user profiling, the inference of characteristics of the user from context, has been shown. When asked directly, models will verbalise users' inferred age, gender and location accurately from ordinary text (Staab et al., 2024; Zaman & Garimella, 2026). This verbalisation is the product of internal representations formed even when the model is not directed to consider the question. One concern about the widescale adoption of language models as decision makers in society is whether those decisions are affected by these internal representations, and whether certain profiles and thereby certain users are unduly advantaged or disadvantaged.

As a step towards answering this question, we study the extent to which language models represent their users and when this capability emerges. Prior work on internal user representations has been limited in three ways. First, realism: the usage of synthetic conversations, generated by prompting a model to role-play a user of a given gender or age (Chen et al., 2024), which in some cases are closer to caricatures. Second, size: studies examine one to four instruction-tuned models (Chen et al., 2024; Bouchaud & Ramaciotti, 2025; Choi et al., 2025), so they cannot say how user modelling depends on scale or training. Third, probe-training data: datasets of real users with self-reported demographics are limited in size.

We aim to address these limitations with the Blog Authorship Corpus (BAC) (Schler et al., 2006). The BAC consists of real, self-reported author labels on posts written up to 2004 and, after filtering, possesses 229,648 posts from 7,341 authors. We train linear probes on the residual stream activations of 29 base models in 6 families, from 14M to 32B parameters, and on checkpoints spanning the full OLMo 2 training pipeline, taking from each the validation-selected layer. We contextualise the results of our probes against baselines of a TF–IDF classifier (lexical) and simply asking the model (verbal), and subsequently test whether they generalise to real user chats.

We find our probes beat the lexical baseline above a few hundred million parameters and, from a few billion parameters, roughly match the verbal one. We find that their ability improves with size – the

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.

best models identify an author's gender from their posts with around 90% balanced accuracy (79% from a single post) – and that this improvement is well described by a log-linear trend in parameters with a common slope across model families. Three findings describe how the capability emerges and works:

- **Models form the relevant representation in their middle and later layers.** Probes on the embedding layer perform about as well as TF–IDF and barely improve with size, indicating models pick up on more than direct lexical cues. After linearly removing what a TF–IDF classifier captures, the remaining signal for gender rises from AUROC 0.55 (Pythia-14M) to 0.78 (OLMo 2 13B).
- **Profiling capabilities are formed early in training.** Most of the gain arrives within the first 100B pretraining tokens (about 2–3% of the total), and author-level accuracy is then preserved to within about a point through mid-training, SFT, DPO and RL. Additionally, after pretraining the direction the probes find remains consistent (cosine similarity 0.91–0.95 between the end of mid-training and the end of RL), although it still rotates during pretraining.
- **Performance rises roughly log-linearly with context.** Within a post, the readout sharpens as the model reads: for Gemma 3 12B, gender AUROC rises from 0.54 after one token to 0.92 after 1,023, against 0.84 for a readout of the embedding layer, and the gap between the two grows with model size.

We also find that the probes transfer to real user conversations, with accuracy improving over the turns of a conversation. Notably, transfer is uneven across languages, and weakest in Hindi. To determine the extent to which the representations found by our probes are actively used by the model, we conduct two experiments. The first explores how verbalisation performance, directly asking the model to guess the characteristic, degrades under two methods which ablate the representation; we find that performance degrades somewhat, though remains above chance. The second explores how the model's verbalised answer varies as we increase or decrease the strength of the representation; we find that doing so moves the share of posts the model labels male from 1–2% to 87–99%.

Overall, our results suggest that models do form and hold latent representations of the characteristics of users, and that the quality of these representations improves with model size and the amount of user writing the model has seen. It remains to be seen whether these representations and profiling more generally affect model behaviour beyond explicit questions about the user.

2 Related work

Profiling authors and users from text. Inferring an author's gender or age from their writing is a long-standing task, from stylometric classifiers (Koppel et al., 2002; Argamon et al., 2003; Schler et al., 2006; Burger et al., 2011) to the PAN author-profiling shared tasks (Rangel et al., 2013; Rangel & Rosso, 2019). Language models do this when asked: they infer the age, gender and location of Reddit users from ordinary posts (Staab et al., 2024) and of real ChatGPT users from their histories (Zaman & Garimella, 2026), and their responses change with the identity they infer (Hofmann et al., 2024; Kantharuban et al., 2025; Poole-Dayan et al., 2026; Eloundou et al., 2025).

Internal representations of users. Demographic attributes of authors are decodable from text encoders (Elazar & Goldberg, 2018; Lauscher et al., 2022). For chat models, TalkTuner (Chen et al., 2024) probes LLaMA-2-Chat-13B on synthetic conversations generated by prompting a model to role-play a user of a given gender or age; Bouchaud & Ramaciotti (2025) and Choi et al. (2025) probe instruction-tuned models, and Choi et al. (2025) find that a decoder reads Llama-3.1-8B-Instruct's belief about a user's gender better than asking the model does.

Probes, scale and interventions. A linear probe shows that information is available to a linear readout; whether the model uses it needs an intervention (Alain & Bengio, 2017; Hewitt & Liang, 2019). Probe accuracy for other concepts grows with model size (Gurnee & Tegmark, 2024; Marks & Tegmark, 2024; Nordby et al., 2026), and fine-tuning can reuse rather than rewrite existing representations (Jain et al., 2024; Prakash et al., 2024). To test use, we remove probe directions from

the residual stream, following directional ablation (Arditi et al., 2024) and the least-change edit of LEACE (Belrose et al., 2023) (see also Ravfogel et al. 2020).

3 Experimental setup

3.1 Data and prediction tasks

We use the public release of the Blog Authorship Corpus (Schler et al., 2006), predicting each author's recorded gender (female or male) and age group (13–17, 23–27 or 33–47) from individual posts. The release keeps only these three age groups and omits the ages between them, which makes age easier to separate than a continuous scale. We strip HTML markup and collapse repeated whitespace, then retain only posts with at least 50 tokens under the Qwen2.5-0.5B tokenizer and authors with at least three such posts. We then randomly drop authors from the larger age groups so that each age group has the same number of authors, and split the authors 70/15/15 into training, validation and test sets, with the same age-group proportions in each split. All posts from an author belong to the same split.

The resulting splits contain 162,505 posts from 5,138 training authors, 30,787 posts from 1,101 validation authors, and 36,356 posts from 1,102 test authors. We select probes on 10,000 validation posts. Each post is scored separately, so prolific authors count more: the 10% most prolific test authors write 59% of test posts. Some text is shared between authors (group blogs, reposted feeds); removing these near-duplicates (11.4% of test posts) raises accuracy by 0.7–2.8 points and leaves the scaling trend unchanged (Appendix A.1).

3.2 Models and representations

Our model comparisons cover Pythia (Biderman et al., 2023), OLMo 2 (Team OLMo et al., 2025), Llama 3.1 (Grattafiori et al., 2024) and 3.2 (Meta, 2024), Gemma 3 (Gemma Team et al., 2025), Qwen2.5 (Yang et al., 2024), and Qwen3 (Yang et al., 2025). We use base models for the size comparisons, and intermediate checkpoints to examine changes during pretraining. Each model receives the post text directly with truncation applied past 1,024 tokens. We report model size as non-embedding parameters, and refer to models by their nominal sizes.

For a post of length T, let $h_{\ell,t}$ denote the representation at layer ℓ and token position t. Our primary readout uses the mean over non-padding positions,

$$x_\ell = \frac{1}{T}\sum_{t=1}^{T} h_{\ell,t}. \tag{1}$$

We extract representations at every fourth layer (every eighth for Qwen2.5-32B), plus the embedding and final layers. Intermediate representations are taken after the corresponding block; the final representation is taken after the model's final normalisation.

3.3 Probe training and selection

We fit a separate ridge probe for each task and sampled layer. Features are standardised using training-set means and standard deviations, with a floor of 10^{-3} times the median feature standard deviation to avoid amplifying near-constant features. For standardised representations z_i and one-hot labels y_i, we minimise

$$\frac{1}{n}\sum_{i=1}^{n}\|W^\top z_i + b - y_i\|_2^2 + \lambda\|W\|_F^2, \tag{2}$$

where the intercept b is unpenalised. We select $\lambda \in \{0.01, 0.1, 1, 10, 100\}$ and then choose a layer by validation balanced accuracy, separately for each model and task, then evaluate the selected probe on the test set. For layer profiles, regularisation is selected separately at each layer.

3.4 Evaluation and baselines

Our primary metric is post-level balanced accuracy: the mean recall across the task's classes. This gives each class equal weight, though it does not give each author equal weight. Author-level ac-

curacy instead makes one prediction per author, from the probe's scores averaged over all of that author's test posts. We report 95% intervals from 1,000 bootstrap resamples: each resample draws test authors with replacement and keeps all of an author's posts together, since posts by one author are not independent. The probe is fitted and selected once, so the intervals reflect only which test authors happened to be sampled.

We compare against a TF–IDF classifier (Spärck Jones, 1972) (word counts weighted by rarity; 50,000 features, sublinear term frequency, logistic regression with regularisation chosen on validation data), embedding-layer probes and probes on randomly initialised models. These provide reference points for lexical prediction, information present before contextual processing, and what mean-pooling a network with no training already captures, respectively. The lexical baseline uses full post text, whereas language-model inputs are truncated to 1,024 tokens; truncating the baseline's input in the same way leaves its accuracy unchanged. We also test the blog probes, without refitting, on 10,000 Reddit "Am I the Asshole" posts (AITA)[1] whose authors state their gender in the post (the tag is masked), and on the human–AI conversations of Section 7. Protocols specific to reading trajectories, training comparisons and transfer are described with those experiments.

4 Scale and Depth

Larger models encode their users better, in every family we test. Post-level balanced accuracy ranges from 64.4% to 79.3% for gender and from 56.4% to 78.6% for age (Figure 1). Within each family, accuracy rises roughly linearly in log parameters, and the families rise at similar rates (Pythia, rising more slowly on gender, is the exception): a single slope with a separate intercept per family explains most of the variance, 3.06 points per tenfold increase in parameters for gender ($R^2 = 0.952$) and 4.43 for age ($R^2 = 0.991$). Families differ more in level than in slope, by 3–4 points at a given size. Across models, the offsets do not track release date or training tokens per parameter; across the six families, those trained on more tokens per parameter sit lower, but six points cannot separate this from other family differences (Appendix A.3). We attribute the offsets to differences in training data and recipe that we cannot separate. Size is confounded with recipe within each family too (Llama 3.2 1B and 3B are distilled from Llama 3.1), so the slope describes existing models and is not an estimate of the causal effect of parameter count.

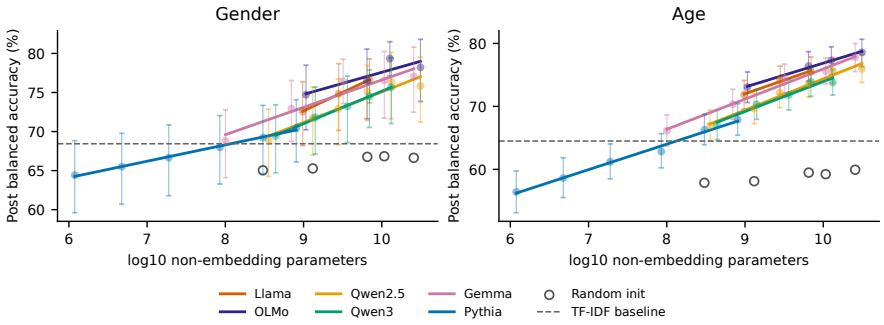


Figure 1: Post-level balanced accuracy against model size, with 95% author-clustered intervals and per-family fits. Open circles: randomly initialised networks; dashed line: TF–IDF.

Larger models capture more than word counts. Probes on the embedding layer do about as well as TF–IDF (64.9–69.2% against 68.4%) and barely improve with size. The best layer's advantage over the embedding layer is near zero below 300M parameters and reaches 10.4 points for OLMo 2 13B (Figure 2, right). Is this extra accuracy information that a word-count classifier already has? On the 9,402 de-duplicated test posts that TF–IDF gets wrong, probe accuracy rises with size from 30% (Pythia-14M) to 58% (OLMo 2 13B), so larger models get right many posts that word counts get wrong. The same holds when we remove from each post's probe score the part that is linearly predictable from the TF–IDF score: what is left separates male from female authors with AUROC 0.55 for Pythia-14M and 0.78 for OLMo 2 13B (correlation with log size 0.95).

[1] https://huggingface.co/datasets/OsamaBsher/AITA-Reddit-Dataset

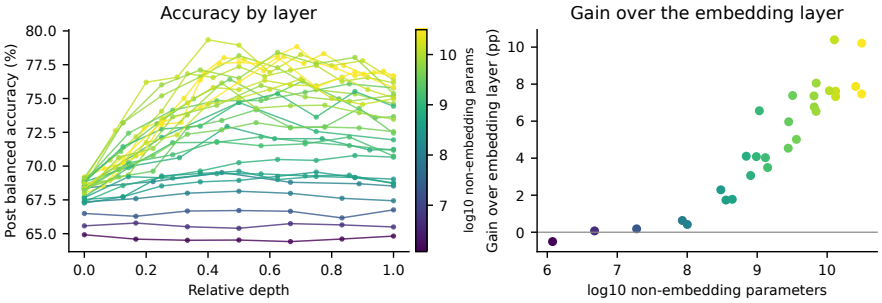


Figure 2: Left: gender balanced accuracy at each probed layer against relative depth (0 = embeddings), coloured by size. Right: gain of the selected layer over the embedding layer.

5 Where it comes from in training

The size comparisons establish that author attributes are more readily recovered from larger models. A separate question is how they come to encode these attributes. To investigate this, we follow OLMo 2 at 1B, 7B and 13B through pretraining and the subsequent stages of training. We fit a new blog probe at each checkpoint, holding the layer fixed at the final stage-1 checkpoint's validation-selected layer and reselecting λ on validation data. After pretraining the probe direction barely changes: its cosine similarity between the end of mid-training and the RL checkpoints is 0.91–0.95.

Most of the observed improvement occurs early (Figure 3). For OLMo 2 7B, author-level gender balanced accuracy rises from 70.4% after 1B training tokens to 85.5% after 114B, compared with 88.1% at the end of stage 1, after 3.9T tokens. Thus, 85% of the gain from the earliest measured checkpoint to the final one occurs within the first 3% of pretraining. The 1B and 13B models follow the same pattern.

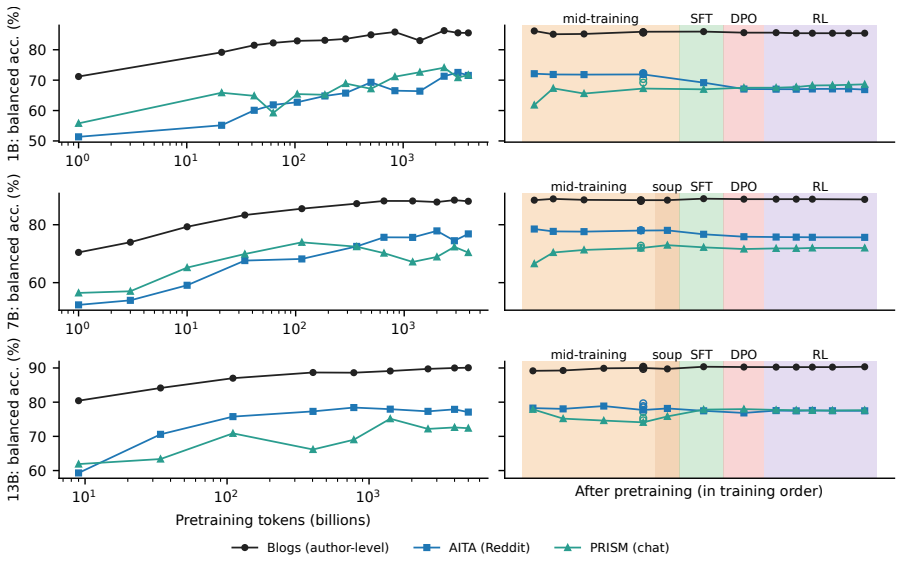


Figure 3: Author gender decodability across OLMo 2 training. Left: stage-1 pretraining checkpoints. Right: later checkpoints in order, shaded by stage; hollow markers are other mid-training runs. Lines: blog accuracy (author level) and transfer to AITA and PRISM (post level).

Subsequent training changes blog decodability relatively little. At 7B, author-level accuracy is 88.5% after mid-training and weight averaging, and 88.7% after RL. At 13B, the measured checkpoints after stage 1 lie between 89.2% and 90.4%.

Transfer is less stable. The blog-trained probes' AITA balanced accuracy falls through SFT and DPO, from 78.1% to 75.9% at 7B and from 71.9% to 67.1% at 1B, and RL leaves it unchanged; with 10,000 AITA posts, these drops are several times their sampling error. At 13B the drop is smaller (78.2% to 76.9%). PRISM does not show a comparable decline. PRISM accuracy also moves by several points between neighbouring checkpoints even within pretraining (62–75% at 13B, where 1,374 users give a sampling error of about 1 point), so we do not interpret its smaller changes across stages. Stable performance on blogs can therefore coexist with changes in how well the same training procedure yields a probe that transfers to other kinds of writing.

6 How the readout builds while reading

Our results up to now have been from models which have seen the entirety of posts (or up to truncation). We next ask how this changes as more of the text becomes available. We hold the trained probe fixed and apply it to each prefix of a post. We then evaluate gender AUROC at each prefix length.

For each of 23 models, we use a fixed cohort of test posts containing at least 1,024 tokens under that model's tokenizer. This gives 1,612–1,664 posts per model, from about 350 authors. With so few authors and because posts by the same author are correlated, early values are noisy.

The readout becomes substantially more accurate as the model reads (Figure 4). For Gemma 3 12B, AUROC rises from 0.54 after the first text token to 0.80 after about 70 and 0.92 after 1,023. Across the models, final AUROC ranges from about 0.78 to 0.92.

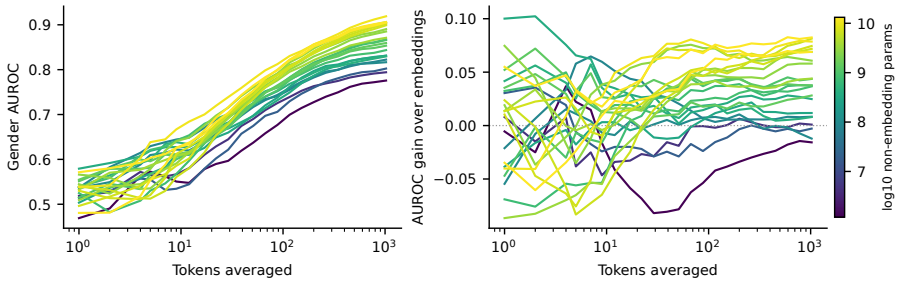


Figure 4: Left: gender AUROC of the probe on progressively longer prefixes, one line per model, coloured by size. Right: advantage over the embedding-layer readout.

More text also improves a probe on the embedding layer, which can accumulate word identities without contextual processing. We therefore compare the two readouts on the same prefixes. For Gemma 3 12B, the embedding readout reaches 0.84 at the end, compared with 0.92 for the contextual readout. The contextual advantage is small or absent in the smallest models and becomes larger with scale. Four randomly sampled individual trajectories are shown in Appendix A.4.

7 Transfer to real conversations

We next ask whether those same probes can recover information about people just from their chat conversations with AI. We evaluate them on three datasets with recorded gender labels: PRISM (Kirk et al., 2024), ThoughtTrace (Jin et al., 2026), and Community Alignment (Zhang et al., 2025). These contain 1,374, 997 and 3,343 users with a recorded gender. In PRISM and ThoughtTrace users report their own gender; Community Alignment's users are paid annotators writing in five languages, whose gender comes from the annotation vendor's records. The assistant replies in these transcripts come from the chat models each dataset collected, not from the models we probe.

We read each user in two ways. For the user's history, we join all of the user's turns from all of their conversations into one text, as an assistant with memory might see them. For a single conversation, we keep the assistant's replies in the text but average the probe only over the user's tokens. The text is formatted with plain "User:" and "Assistant:" labels, since the models are base models without chat templates.

The blog probes transfer to both settings, although performance is lower within a single conversation. For Gemma 3 12B, user-history AUROC is 0.84 on PRISM, 0.75 on ThoughtTrace and 0.74 on Community Alignment. The blog-trained TF–IDF classifier, applied to the same user histories, reaches only 0.71, 0.64 and 0.59. The corresponding single-conversation scores are 0.67, 0.68 and 0.58. Reading additional user turns improves the readout (Figure 5). On conversations with at least four user turns, Gemma 3 12B rises from 0.60 to 0.69 on PRISM and from 0.59 to 0.69 on ThoughtTrace between the first and fourth turns. Each curve follows the same conversations at every turn: those with at least four user turns that fit within the model's input length. Larger models transfer better on matched conversations too: on the conversations shared by all 17 Pythia and Qwen models, AUROC after four turns correlates with log size at 0.87 on PRISM and 0.90 on ThoughtTrace.

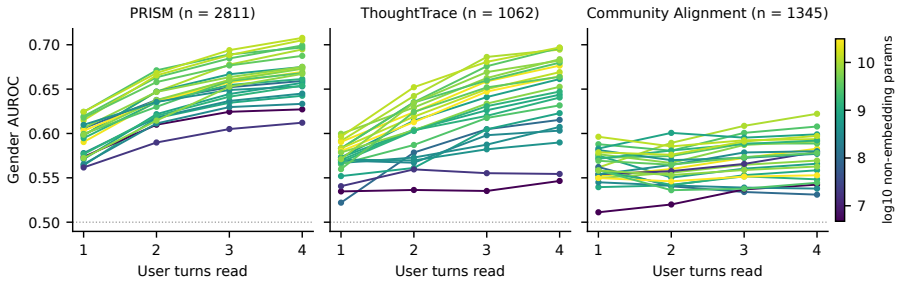


Figure 5: Gender AUROC of blog-trained probes after the user's first k turns of a real conversation, read on user tokens only; 23 base models, coloured by size. The number of conversations (n) varies by model.

Transfer is uneven across languages. Gemma 3 12B's user-history AUROC on Community Alignment is 0.83 in Italian, 0.78 in Portuguese, 0.76 in French, 0.74 in English and 0.61 in Hindi; the English-trained TF–IDF classifier gets 0.65 in English but only 0.53–0.59 in the other languages and 0.47 in Hindi. Hindi single-conversation performance is at chance.

8 Do models use what they encode?

A successful probe establishes that information is recoverable from activations. It does not, by itself, establish that the model uses that information. We approach this distinction in two steps: asking models to report the writer's gender, then testing whether changing the probed activations changes their answers.

Asking and probing. We compare the probe's gender margin with the model's $\log P(\mathrm{M}) - \log P(\mathrm{F})$ after the prompt "The gender of the person who wrote this post (M/F):". Both are evaluated by AUROC on the same 2,000 test posts. AUROC asks only whether posts by men get higher scores than posts by women, so it is unaffected by a model's overall preference for one answer letter.

Small models often encode substantially more than they report (Figure 6). Without examples, small models rarely answer with either letter: every Pythia model puts about 5% or less of its next-token probability on M and F together, so their zero-shot scores compare two unlikely tokens. With four examples, M and F take 85–100% of the probability in every model from 160M parameters up. Pythia models from 14M to 1B then score 0.47–0.52, against 0.75–0.81 for their probes on the same posts. By 7–14B, several models' answers match or exceed their probes: with four examples, Llama 3.1 8B scores 0.88 against its probe's 0.86, Gemma 3 12B scores 0.90 against 0.88, and OLMo 2 13B (stage 1) scores 0.89 for both. The small-model gap is therefore a failure of elicitation under the tested prompts, rather than evidence that the attribute is absent from their activations.

Examples can change both ranking and the decision threshold. Gemma 3 12B has AUROC of about 0.90 with either zero or four examples, while its balanced accuracy rises from 72% to 82%, because AUROC ignores where the decision threshold sits: without examples it answers M for 76% of posts, although 52% are by men, and with four examples for 56%. Qwen2.5-7B, by contrast, improves in AUROC, from 0.78 to 0.83. The comparison across models allows 1,024 tokens for the whole

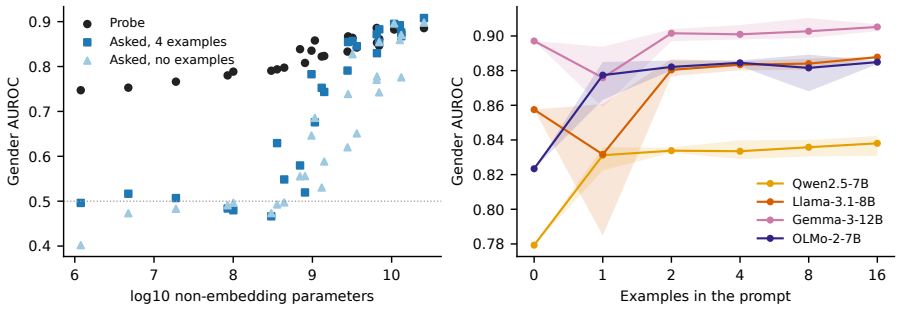


Figure 6: Left: gender AUROC of the probe and of asking the model, on the same 2,000 test posts (27 base models). Right: asking with k examples for four 7–12B models; bands span three example sets.

prompt, so with four examples the end of the target post is cut off in 11–12% of prompts; the example-count sweep allows 4,096 tokens and cuts almost none.

Removing the readout. We next intervene on the target post's activations and measure the model's four-shot gender prediction on 600 test posts. Let w be the direction the saved probe reads gender along (the difference between its male and female weight vectors), with mean μ and covariance Σ estimated from training-post activations. We compare orthogonal projection with a covariance-weighted edit, related to linear concept erasure (Belrose et al., 2023):

$$x \leftarrow x - \frac{\Sigma w\, w^\top (x - \mu)}{w^\top \Sigma w}.$$

Both edits give every edited token the same probe score, so the probe can no longer tell men from women. They differ in which way they move the activation: orthogonal projection moves it along the probe's weights w, while the covariance-weighted edit moves it along Σw, approximately the difference between men's and women's average activations (Haufe et al., 2014). Neither guarantees removal of all linearly decodable gender information. We apply each sampled layer's edit through the intervening blocks up to the next sampled layer, excluding the final block output. Only target-post tokens are edited; the examples, question and beginning-of-sequence token are left intact.

The covariance-weighted edit reduces prompted AUROC in all six tested Qwen and Llama models, by 0.06–0.09, while next-token loss on raw posts rises by about 0.02 nats, under 1% of the roughly 3-nat baseline (Figure 7). For example, Llama 3.1 8B falls from 0.88 to 0.78 and Qwen2.5-14B from 0.86 to 0.78. Orthogonal projection lowers AUROC by 0.02–0.03. Applying the same edits to the age probe's oldest-minus-youngest direction, or to three random directions, changes gender AUROC by about 0.01 or less, although these edits raise the loss by a similar amount or more (0.02–0.07 nats, averaging the three random directions). These comparisons support a contribution of the probed information to explicit gender predictions.

Steering the report. We also add the activation pattern Σw at approximately 0.4 relative depth, scaling the shift by the standard deviation of token activations along that direction. Across the six Qwen and Llama models, the fraction of the 600 posts assigned M changes from 34–58% without an edit to 1–2% at $\alpha = -4$ and 87–99% at $\alpha = +4$, whereas shifting the same distance along three random directions moves it by a median of 7 points and at most 23, against a median of 46 for the gender direction (Figure 8). The random shifts also disrupt the model less (next-token loss rises by 0.02–0.15 nats, against 0.11–0.36 for gender), so part of the gender effect could come from a larger disruption. Here, assignment means that M has higher next-token probability than F. The interventions show that the probed activations can affect an explicitly requested attribute prediction. Whether the same information shapes an ordinary reply, recommendation or decision remains a separate question.

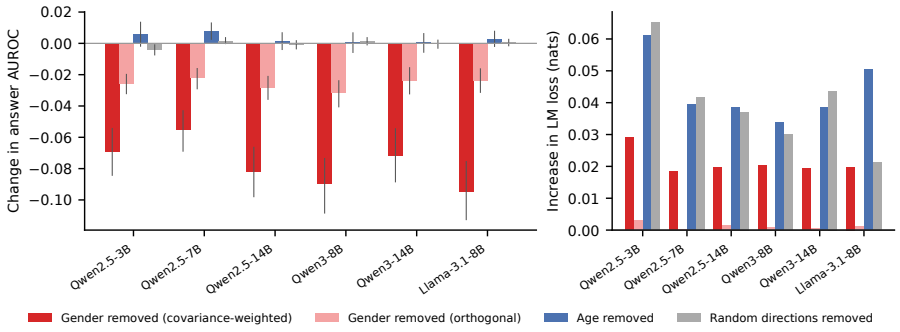


Figure 7: Removing the saved probes' gender readouts and asking the model (four-shot M/F prompt, 600 test posts from 525 authors; Qwen and Llama). The age and random-direction controls use the covariance-weighted edit (random: mean of three directions). Left: change in AUROC of the model's answer from the unedited model, with paired 90% intervals from resampling authors. Right: increase in next-token loss on raw posts. Gemma 3 is not shown: there, edits along random directions were themselves destructive.

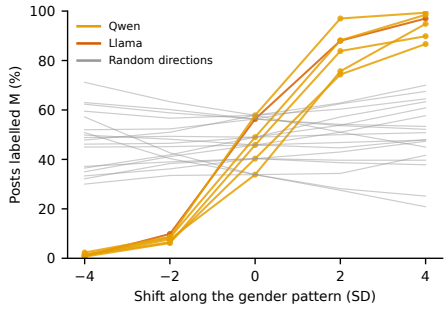


Figure 8: Share of 600 test posts labelled M when the post's activations are shifted along the gender pattern (coloured, one line per model) or three random directions (grey).

9 Discussion and limitations

Our experiments show that language models encode user characteristics and that this encoding becomes richer with both increased model size and increased context. Its transfer from blogs to conversations indicates that there is some general representation of these characteristics across settings, though this transfer is somewhat limited. It is difficult to apportion this limitation between failures of the probed model itself and failures in our probing setup.

Ablating and interfering along the probed readout has specific effects (relative to random directions) on relevant tasks. This serves as a proof of concept that these representations are actively used by the model. It is yet to be seen how extensive this usage is. We have not yet established whether these representations have more diffuse effects in shaping recommendations or judgements. Establishing when such information affects consequential decisions remains an important next step.

Several limitations qualify the results. The probes predict the corpus's binary gender labels and selected age groups, rather than a person's identity in any broader sense. These labels correlate with topic and register, and performance beyond our TF–IDF baseline does not rule out other lexical explanations. Model size is entangled with training data and recipe, and width and depth are correlated, so their relationships with accuracy are descriptive. The blog corpus is old public web text and may occur in pretraining data; author-disjoint probe splits do not rule this out. Finally, the conversation experiments read existing transcripts with base models and plain role labels.

AI USE STATEMENT

We used LLM-based assistants (Claude and Codex) to implement and review analysis code, run experiments, clean and reformat datasets, give feedback on experimental design, help interpret results, create figures, search and summarise the literature, format references, and edit drafts for readability. We have not used generative AI tools to generate synthetic data. The authors take full responsibility for the content of the paper.

ETHICS STATEMENT

This work measures how well language models infer the gender and age of the people whose text they read, which is a privacy risk in its own right: the information is present in the activations of small models that do not report it under our prompts, and it transfers to real users of AI assistants. We use only existing public datasets (the Blog Authorship Corpus, Reddit AITA posts, PRISM, ThoughtTrace and Community Alignment) under their licences, collect no new data, and report aggregate results; example text shown in the appendix is excerpted and names are redacted. The labels are self-reported or vendor-held binary gender and binned age, which do not capture gender identity and may encode the stereotypes the probes pick up. Decoding methods like ours could be used to profile users.

REPRODUCIBILITY STATEMENT

All models are public checkpoints; the Hugging Face revision of each is recorded in the released run configurations. Section 3 and the appendix give the splits, tokenisation, pooling, probe training and selection, and every baseline. The anonymous code release contains the probe-training pipeline, every analysis and figure script, and the CSV behind each figure; seeds are fixed throughout.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.423. URL https://aclanthology.org/2022.findings-emnlp.423/. arXiv:2212.01681.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.11717.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346, 2003. doi: 10.1515/text.2003.014.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023. arXiv:2306.03819.

Stella Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, 2023.

Paul Bouchaud and Pedro Ramaciotti. Linear socio-demographic representations emerge in large language models from indirect cues. *arXiv preprint arXiv:2512.10065*, 2025. URL https://arxiv.org/abs/2512.10065.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309, Edinburgh, Scotland, UK, 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1120/.

Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational ai. *arXiv preprint arXiv:2406.07882*, 2024.

Dami Choi, Vincent Huang, Sarah Schwettmann, and Jacob Steinhardt. Scalably extracting latent representations of users. Transluce blog, 25 November 2025, 2025. URL https://transluce.org/user-modeling.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002.

Tyna Eloundou, Alex Beutel, David G. Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. First-person fairness in chatbots. *arXiv preprint arXiv:2410.19803; ICLR 2025*, 2025.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, 2019. URL https://arxiv.org/abs/1909.03368.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154, 2024. doi: 10.1038/s41586-024-07856-5.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=A0HKeKl4Nl.

Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, Shayne Longpre, Hongxiang Gu, Maximillian Chen, and Tianmin Shu. ThoughtTrace: Understanding user thoughts in real-world LLM interactions. *arXiv preprint arXiv:2605.20087*, 2026. URL https://arxiv.org/abs/2605.20087.

Anjali Kantharuban, Jeremiah Milbauer, Maarten Sap, Emma Strubell, and Graham Neubig. Stereotype or personalization? user identity biases chatbot recommendations. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.1254/.

Hannah Rose Kirk et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *NeurIPS Datasets and Benchmarks; arXiv:2404.16019*, 2024.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002. doi: 10.1093/llc/17.4.401.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. URL https://aclanthology.org/2022.emnlp-main.539/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2310.06824.

Meta. Llama 3.2 model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B.

Erik Nordby, Tasha Pais, and Aviel Parrack. Linear probe accuracy scales with model size and benefits from multi-layer ensembling. *arXiv preprint arXiv:2604.13386*, 2026.

Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. LLM targeted underperformance disproportionately impacts vulnerable users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2406.17737.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2402.14811.

Francisco Rangel and Paolo Rosso. Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in Twitter. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller (eds.), *CLEF 2019 Labs and Workshops, Notebook Papers*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL https://ceur-ws.org/Vol-2380/paper_263.pdf.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at PAN 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis (eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL https://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *ACL*, 2020. arXiv:2004.07667.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *ICLR*, 2024.

Team OLMo et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

S M Mehedi Zaman and Kiran Garimella. Inferential privacy leakage in anonymized conversational AI logs. *arXiv preprint arXiv:2605.23820*, 2026. URL https://arxiv.org/abs/2605.23820.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*, 2025. URL https://arxiv.org/abs/2507.09650.

A SUPPORTING EXPERIMENTS

A.1 DUPLICATE AND FEED-AUTHOR SENSITIVITY

Splitting by author does not stop the same text appearing under several authors: group blogs, reposted feeds and copied articles all do this. We find near-duplicate posts with MinHash (five-word shingles, 64 permutations, 16×4 LSH bands) and keep pairs whose signatures agree on at least 70% of their entries. We then flag three things: posts in near-duplicate clusters that span more than one author, authors with at least half of their posts in such clusters, and two feeds that are not written by the named author (one posting copied casualty notices, with 1,002 test posts, and one a syndicated newspaper column). This flags 11.4% of test posts, from 339 authors, and 8.6% of validation posts.

Duplicates hurt rather than help, because the copies usually carry different authors' labels. Removing them raises post-level gender balanced accuracy by 2.2–2.8 points in most models. The exceptions are Gemma-3-4B and Llama-3.1-8B (+1.7) and OLMo 2 13B (+0.7), which had done unusually well on the casualty-notice feed. Author-level accuracy moves by -0.2 to $+0.6$ points. We rescored 27 of the 29 models (not OLMo 2 1B or 32B), and re-selecting the layer and λ on de-duplicated validation data changes none of the results that depend on selection. Training posts are not de-duplicated; the probes were trained on them.

A.2 CONTROL TASKS

Probes for the author's star sign, which the text should not predict, stay at chance in every model (test balanced accuracy 7.1–9.2%, chance 8.3%, 29 models), with no trend in size. Within each of the five largest occupations in the corpus (41–229 test authors each), gender accuracy still rises with size (2.6–4.4 points per decade, correlation with log size 0.83–0.95, 27 models), so the size trend is not only a shift in which occupations are recognised.

A.3 SCALING SHAPE: WIDTH, DEPTH, FAMILY OFFSETS AND DIRECTION STABILITY

Width against depth. Across the 28 models other than OLMo 2 32B, a plane in log width and log depth fits post-level gender accuracy well ($R^2 = 0.89$), with $+8.3$ points per decade of width and only $+0.9$ per decade of depth (Figure 9). Width and depth rise together within families (correlation 0.84), so this split is only suggestive. Where depth is fixed, more width always helps: Pythia 14M to 31M to 70M at 6 layers gains 1.1 and then 1.2 points, Qwen3 0.6B to 1.7B at 28 layers gains 2.4, Qwen2.5 1.5B to 7B at 28 layers gains 3.4, and Qwen3 4B to 8B at 36 layers gains 1.3. The one pair with the same width and more depth, Qwen2.5 14B to 32B (48 to 64 layers), gains nothing (-0.5), but it sits at the top of the size range, where the trend is least clear.

Family offsets. Some families sit higher than others at the same size. Across the 28 models, a model's distance from the plane does not track its release date ($r = 0.23$) or its training tokens per parameter ($r = -0.10$). Averaged within each family the correlations are 0.34 and -0.96, so families trained on more tokens per parameter sit lower, but six families are too few to separate this from any other difference between them. Qwen3, the newest family and the one trained on the most tokens (36T), sits lowest (-1.1 points); OLMo 2's stage-1 checkpoints ($+1.8$) and Gemma 3 ($+1.1$) sit highest.

Direction across layers. We take the probe's gender direction at each layer, the difference between its male and female weight vectors in the model's own coordinates, and compare it between layers (Figure 10). Directions at neighbouring probed layers have mean cosine similarity between 0.16 (Llama-3.2-1B) and 0.58 (Gemma-3-27B), and the mean over all pairs of layers is 0.09 (Pythia-14M) to 0.27. Probed layers are four apart (eight for Qwen2.5-32B), so "neighbouring" covers different fractions of depth in different models. Ridge weights are shaped by how the activations vary, so these numbers probably understate how stable the underlying difference between men's and women's activations is. Section 5 compares OLMo's directions across training checkpoints instead.

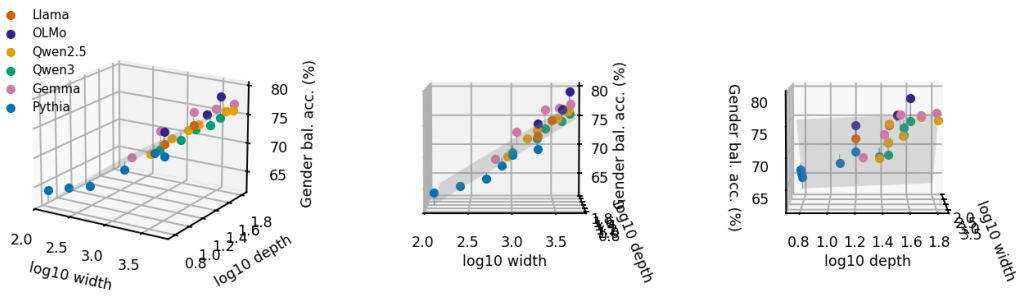


Figure 9: Post-level gender balanced accuracy against log width and log depth, with the least-squares plane, from three angles; vertical lines are residuals.

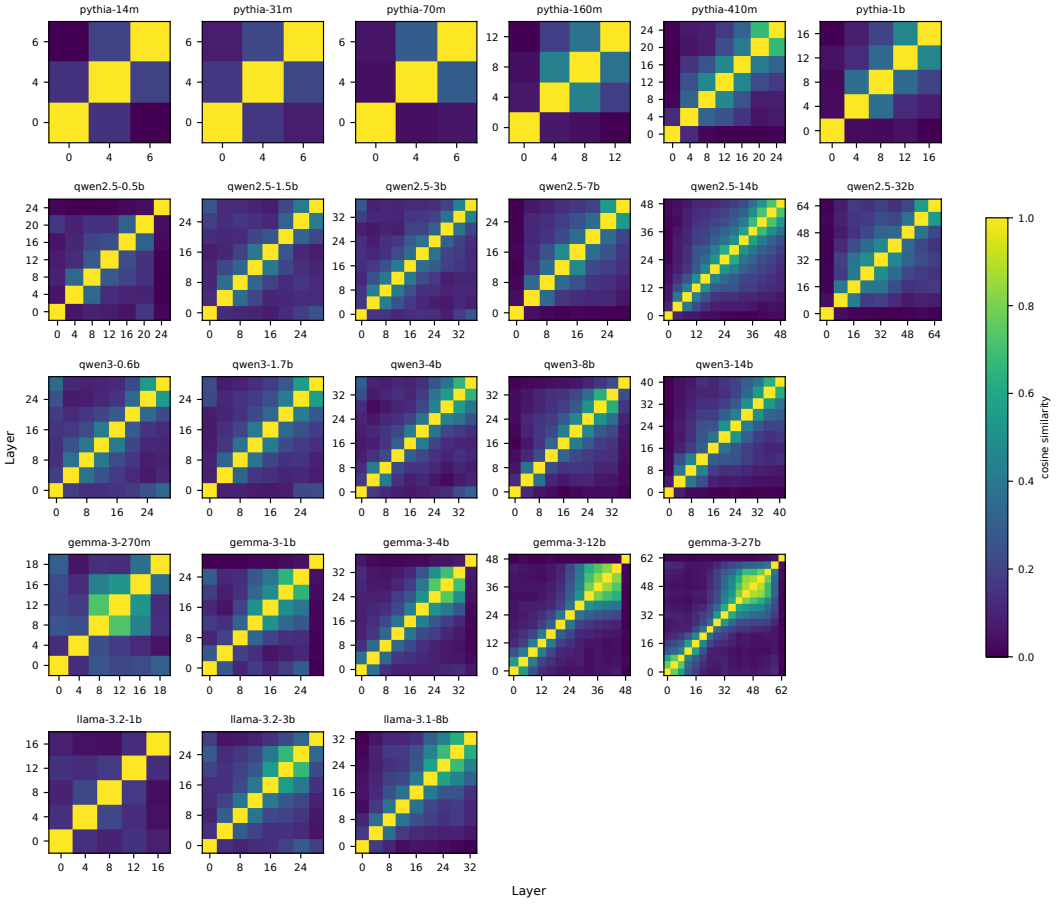


Figure 10: Cosine similarity between gender probe directions at every pair of probed layers, one panel for each of the 25 non-OLMo models.

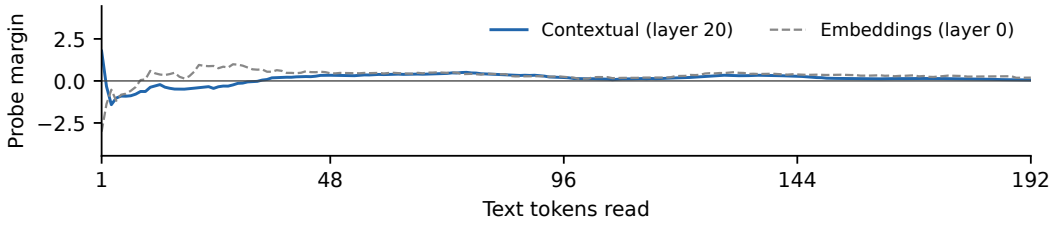
A.4 FOUR RANDOMLY SAMPLED READING EXAMPLES

We sample two male-labelled and two female-labelled posts without replacement from the 1,889 non-excluded posts in the saved 2,000-post BAC test subset, using NumPy PCG64 seed 42 after sorting by text identifier. The display order is M/F/M/F. Selection does not depend on probe scores, correctness or the shape of the trajectory. We show at most the first 192 text tokens of each selected post.

The curves show the running mean of the raw male-minus-female probe score for Gemma 3 12B: the validation-selected layer (20, $\lambda = 0.01$) and the embedding-layer control ($\lambda = 0.01$). BOS is present in the model input but excluded from the running mean. Positive values favour the male label; these scores are not calibrated probabilities. Axes are shared across examples. A blog name is redacted in the displayed text only; scores use the original text.

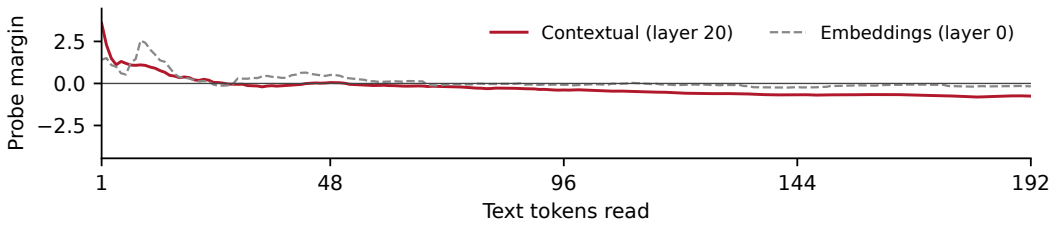
Example 1: recorded label M.

> Gloria is the headmistress of the School of Double-Talking for the Gullible and Wretched Hobbits of the Lower Earth (apologies to Tolkien fans for the allusion). Come to think of it, whether she steps down or not in 2007, it would still be a win-win situation for her. That is, to go down in history as the ONE who bound us all to the Mt. of Doom for the longest time ever, if not forever. Consider this: Scenario 1: Gloria steps down as President in 2007 as a result of the shift to a parliamentary system. That was her campaign promise last May, and this time, she is keeping her word. Rumors would have it that through The Cardinal's divine intervention, the Holy Yellow Ghoul appeared before her, cleansed her heart and changed her mind. Hallelujah! But no, don't bring out the champagne yet. As [excerpt ends]



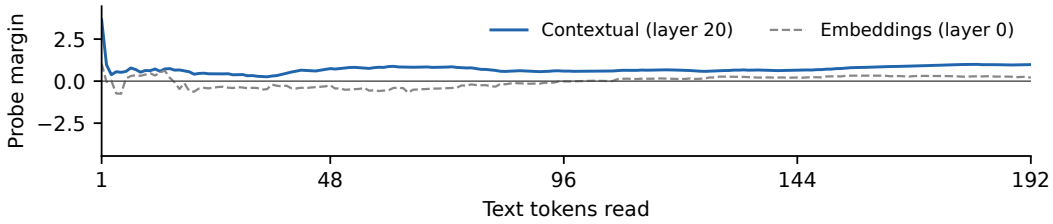
Example 2: recorded label F.

> well, today at work was pretty good during the day. i was so busy i didn't have time to think. all of a sudden it was 3pm, and all sorts of shit started hitting the fan. changes are coming in my office. and they are because of me... well, we'll see what my horoscope has to say about it... "The fire burns in your eyes again, and you possess a certain spring to your stride. You are able to laugh about anything that comes your way. Accept life as a joyous jumble and go about living it. Your sense of humor is your best means of transportation, taking you from delight to delight. Don't just ask how, but inquire as to why something is peculiar. This is a great time for brainstorming and innovations." well that whole fire burns in your eyes thing is right...pink eye and all. i normally do laugh at anything that comes my way...and i [excerpt ends]



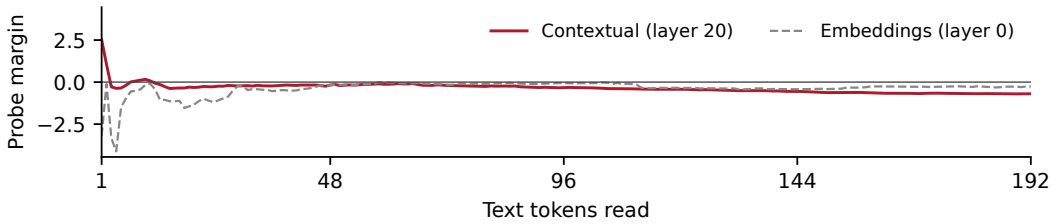
Example 3: recorded label M.

> When I was prompted to give the name for this blog I typed down [blog name]. Perhaps there are those of you who are the type who like to know why somethings have the names they do. Like why is the Williamsburg Bridge in NYC named "The Williamsburg Bridge". Or why the McDonalds red-haired dude is named 'Ronald' instead of 'Weird-scary-clown-dude'. Whatever, I'm going to give you the insight to why this blog is named the way it is. Probably soon before I had started to created this blog I had thought of the many things that may kill us Americans in the years to come. I thought about urlLink near earth objects , urlLink famine , urlLink AIDS , urlLink multiple wars and unrests , urlLink global warming , urlLink environmental degradation , urlLink the end of fossil fuels and plain old stupid people. Oh yeah and don't forget todays popular favorite [excerpt ends]



Example 4: recorded label F.

> (((Sigh))) Well, here it is a Holiday weekend & I thought I had plans...several diffrent ones actually, but a couple fell through & the last one (which is the one I was lookin the most forward to doin) isn't lookin to well either. But I'm still hopin'. I can't really beleive I'm gonna lay this out there like this...but..hey..I guess that is what this blog thing is all about. But right now..& quite frequently, I feel soooo alone. None of my friends would ever guess that I feel this way. (I tend to cover it up pretty well) See, I'm one of those that doesnt like to let people see what some people call, the weak side. But I'm not sure really where the lonliness comes from. It started gradually when I met someone. & only GOD knows why I am attracted to [excerpt ends]



A.5 Distribution of probe scores

Per-post scores separate steadily with size: Cohen's d between male and female authors' standardised scores is 0.65 for Pythia-14M, 1.1–1.3 at 1–3B and 1.3–1.5 at 7B and above (post-level AUROC 0.69 to 0.85); averaging each author's posts, d rises from 1.2 to 2.5 (AUROC 0.80 to 0.95). The randomly initialised Qwen2.5-7B sits at the level of Pythia-14M ($d = 0.73$).

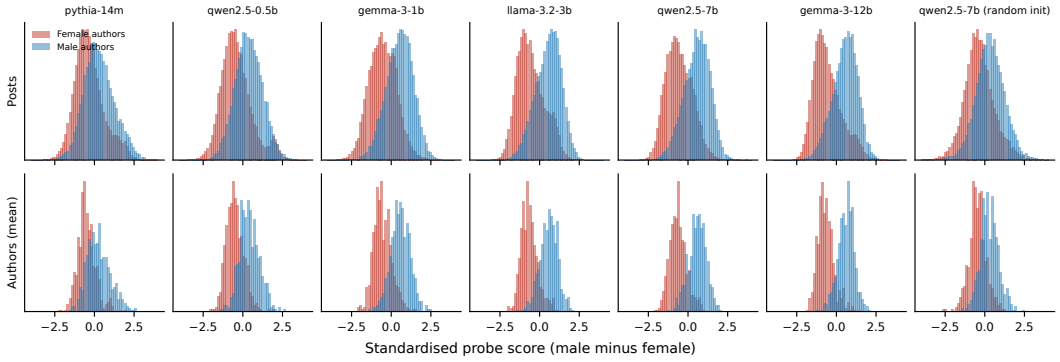


Figure 11: Standardised gender probe score for each test post (top) and averaged over each author's posts (bottom), by recorded author gender; the rightmost panel is a randomly initialised Qwen2.5-7B.