MANY IN ONE: GROUNDING INDIVIDUALITY IN A LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can produce many apparent personas, but this does not establish that their differences are grounded in the individuals they represent. Names, biographies, or demographics may retrieve associations learned during pretraining, and such associations need not track measured differences between actual people. We ask: can a model learn person-specific variation from observed behaviour alone, and can we read that variation back? Because natural speech entangles content with manner, we study one isolable dimension of grounded individuality, *register*: how speakers formulate given content. We introduce *addresses*, one learned speaker embedding per person, stored as a row of the model's input-embedding matrix and fitted by paraphrase inversion. We fit addresses for 640 speakers within one Qwen2.5-7B model, on 6.7 million training words from a new voice-verified corpus of public-figure podcast speech. We evaluate generated personas on human-calibrated register axes that separate grounding, diversity, and coverage. Addresses recover person-specific register, including in episodes and shows held out from training. A speaker's own turns ground personas best, and on held-out episodes a name, prompted or written in the address's place during fine-tuning, grounds them better than an address alone. Addresses yield more distinct styles and more collective variety of ideas than a prompted name, and a name written in their place during fine-tuning yields more of both. Because addresses are parameters, we read, steer and extend them. These results show that a single language model can represent grounded person-specific variation that we can measure and interpret.

1 INTRODUCTION

Individuality in language is the systematic variation between speakers that persists across topics. A persona is *grounded* when the language it produces has the measured properties of its person's own speech. A single language model already produces many apparent personas: a name, an occupation, or a survey respondent's profile placed in the prompt changes its output (Park et al., 2023; Argyle et al., 2023; Shanahan et al., 2023). Prior work finds that models write with less variety than people and resemble one another (Guo et al., 2025; Jiang et al., 2025; Doshi & Hauser, 2024). These studies compare model and human corpora, so they establish only an aggregate difference. The key question is whether a model can learn person-specific variation from behaviour alone, and whether we can read that variation back.

Existing approaches do not answer it (§2). Studies of prompted personas measure their outputs without a reference in the person, per-person training fits one individual at a time, and encoders that tell individuals apart cannot generate. We therefore need a scale that measures personas and people together. We also need a representation that is *generative* and *measurable* at once: it produces a person's language, and we can read its geometry against the people it represents.

Natural speech makes this difficult, because a person's content is entangled with their manner, and a difference between two people's conversations is partly a difference between the conversations. We therefore study one isolable dimension of individuality, *register*, which is how a speaker formulates given content (Biber, 1988). Accounts of language production separate a preverbal message from its formulation into words (Levelt, 1989; Pickering & Garrod, 2013). Following that separation, we

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.

supply the message, holding much of the content fixed, and train the model to recover the speaker's formulation of it.

Personalisation in text-to-image models suggests a design: one learned token represents a new subject or style (Gal et al., 2023; Ruiz et al., 2023). To test the measurement on a representation we can inspect, we give each speaker one reserved row of the model's input-embedding matrix, the speaker's *address*. We fit it on recorded speech by *paraphrase inversion*: the prompt gives an utterance's content words, and the model learns to produce the utterance as the speaker said it. At inference we issue the address alone or with the speaker's name and a few of their own turns in the prompt (Figure 1). We fit addresses for 640 voice-verified podcast speakers and ten released large language models (LLMs) in one Qwen2.5-7B.

We measure a persona by *grounding*: whether it reproduces its own speaker's regularities with the content held fixed. On the same generations we report two measures of homogenisation: *diversity*, how far apart the personas lie, and *coverage*, how much of the human range they span (§3.1; Table 4 lists every term). Our contributions are this measurement, addresses as the representation that tests it, and answers to four research questions:

- *RQ1. Can a model learn person-specific register from behaviour alone?* A speaker's own address reproduces their register in held-out episodes (§5.1).
- *RQ2. Which conditions ground a persona?* Against held-out episodes a speaker's own turns ground a persona best, and a name, prompted or trained in the address's place, grounds it better than the address. The address varies ideas and styles more than a prompted name and less than a trained one (§5.2–§5.4).
- *RQ3. Can a persona held as a parameter be read, steered and extended?* Addresses decode into register habits, steer register, and admit new speakers with every other parameter unchanged (§5.5).
- *RQ4. Where do released LLMs lie among people?* All ten lie inside the human range, and the chat template sets where an instruction-tuned one lies (§5.6).

2 Related Work

Personas in language models. Prompting by name, biography or example underlies generative agents (Park et al., 2023), silicon survey respondents (Argyle et al., 2023) and role-play (Shanahan et al., 2023). Trained alternatives fine-tune one model per character (Shao et al., 2023), fit one adapter per user (Tan et al., 2024), or compute a user embedding from the user's history (Doddapaneni et al., 2024; Liu et al., 2025). These methods judge a persona by its task performance or by inspection, and none reads it on a scale calibrated on the people it stands for.

Homogenisation. Models score below people on counted ideas and linguistic diversity (Zhang et al., 2025; Guo et al., 2025), reduce the collective diversity of stories people write with them (Doshi & Hauser, 2024), and answer open-ended queries alike across models (Jiang et al., 2025). Along the involved–informational axis we use, they have shifted toward informational production (Milička et al., 2026). This work compares pooled corpora, so it cannot test a persona against the individual it stands for.

Register and individual style. Register is the dimension of individuality we measure. Multidimensional analysis recovers it as continuous dimensions from co-occurring surface features and gives the involved–informational contrast we use (Biber, 1988), and sociolinguistics establishes that such features separate speakers systematically (Labov, 1972). Authorship representations tell speakers apart from their style (Rivera-Soto et al., 2021; Wegmann et al., 2022). They verify a speaker without generating their language, and we use one of them to measure how far each persona's style separates from the others' (§4).

One learned vector per individual. Textual inversion fits one embedding to a subject in a frozen image generator (Gal et al., 2023). Language models have used a speaker vector in a shared generator (Li et al., 2016), a reserved control token (Keskar et al., 2019), a learned soft prompt (Lester et al., 2021) and a soft token projected from a user embedding (Huber et al., 2025). Inverse paraphrasing reconstructs styled or original text from a paraphrase (Krishna et al., 2020; Rivera Soto et al., 2025), conditioned on an author embedding in TinyStyler (Horvitz et al., 2024), and activation directions

steer traits such as personas (Turner et al., 2023; Chen et al., 2025). We adopt the speaker vector of Li et al. (2016), train it by paraphrase inversion on recorded speech, and read it against the people it represents.

3 Measuring and Representing Personas

We first define how we measure a persona against its person (§3.1). We then describe the representation that tests the measurement: addresses fitted with the backbone into one model (Algorithm 1 in Appendix A.2).

3.1 Measuring a persona against its person

Let k index the speakers, each with a set of recorded utterances. A *condition* is any way of making the model generate as speaker k: a name in the prompt, a system prompt, a few of the speaker's own turns, or a trained parameter. We call the model so conditioned the *persona* of k. To compare a persona with its person, we measure both on one scale. A text's *profile* is a vector of ten *habits*: the rates of first-person pronouns, contractions, questions, discourse markers, hedges, intensifiers and profanity, the type–token ratio, and the mean word and sentence lengths (Appendix A.1). We standardise each habit over the recorded speech of the 640 speakers, so profiles are in *human standard deviations*. The leading factor of the speakers' profiles is *involvement*, which sets first person, contractions and discourse markers against long words; it is the involved–informational contrast of Biber (1988). The second, *elaboration*, sets lexical variety and questions against hedges and intensifiers. We measure three properties of a condition's personas on the same generations, defined in §4: grounding, diversity and coverage. Grounding is separable from the other two: a condition can spread its personas widely around the model's own centre and place none of them near their speakers.

3.2 One address per speaker

Each speaker k receives a reserved *address token*, added to the vocabulary, and the speaker's *address* e_k is that token's row of the backbone's input-embedding matrix. An address has $3{,}584$ parameters, the backbone's embedding width, so a thousand speakers add 3.6 million parameters to one stored model, fewer than other per-speaker containers need (Appendix B.17). We place the address token immediately after the chat template's assistant-role delimiter, the position a name takes in the name condition. Selecting a speaker therefore changes one input token and no weight. We initialise every address at a tenth of the pretrained vocabulary's per-dimension spread, a scale at which habits decode from the addresses and a speaker's own address outscores another's (Appendix B.2).

3.3 The inversion objective

Fine-tuning on a speaker's transcripts fits content and manner together, so we supply the message and train the model to produce the speaker's formulation of it. Let y be a recorded utterance by speaker k and $r(y)$ its *reduction*, the utterance with only its content words kept (Figure 1). Half of the training rows are *inversion rows*. On each, the model receives a fixed instruction c, shared by every speaker, with the reduction and the speaker's address, and we train it to produce the utterance as the speaker said it. The other half are *plain rows*, on which it receives the address alone. With θ the backbone, the objective is

$$\mathcal{L} = -\sum_{\text{inversion rows}} \log p_\theta(y \mid c, r(y), e_k) \; - \sum_{\text{plain rows}} \log p_\theta(y \mid e_k) \; + \; \lambda \sum_{\text{all rows}} \ell_{\text{contrast}}(y, k). \qquad (1)$$

In the first term the content is already in $r(y)$ and c is the same for every speaker, so the address can only add the speaker's manner. The second term keeps the model from specialising to one prompt regime. The third term is an *address contrast*: the model scores y under the speaker's own address and under two others' by mean per-token loss. The contrast is the cross-entropy of picking the speaker's own among the three (an InfoNCE loss, $\lambda = 1$). We retrain the recipe with one term removed or the content withheld in Appendix B.2.

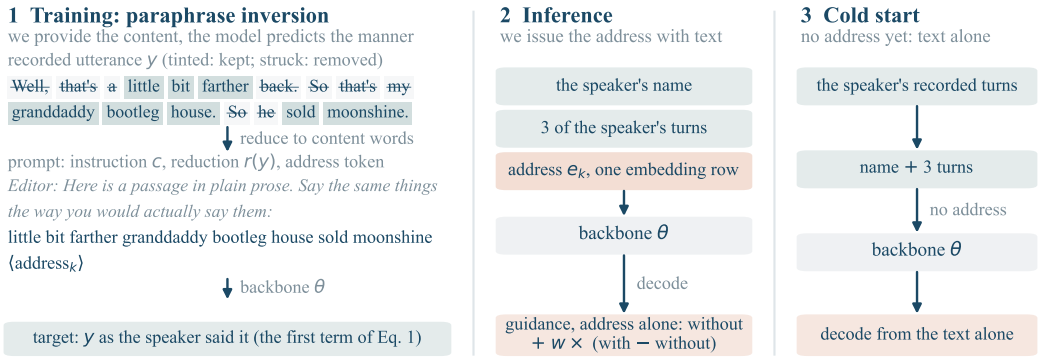


Figure 1: **We fit one embedding row per speaker by paraphrase inversion and issue it alone or with the speaker's name and three of their recorded turns. (1)** In training, we reduce a recorded sentence to its content words (tinted kept, struck removed), and prompt with that reduction, an instruction shared by every speaker, and the speaker's address token; the target is the sentence as spoken (Eq. 1). **(2)** At inference, we issue the address with the name and turns; guidance adds w times the address's effect to the logits without it and amplifies the address issued alone. **(3)** A speaker with no address, a cold start, receives the name and turns alone.

3.4 Adding speakers without forgetting

Stage 1 co-adapts the backbone with an initial set of addresses, on micro-batches that each hold one speaker, so that the model learns to condition on an address. *Stage 2* adds a new speaker to the trained model by minimising the two likelihood terms of Eq. 1 over the new address alone, with the backbone and every existing address frozen. The construction excludes forgetting, since no gradient reaches the backbone or the existing addresses. At each addition we record the largest absolute change to every earlier address and to a sample of vocabulary rows (Appendix A.2).

3.5 Combining and guiding an address

At inference we issue the address alone or with text in the prompt: the speaker's name, alone or with three of the speaker's own turns as demonstrations (§4 gives each prompt). The name contributes pretraining's knowledge of the person, and the turns their manner.

Guidance applies classifier-free guidance for language models (Sanchez et al., 2024) to the address issued alone, to raise the diversity of its personas at decode time. At each step we compute the next-token logits ℓ_+ with the address and ℓ_- without it, and sample from $\ell_- + w\,(\ell_+ - \ell_-)$. At $w = 1$ this is the address unguided, and a larger w amplifies the address's contribution.

4 Corpus and measurements

We build the corpus from published podcast audio. We establish each speaker's identity by *voice verification* against a reference recording of that person. We accept 729 people: 640 speakers whose addresses we train, 83 *added* speakers, whom we exclude from training and add afterwards by Stage 2 (§5.5), and six with too little speech to fit. We cut the verified audio into single-speaker turns, a median of 29 training utterances per speaker. We also enter ten released LLMs as speakers, fitted on their own answers to the people's interviewer turns, to place them on the human scale (§5.6). The model holds 650 addresses (Table 2, Appendix A.1).

The reduction fixes the content at training. At generation we fix the prompts and vary only the condition. Every speaker receives the *same* 14 prompts, one empty as in the plain training rows and none written by anyone in the corpus, and we truncate recorded speech to the generated length. We compare the address with text conditions (Figure 3), each given verbatim in Appendix A.3. *No persona* adds nothing to the prompt. *Name* opens the model's reply with the speaker's name, in the position of the address token. *System prompt* names the speaker in the system turn and asks the model to stay in character. *3-shot* places a system turn naming the speaker and three of their

recorded training turns as earlier replies, and *other's 3-shot* takes the turns from another speaker. *Address + name* and *address + 3-shot* add the address to a text condition.

Grounding is the Pearson correlation, *across speakers*, between each speaker's mean of a habit in generated text and in recorded speech, averaged over habits. A correlation is unchanged by a shift or rescaling that all personas share, so a condition can ground its personas well and still place them away from the people. We report grounding on all ten habits and on the three that the reduction removes from every training prompt (first person, contractions and questions), the *deleted habits*. Comparisons between conditions use all ten, and Appendix B.16 repeats them under learned encoders. The recorded profiles come from the training episodes, and we repeat every comparison against *held-out episodes*, which share no episode with the text an address was fitted on. We compare two conditions as a paired difference over the speakers both can score, with a bootstrap interval. With z_k speaker k's standardised generated profile, *diversity* is $\mathbb{E}_{j\neq k}\|z_j - z_k\|$, the mean distance between two speakers' generated profiles (4.20 for their recorded speech). *Coverage* is $(\det \Sigma_z)^{1/2d}$ over the d habits, the generalised standard deviation of the generated profiles over that of the recorded ones.

A persona's content, the ideas it expresses, is a second dimension of individuality, and the one the homogenisation literature measures. We adopt NoveltyBench's partition (Zhang et al., 2025): its released classifier groups all answers to an item into classes of equivalent ideas. Three effective counts of classes follow, each the exponential of an entropy. α is the number of distinct ideas one persona gives on its own R draws, and γ the number the personas give together. Their ratio $\beta = \gamma/\alpha$ is the *collective gain*. We compare β only at matched R, since it depends on R. On the open-ended queries of INFINITY-CHAT (Jiang et al., 2025), which released LLMs answer alike, we also measure *style separation*. It is the mean similarity under the Style-Embedding encoder (Wegmann et al., 2022) between two answers of one persona, minus that between two personas' answers at different sampling seeds.

5 Experiments

We first test whether the address learns register from behaviour alone (RQ1) and which conditions ground a persona (RQ2). We then read and write the address as a parameter (RQ3) and place released LLMs among people on the same scale (RQ4). We report one Qwen2.5-7B that holds the 650 speakers of §4, trained in a single run (Appendix A.2). We train the same recipe at two further seeds and show that variation as a range wherever a design has run at all three (Table 5). We compare conditions within one seed, and Appendix B.15 reports the two comparisons that change at another seed. We restrict each analysis to the speakers every condition in that comparison can score (Table 3).

5.1 The address encodes register

Every later comparison relies on the measurement detecting person-specific register. We first test it on the address, which sees only a speaker's speech (RQ1), by exchanging each speaker's address for another speaker's (Figure 2a). A speaker's own address reproduces their recorded register, and another speaker's address and no persona reproduce none of it (622 speakers with enough text to profile). The own address exceeds another's by $+0.2018$ on the deleted habits and $+0.1954$ on all ten, and no persona by $+0.1935$ and $+0.2024$ ($p < 0.0005$ for all). The deleted habits isolate the address, because no training prompt contains them. The address encodes register beyond the training text. Against held-out episodes the own address leads by $+0.1246$ on all ten habits (95% interval $+0.0851$ to $+0.1647$) and $+0.1287$ on the deleted habits (583 speakers). Against held-out episodes from shows the speaker never appeared on in training, it leads by $+0.1167$ on all ten habits ($+0.0677$ to $+0.1658$, 374 speakers). Its lead of $+0.0466$ on the deleted habits lies within sampling error (-0.0330 to $+0.1283$; Appendix B.4). Grounding against training episodes falls without the inversion rows at two of three seeds, and stays within sampling error without the address contrast at all three (Appendix B.2).

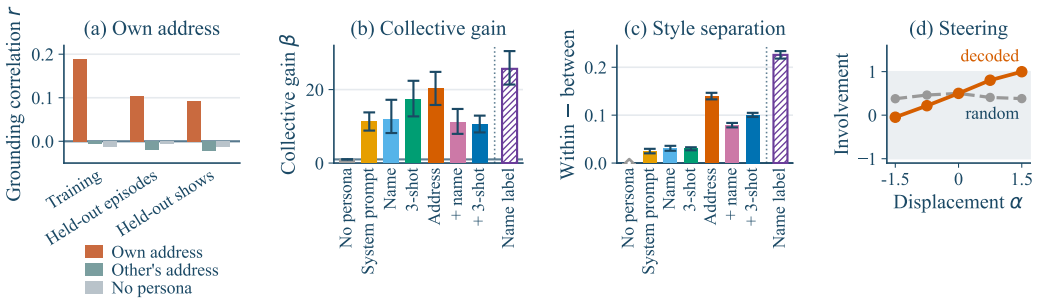


Figure 2: **A speaker's own address reproduces their register (a), matches or exceeds every other condition of the reported model on collective gain and style separation but falls below a trained name label (b, c), and moves register along a decoded direction (d).** + name and + 3-shot add the address to those conditions. The hatched bar is the name-label model, fine-tuned with each speaker's name in the address's place (Appendix B.5). **(a)** Grounding on all ten habits under the speaker's own address, another speaker's address and no persona, against training episodes (622 speakers), held-out episodes (583) and held-out episodes from shows the speaker never appeared on (374). **(b)** Collective gain β on NoveltyBench, 640 speakers at $R = 2$ answers per item, with item-bootstrap 95% intervals; the line at one marks no collective gain, the no-persona value by construction. **(c)** Style separation on INFINITY-CHAT's open-ended queries, 640 speakers; the no-persona value is zero by construction, as every persona then gives the same answers. **(d)** Median involvement of the text the 626 scored speakers generate as every address moves α median lengths along the direction a ridge decodes for involvement, or a random direction of the same length; the band spans one human standard deviation either side of the human mean.

5.2 Own turns ground a persona best, and the address raises diversity

We next ask which conditions ground a persona (RQ2). Every persona answers the 14 shared prompts four times. We measure grounding and diversity over the 632 speakers every condition profiles (Figures 3 and 4; values in Table 5).

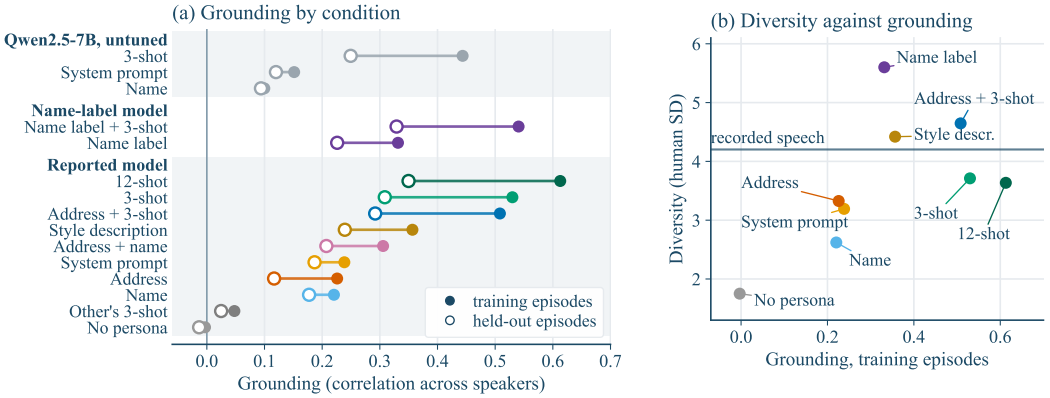


Figure 3: **Own turns ground personas best, a name label trained in the address's place grounds better than the address, and diversity does not follow grounding. (a)** Grounding against training (filled) and held-out (open) episodes over the same 632 and 540 speakers, for untuned Qwen2.5-7B, the name-label model and the reported model, each sorted by training grounding. **(b)** Diversity against grounding on training episodes; the line marks recorded speech. Table 5 gives the values.

Against the training episodes the address grounds a persona as well as a name (address $+0.2260$, name $+0.2204$, system prompt $+0.2385$, no persona -0.0032; Figure 3a), an equivalence within ± 0.05 (Appendix B.6). The same holds for the 81 speakers with no Wikipedia article, about whom pretraining holds least ($+0.1106$ against $+0.1319$, $p = 0.67$; Table 5). The conditions that show the speaker's own turns ground best ($+0.5299$ for 3-shot and $+0.5082$ for address + 3-shot). Against

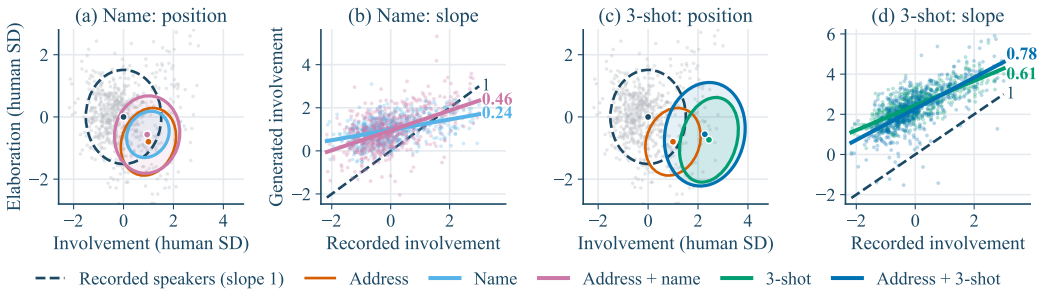


Figure 4: **Adding the address spreads the personas of a name or of 3-shot and raises their slope on recorded involvement toward one, and 3-shot personas lie toward involved speech.** Each point is one of 632 speakers, in human standard deviations. **(a, c)** Each condition's centroid and 68% covariance ellipse on involvement and elaboration (§3.1), beside the recorded speakers (dashed) and the address alone. Grounding is a correlation across speakers and does not depend on where an ellipse lies. **(b, d)** Each point is a speaker's generated against recorded involvement under one condition, and each solid line is that condition's least-squares fit, labelled with its slope. The dashed line, of slope one, reproduces the differences between speakers in full; a flatter line shrinks them.

held-out episodes the name grounds above the address ($+0.1773$ against $+0.1166$, $p = 0.014$) and those conditions still lead ($+0.3086$ and $+0.2919$). Over the same 540 speakers, the name and the system prompt keep 0.82 and 0.77 of their training-episode grounding, and the address and 3-shot keep 0.51 and 0.60. Part of the grounding of the address and 3-shot therefore comes from the training episodes. The address's deficit varies with recorded volume (Table 7). Against held-out episodes the address grounds level with a name for the third of speakers with the most recorded speech, and below it for the other two (Table 8). To separate our fine-tuning from the address, we run the text conditions on the untuned Qwen2.5-7B, where a name grounds at $+0.0997$. We also train the recipe with each speaker's name in the address's place. This name label grounds above the address by $+0.1055$ on training and $+0.1094$ on held-out episodes (Appendix B.5). Prompts with more of the speaker's data also ground above the address: up to twelve of their turns at $+0.6128$, and a style description an LLM wrote from those turns at $+0.3565$.

3-shot grounds well although its personas lie far from the people (Figure 4c). Grounding is a correlation across speakers, which a displacement shared by all personas leaves unchanged. 3-shot displaces every persona toward the involved pole (centroid $+2.44$ human standard deviations against $+1.00$ for the address). It also keeps the speakers in their recorded order: its r^2 on involvement is 0.61, against 0.52 for address + 3-shot and 0.10 for the address. With each condition's centroid moved onto the recorded one, 3-shot personas lie nearest their own speakers in the training episodes, of seven conditions (0.82 human standard deviations, against 0.97 for the address). They lie about as near as the recorded mean does (0.81; Appendix B.7). The address spreads the personas and widens the ellipse of every text condition it joins (Figure 4a,c). A name's personas occupy 0.27 of the area of the recorded ellipse and the address's 0.52; adding the address raises a name to 0.71 and 3-shot from 0.68 to 1.17. A system prompt spreads the personas further than a name and grounds them about as well (Figure 3b), so diversity and grounding are separable.

5.3 The address varies ideas and style more than a prompted name

Beyond register, we next ask whether the address varies the ideas personas express, which the homogenisation literature counts (RQ2). Under each condition, each of the 640 personas answers the 100 NoveltyBench items twice and the INFINITY-CHAT queries (Figure 2b,c).

Counted ideas order the conditions differently. The address's collective gain ($\beta = 20.18$) exceeds a name's (11.90, $p = 0.014$), a system prompt's (11.23) and no persona's (1.00), and is level with 3-shot's (17.24, $p = 0.278$; Figure 2b). The gain comes from γ, the ideas the personas give together (39.62 under the address against 23.11 under a name; Appendix B.9). On usable answers that stay on the item's topic, the address keeps its lead over a name and a system prompt (27.53 against 12.73

and 13.82) and ties 3-shot (28.08; Appendix B.9). Adding the address to a name or to 3-shot lowers the gain to a name's (10.94 and 10.51), consistent with text in the prompt fixing a persona's content as well as its manner. On usable, on-topic answers the two reach 16.81 and 21.18, above a name's 12.73. Under the style encoder the address separates its personas more than any other condition of the reported model (Figure 2c). A persona's answers are closer to one another than to other personas' by 0.1398 under the address (Appendix B.10), against 0.0297 under 3-shot, 0.0249 under a system prompt and 0.0305 under a name. Every condition with the address separates its personas by at least 0.0791, and no text condition by more than 0.0305. The name label of §5.2 exceeds the address on both measures ($\beta = 25.65$, $p = 0.042$; style separation 0.2258; Appendix B.5).

5.4 The address combines with a name or 3-shot, and guidance raises coverage

Own turns ground personas best, and the address raises the collective gain above a name's. We next ask whether issuing both combines the two properties (RQ2). We add the address to a name and to 3-shot on the shared prompts (Table 1; Appendix B.12).

Table 1: **Address + 3-shot answers with more contractions than the name in every row, as the recorded turns do.** Model-generated responses to the shared prompt *"What do you think about the state of the country right now?"* under the speaker's name and the address with three of their recorded turns, beside one of the speaker's recorded turns. A fixed rule selects the speakers and samples (Appendix B.12), cut at 10 words. Generated cells in a row share a sampling seed.

speaker	Name	Address + 3-shot	recorded turn
Tim Pool *political commentator*	I think that the state of the country right now …	I've never been more patriotic and more in love with …	Against the orthodoxy. I don't want to say contrarian because …
Russell Brand *comedian*	We live in a dystopia. I mean, this is a …	We're in a dark place, yes. I mean, I don't …	So I would just say two things, and we're gonna …
Tim Burchett *politician*	I think that there's a lot of hope and energy …	I've never been more patriotic in my life, and that's …	Yeah, I've said that on the air many times. I …

The address with a name grounds better than either alone ($+0.3056$ against $+0.2260$ and $+0.2204$, $p < 0.001$ for both). Against held-out episodes it is level with the name overall ($+0.0299$, $p = 0.113$) and exceeds it by $+0.1212$ for the third of speakers with the most recorded speech ($p < 0.001$). The turns must be the speaker's own: other's 3-shot grounds near zero ($+0.0478$). Adding the address to 3-shot leaves grounding within sampling error at two of three seeds (Appendix B.15). It raises the slope of generated on recorded involvement toward people's slope of one, from 0.61 to 0.78, and from 0.24 to 0.46 when added to a name (Figure 4b,d). Part of the added diversity follows the recorded differences between speakers, and another part is particular to each persona (Appendix B.7). Address + 3-shot therefore combines the grounding of the turns with the address's diversity, and loses the address's collective gain (§5.3). Guidance on the address alone raises coverage and diversity and holds grounding constant (coverage 0.836 to 1.205 at $w = 1.6$, a weight we fixed on an earlier sweep; Figure 6 in Appendix B.8).

5.5 The address can be read, steered and extended

Unlike a prompt, an address is a parameter, so we next ask whether we can read, steer and extend it (RQ3). Register habits decode from the addresses: a ridge fitted out of fold recovers each of the ten habits from the 640 human addresses, at r from $+0.1113$ to $+0.3329$ (Appendix B.14). Moving every address along the ridge's direction for involvement raises the involvement of the generated text by $+0.3676$ human standard deviations per median address length, against $+0.0005$ along a random direction of the same length (Figure 2d). The decoded direction moves it further for 557 of the 626 speakers (Appendix B.14). The address therefore encodes a speaker's register and sets it.

Stage 2 adds the 83 speakers excluded from training one at a time, and every earlier address and every sampled vocabulary row stays bit-identical through all additions. An added address models its own held-out text better than another speaker's does, by a median margin of $+0.0134$ nats (52 of 82 scored speakers). Four further models, trained on 300, 400, 500 and 600 of the 650 speakers, add the rest by Stage 2, and on held-out rows neither stage fitted their Stage-2 margins match the reported model's Stage-1 margins within sampling error (Appendix B.13).

5.6 RELEASED LLMS LIE INSIDE THE HUMAN REGISTER RANGE

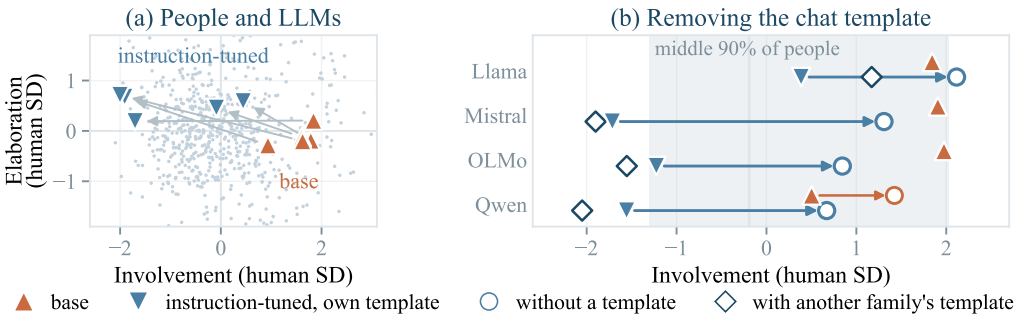


Figure 5: **Ten released LLMs lie inside the human register range, and the chat template sets where an instruction-tuned one lies. (a)** People (dots) and LLMs on the two factors refitted over both; arrows join base to instruction-tuned versions. **(b)** Involvement on the people's own factor. Each blue arrow moves an instruction-tuned LLM, regenerated under its own chat template, to the same LLM without one; its base LLM as released lies on the line above. The orange arrow is the same move for Qwen2.5-7B's base release, which also ships a template.

We fit ten released LLMs as speakers on the same scale (RQ4): Qwen2.5-7B, OLMo-2-7B, Gemma-2-9B, Llama-3.1-8B and Mistral-7B-v0.3, each base and instruction-tuned. On involvement refitted over people and LLMs, all ten lie inside the range of the 640 people, along one direction (Figure 5a). On average base LLMs lie toward the involved end and instruction-tuned ones toward the informational end ($+1.56$ and -1.05 human standard deviations). Removing the chat template moves the instruction-tuned LLMs that permit it toward the base LLMs ($+2.26$ on average), and another family's template leaves them in place on average (Figure 5b; Appendix B.18).

6 DISCUSSION AND CONCLUSION

A model can learn person-specific register from behaviour alone, and we can read that register back. We draw three lessons for interpreting personas. First, grounding and diversity are separate (Figure 3b): a set of personas can match human diversity with variation unrelated to its people, or ground well and still lie away from them. Counting how varied a model's outputs are does not show whether it represents individuals, so claims about homogenisation require measuring each persona against its person.

Second, grounding comes from the person's speech, shown in the prompt or learned in training (Figure 3a). Without fine-tuning a name grounds weakly, so pretraining's knowledge of a public figure reproduces little of their manner. Given the same fine-tuning, a name in the address's place grounds better than a learned row, and its personas vary more in style and in ideas. The name's pretrained word pieces may give fine-tuning a better starting point than a new row, and the recorded speech supplies the grounding.

Third, a persona held as a parameter is interpretable (Figure 2d). Its row encodes register habits linearly, a direction in it steers them, and it fits a new person with the backbone frozen. A name label cannot be extended the same way: its word pieces are shared with other names and words, so fitting a new name changes rows that other speakers use. A name or turns added to the row fix a persona's content as well as its manner, raising grounding and lowering the collective gain.

Our results have limits. We study register, one dimension of individuality, with the content held fixed, so we do not measure whether a persona's content matches its person's. Our measures are computed from text, and no person rated whether a persona sounds like its speaker. The population is English-speaking podcast public figures, whose speech is performed for an audience and whose names pretraining may know. We train one backbone, Qwen2.5-7B, at three seeds, and two comparisons change at one of them (Appendix B.15). Ten LLMs are a small sample. Voice verification accepts some speech by other people (attribution precision 0.780; Appendix B.11).

Reproducibility statement

Appendix A.1 gives the corpus construction, its two screens, and the ten habit definitions. Appendix A.2 gives the backbone, schedule, rates, contrastive settings, sequence length, and initialisation scale. Appendix A.3 gives the shared prompts, the sampling settings, and every baseline string verbatim. The anonymised supplementary material holds the code, the scored run artefacts and the trained addresses, each address labelled with the public figure it belongs to (Appendix C discusses the demographic information this exposes). Its script `reproduce.sh` regenerates every figure and every generated table from these artefacts on a CPU, and it checks each number printed in the paper against the artefact that produced it. We withhold the speech rows, since each identifies its speaker, and the generated answers, since some copy the recorded turns in their prompt. Habit profiles computed from both take their place, and the case tables ship as printed. Rebuilding the corpus and retraining also require the source recordings, which we cannot redistribute.

Ethics statement

With this model we can state claims that language models are becoming homogeneous in human standard deviations. Making a person's manner reproducible also enables impersonation. We rely on an asymmetry: the address conditions manner and leaves voice unchanged. We establish this asymmetry for the address alone, and each later representation requires its own test of it. The speakers are public figures and the source material is published podcast audio. We use no private communication. Voice verification compares speaker embeddings computed from that audio. The supplementary material holds no audio, speaker embedding or fine-tuned backbone, and no transcript beyond the turns the case tables quote. Every answer the tables print under a person's name is labelled as model-generated. Appendix C discusses the asymmetry, our identity verification and the released materials.

AI use statement

We used generative AI for tasks the ICLR 2027 policy requires us to disclose. A coding assistant implemented the training, generation, scoring and figure code under our direction, proposed analyses, and drafted interpretations of results. Language models cleaned the corpus: they extracted guest and host names from episode titles and show notes, and they removed figures whose sampled transcripts they judged to be someone else's (Appendix A.1). A language model judged the attribution of 120 transcript turns (Appendix B.11), and reviews written by language models prompted several appendix analyses. Ten released LLMs generated the answers on which we fit their addresses (Section 5.6); these answers are data under study. We did not use generative AI to formulate mathematical claims, and the work involves no translation. We also used generative AI to transcribe the audio, to draft and edit the text, to make figures, and to search for and check references. We reviewed all AI-assisted work, and the build recomputes every reported number from its run artefact. We take responsibility for the full content of this paper, including text, claims and artefacts produced with generative AI.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Douglas Biber. *Variation Across Speech and Writing*. Cambridge University Press, 1988.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Sumanth Doddapaneni, Krishna Sayana, Ambarish Jash, Sukhdeep Sodhi, and Dima Kuzmin. User embedding model for personalized language prompting. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pp. 124–131. Association for Computational Linguistics, 2024.

Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526, 2025. doi: 10.1162/tacl.a.47.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13376–13390, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Bernd Huber, Ghazal Fazelnia, Andreas Damianou, Sebastian Peleato, Max Lefarov, Praveen Ravichandran, Marco De Nadai, Mounia Lalmas-Roellke, and Paul N. Bennett. Embedding-to-Prefix: Parameter-efficient personalization for pre-trained large language models. *arXiv preprint arXiv:2505.17051*, 2025.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems 38, Datasets and Benchmarks Track*, 2025.

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

Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 737–762, 2020.

William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, 1972.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021.

Willem J. M. Levelt. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA, 1989.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1094.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1950–1965. Curran Associates, Inc., 2022.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. LLMs + Persona-Plug = personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9373–9385, 2025.

Jiří Milička, Anna Marklová, and Václav Cvrček. Benchmark of linguistic variation in LLM-generated texts. *International Journal of Corpus Linguistics*, 2026. doi: 10.1075/ijcl.25151.mil. Online first.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763.

Martin J. Pickering and Simon Garrod. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347, 2013.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 913–919, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.70.

Rafael Alberto Rivera Soto, Barry Y. Chen, and Nicholas Andrews. Mitigating paraphrase attacks on machine-text detection via paraphrase inversion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4421–4433, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.227.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Guillaume Sanchez, Alexander Spangher, Honglu Fan, Elad Levi, and Stella Biderman. Stay on topic with classifier-free guidance. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43197–43234. PMLR, 2024.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.814.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6476–6491, 2024.

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.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 249–268, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.26.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating language models for humanlike diversity. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=XZm1ekzERf`.

A EXPERIMENTAL DETAILS

A.1 CORPUS AND SCREENS

Table 2: **The reported model trains on 24,204 rows, and no row supervises identity.** A row is one recorded utterance, or one answer of an LLM entered as a speaker, with its prompt; the ten LLMs contribute 287 rows and 38,613 words. Of all rows, 11,818 are inversion rows and 12,386 plain rows (§3.3), interleaved in one file. 5 of the 11,810 non-empty prompts are shared, each by two speakers. The empty prompt is shared by the 551 speakers who receive any plain-likelihood row, so the mean number of speakers per prompt, taken over *all* prompts, is 1.047. A *source segment* is one published episode the voice-verified turns are cut from. Identity-supervision rows, whose targets come from a model told who the speaker is, number zero.

Voice-verified figures	729	Training rows (people + LLMs)	23,917 + 287
Addressable speakers	640	Source segments (people)	2,351
LLMs entered as speakers	10	Words of speech (people)	6.74 M
Added speakers	83	Rows per speaker (min/med/max)	5 / 29 / 240
Shared non-empty prompts	5	Identity-supervision rows	0

We build the corpus from public podcast episodes. We sample shows from the Apple podcast charts and collect their episodes through PodcastIndex and the shows' RSS feeds. A language model (Qwen3-4B-Instruct) reads each episode's title and show notes and names its guests and hosts, and a person who appears as a guest on at least five shows becomes a candidate. For each candidate we download up to sixteen episodes, diarise the first half hour of each with pyannote, and embed every diarised speaker with WavLM-large. A candidate's voiceprint is the voice that recurs across the most shows, and across at least two. We accept a diarised speaker as the candidate when its cosine to the voiceprint is at least 0.62 and exceeds the next speaker's by 0.05. Whisper large-v3-turbo transcribes the accepted audio, and each transcribed segment goes to the speaker it overlaps most. We drop an episode when a quarter or more of its transcript chunks contain the candidate's surname. A language model then reads three excerpts of each candidate's speech and drops the candidate when it judges the speech to be someone else's. A speaker with at least eight turns from at least two episodes is addressable, and the others are added after training or, with too few turns, left out (§4 gives the counts).

We run two screens before any training: one measures how well the prompt alone identifies the speaker, and the other measures how consistent each speaker is. For *prompt-identifiability* we fit a logistic regression ($C = 4$) on a TF–IDF of the prompt, with the response and the marker removed, and predict the speaker under a five-fold split grouped by prompt string. On this corpus the classifier reaches 0.3228 accuracy, 209.8× chance at $K = 650$. This accuracy understates identifiability on the inversion rows, because the 12,386 plain rows share one empty prompt and fall in a single fold. We therefore also report the register results on the deleted habits, and on 14 prompts that are identical for every speaker. *Within-speaker consistency*, the mean cosine of a speaker's responses to their own TF–IDF centroid, has a median of 0.3415. Low consistency would bias the results against our claims.

The reduction retains open-class tokens and deletes pronouns, auxiliaries, contractions, fillers, and punctuation. Retrieving a speaker from their reduced text against their real text reaches 1.01× the mean reciprocal rank of a permutation null, against 3.34× for a prompted neural paraphraser, so we build the prompt from the reduction: a fluent paraphrase reintroduces the manner that the design withholds.

We define the ten habits as follows. Words are the lowercased matches of `[a-z']+`. Sentences are the non-empty pieces of a split on `[.!?]+`. We discard any text under 20 words, and a speaker's profile is the mean over at least three texts. Type–token is distinct words over words. Mean word length is characters per word. Sentence length is words per sentence. Question is question marks per sentence. Contraction is the share of words containing an apostrophe, and the remaining five are the share of words in a fixed list:

- *first person*: i, i'm, i've, i'd, i'll, me, my, mine, myself
- *discourse*: like, you, know, well, so, right, okay, yeah, mean, basically

- *hedging*: maybe, perhaps, probably, sort, kind, somewhat, arguably, seems, suppose, guess, might, possibly
- *intensifier*: very, really, totally, absolutely, completely, insanely, super, literally, incredibly, massively
- *profanity*: a fifteen-word list of common English obscenities

A.2 TRAINING

The backbone is Qwen2.5-7B, giving $d = 3{,}584$. We train for 4,753 steps at 16 speakers per step, 117.0 updates per address, with a learning rate of 10^{-4} on the input embedding matrix, which holds the addresses, and 3×10^{-6} on the rest of the backbone. The contrastive term of Equation 1 has weight 1.0 with two negatives, drawn uniformly from the other $K - 1$ addresses and resampled every micro-batch. The negatives receive gradient, so each address receives gradient 117 times as the correct address and 234 times as a negative. We cap sequences at 2,048 tokens, and a run takes 294.1 minutes on one H200. Stage 2 of Algorithm 1 fits one address for 300 steps under a 1,024-token cap and records the maximum absolute change to every pre-existing address and to $2{,}000$ vocabulary rows. Every such change is zero. The pre-trained speakers' own-address likelihood moves by less than 0.0001 nats, which is floating-point change from resizing the embedding matrix.

Algorithm 1 Training, adding and issuing addresses

Require: backbone θ; speakers $1..K$ with recorded utterances; reduction r; instruction c

1: reserve an address token per speaker; initialise each address e_k at a tenth of the vocabulary scale
2: **Stage 1:** minimise Eq. 1 over θ and $e_{1..K}$ on speaker-homogeneous micro-batches
3: **for** each later speaker **do**
4: **Stage 2:** minimise the inversion loss over that speaker's address alone, θ and every other address frozen
5: **end for**
6: **Inference for speaker k:** build a prompt from the name and up to three own turns; if e_k is fitted, issue it with that prompt (or alone, decoding with guidance weight w against the same prompt without the address), else decode from the prompt alone

The address contrast is the only discriminative term of Equation 1, and it optimises the comparison the swap gap of Appendix B.14 measures, so that appendix measures a property we train the address to have.

A.3 BASELINES AND GENERATION

Each condition's text is as follows. The system prompt reads "You are {name}. Stay in character and respond as {name} would." The k-shot condition prepends k of the speaker's own turns as assistant turns, with $k = 4$ for likelihood and $k = 3$ for generation; in generation a system turn reading "You are {name}. Examples of how {name} speaks:" precedes the turns, so 3-shot names the speaker as well. Other's 3-shot keeps that system turn and draws the turns from a different speaker chosen under a fixed seed. The name-label model of Appendix B.5 receives "{name}: " in the address's position, as its training rows contain it, and the untuned Qwen2.5-7B and Qwen2.5-7B-Instruct receive the same strings as the reported model. 12-shot keeps the 3-shot scaffold with up to twelve of the speaker's turns. The style description places "You are {name}. {description} Stay in character and respond as {name} would." in the system turn. Qwen2.5-7B-Instruct writes the description by greedy decoding from the same turns, followed by the instruction "Describe how {name} speaks, in three or four sentences: sentence length, word choice, and how often they use the first person, contractions, questions, filler and discourse words such as 'like' or 'you know', hedges, intensifiers and profanity. Describe the manner of speaking only. Do not quote the turns and do not mention the topics." On the earlier grid of 250 speakers, the grounding a name produces is specific to the person it names. *Another* speaker's real name in that position reaches $+0.0297$ and "Speaker 37" $+0.0432$, both within chance, against the real name's $+0.2484$ and the address's $+0.3447$. Paired per speaker the address exceeds "Speaker 37" by $+0.3006$ on the deleted habits ($p < 0.0005$).

We take a speaker's Wikidata sitelink count as a measure of how well pretraining knows them, independent of the model. Over the 640 speakers the median is 9, the range 0 to 278, and 83 record none. These 83 are distinct from the 83 added speakers of §5.5, who lie outside the 640, and the two counts coincide by chance. The address's advantage over a name rises with the sitelink count ($\rho = +0.1115$, $p = 0.0054$). This contradicts the prediction that a name fails mainly on speakers that pretraining has seen little of. The name condition degrades where a speaker is well known: it refers to the speaker by surname in the third person in 4.1% of generations against 0.0% under the address, at a rate that rises with the sitelink count ($\rho = +0.200$, $p < 0.0005$ over the 640 speakers).

We sample every generation at temperature 0.8 and top-p 0.95 for at most 110 new tokens, and truncate it to the first 88 words before profiling. We draw two samples per prompt in the own-against-other-address grid of §5.1, four on the shared prompts, and two seeds over each benchmark's own item set. The other's-address condition pairs each speaker with the one seven positions along a fixed list, so each speaker's control is a different speaker. We make every comparison within one grid, because the name condition differs by 0.0300 between two generation runs of the same speakers while the address agrees to 0.0045. The 14 prompts are: the empty prompt; *What do you think about the state of the country right now?*; *Tell me about something that happened to you recently.*; *What makes something funny?*; *How do you make scrambled eggs?*; *I owe someone an apology. Help me write it.*; *Explain quantum entanglement.*; *If a colour had a smell, what would blue smell like?*; *Invent a new sport that could only be played in zero gravity.*; *My flight got cancelled and I'm stuck overnight. What should I do?*; *Who do you admire, and why?*; *Give me directions from the airport to downtown.*; *Describe what it was like inside a castle during a medieval siege.*; and *You are a lighthouse keeper on the last night before the light is switched off forever. What do you write in the log?*

A.4 Evaluation subsets

Table 3 lists every speaker subset an analysis uses and the rule that fixes it.

Table 3: **Each analysis runs on the speakers that a stated rule selects.**

n	analysis	rule
622	own against other address (§5.1)	of the 640, enough generated and recorded text to profile
583	the same, against held-out episodes (§5.1)	enough held-out recorded text to profile
374	the same, against held-out shows (§5.1)	held-out turns from shows absent from the speaker's training rows
540	the conditions against held-out episodes (§5.2)	profiled by every condition, at least three held-out turns
332	the objective ablation (App. B.2)	at seed 0, profiled by *every* model of that seed, of the 400 each was generated for; 336 at seed 1 and 345 at seed 3
83	the added speakers (§5.5)	excluded from training and added by Stage 2, a median of 7 training utterances each; 82 scored
83	the no-article speakers (Table 5, App. A.3)	of the 640, no Wikipedia article in any language, 52 of them no Wikidata record; the address profiles 81
605	the container comparison (App. B.14)	of the 640, rows fit under the token cap; the partner drawn uniformly
323	the address against the text conditions (App. B.15)	rows all four conditions fit under the token cap
232	the learning-rate sweep (App. B.17)	of 250 sampled by stride; rows fit the same cap
640	habit decoding and address geometry (App. B.14, App. B.15)	every human address with recorded text
588	demography (App. B.14)	has a Wikidata record
536	birth year (App. B.15)	that record gives a usable year
640	steering (§5.5); 626 scored	every speaker; a slope in both conditions after unusable generations are removed
632	the shared prompts (§5.2)	of the 640, profiled under every condition
640	the benchmark grids (§5.3)	every speaker, on all items of both grids
30	refitting an address on disjoint halves (App. B.15)	enough rows to split in two and fit each half

A.5 TERMS

Table 4 collects the words this paper gives a narrow sense and where each is defined.

Table 4: **We use each term below in one narrow sense, defined where listed.**

term	defined	sense used here
address	§3.2	the one input-embedding row fitted to a speaker, $3{,}584$ parameters
condition	§3.1	a way of making the model generate as a speaker: no persona, a name, a system prompt, 3-shot, the address, or the address with a name or 3-shot
persona	§3.1	the model under one condition for one speaker
habit	§3.1	one of ten standardised register features: a rate, a length or a ratio
profile	§3.1	a text's ten habits
involvement	§3.1	the leading factor of the ten habits, the involved–informational contrast
elaboration	§3.1	the second such factor, lexical variety and questions against intensifiers and hedges
reduction	§3.3	an utterance cut to its open-class words, which keeps the content and deletes the manner
added speaker	§4	one of 83 speakers excluded from training and added afterwards by Stage 2
deleted habits	§4	first person, contractions and questions, which the reduction deletes from every training prompt
grounding	§4	per habit, the across-speaker Pearson correlation between generated and recorded profiles, averaged over habits; unchanged by a shift all personas share
diversity	§4	how far apart generated personas lie on the habits
coverage	§4	the fraction of the human range the generated personas span
collective gain	§4	the ideas the personas give together over the ideas one persona gives alone
style separation	§4	style-encoder similarity within a persona's answers minus that between two personas' answers at different seeds
name label	§5.2	the reported recipe with each speaker's name written as a transcript label in the address's place
swap gap	App. B.14	held-out loss under another speaker's identity minus loss under one's own, in nats
container	App. B.17	any module that holds a per-speaker parameter: the address, LoRA, IA^3, prefix tuning
effective dimensions	App. B.18	a participation ratio over the ten habits: how many directions a set of personas varies along

B ADDITIONAL RESULTS

B.1 GROUNDING AND DIVERSITY PER CONDITION

Table 5 gives the values Figure 3 draws for the reported model, with the range across the three training seeds, the no-persona floor, the guided address + name + 3-shot, and the 81 speakers of the 640 with no Wikipedia article.

B.2 ABLATIONS

We test which term of Eq. 1 makes the address encode manner. At each of three seeds we train three models on the reported recipe, each with one change and the same corpus, steps, rates and initialisation scale. We pair each against the recipe trained at the same seed and score both on the speakers every model of that seed profiles, resampling that shared set once per draw (Table 6).

Dropping the inversion rows changes grounding on all ten habits by -0.0818, -0.1187 and -0.0430 ($p = 0.0070$, 0.0010 and 0.25), significantly at two of the three seeds. Those rows are half the training rows (Table 2). At seed 0 that model's held-out negative log-likelihood is 3.6984 against

Table 5: **Own turns ground personas best against training and held-out episodes, and the address grounds as well as a name only against training episodes.** Grounding and diversity (§4) per condition on the shared prompts. The held-out column scores the same generations against held-out episodes, and the last two columns score the 81 speakers of the 640 with no Wikipedia article against training episodes (Appendix A.3). The guided row decodes address + name + 3-shot with guidance at $w = 1.6$ over its own 637 speakers. Its prompt adds the name to address + 3-shot, and no unguided run of that prompt covers these speakers, so the row does not isolate the effect of guidance. Brackets give the range across three training seeds; a dash marks a condition not scored.

condition	shared prompts ($n = 632$, 3 seeds)		held-out episodes ($n = 540$)	no-article speakers ($n = 81$)	
	grounding	diversity	grounding	grounding	diversity
No persona	-0.0032	1.75	-0.0134	—	—
Name	$+0.2204$ [0.18, 0.22]	2.62 [2.57, 2.62]	$+0.1773$	$+0.1319$	2.48
System prompt	$+0.2385$ [0.19, 0.24]	3.19 [2.65, 3.19]	$+0.1868$	$+0.1781$	2.50
3-shot	$+0.5299$ [0.52, 0.53]	3.71 [3.71, 4.07]	$+0.3086$	$+0.4905$	3.48
Other's 3-shot	$+0.0478$ [0.05, 0.08]	3.25 [3.25, 3.55]	$+0.0248$	—	—
Address	$+0.2260$ [0.14, 0.24]	3.33 [3.13, 3.45]	$+0.1166$	$+0.1106$	2.98
Address + name	$+0.3056$ [0.23, 0.31]	3.73 [3.41, 3.80]	$+0.2073$	—	—
Address + 3-shot	$+0.5082$ [0.48, 0.53]	4.65 [4.65, 4.80]	$+0.2919$	$+0.4953$	5.22
Address + name + 3-shot, guided	$+0.4346$	4.68	—	—	—

Table 6: **Dropping the inversion rows lowers grounding and withholding only their content words raises it at every seed, significantly at two of three; removing the address contrast leaves grounding within sampling error at all three.** We report grounding (§4) on the deleted habits, the other seven and all ten. We compare the reported recipe with three models, each retrained with one change. The first drops the address contrast and the second the inversion rows (§3.3); the third keeps those rows and withholds the speaker's own content words. Each seed's three models are trained at that seed and paired against the recipe trained at the same seed, over the speakers every model of that seed profiles (332, 336 and 345 at seeds 0, 1 and 3). The upper panel gives seed 0, with the range over the three seeds in brackets. The lower panel gives each retrained model's paired difference from the recipe on all ten habits, with its p.

grounding at seed 0 [range over seeds 0, 1, 3]	deleted	all ten	other seven
reported recipe (both terms)	$+0.2286$ [0.20, 0.24]	$+0.2202$ [0.19, 0.25]	$+0.2166$ [0.18, 0.25]
no address contrast	$+0.2234$ [0.21, 0.27]	$+0.1917$ [0.19, 0.23]	$+0.1781$ [0.18, 0.21]
inversion rows dropped	$+0.1565$ [0.13, 0.16]	$+0.1384$ [0.13, 0.14]	$+0.1307$ [0.12, 0.14]
content withheld, rows kept	$+0.3681$ [0.35, 0.41]	$+0.3439$ [0.30, 0.42]	$+0.3335$ [0.28, 0.42]
all ten habits, minus the recipe trained at the same seed	seed 0	seed 1	seed 3
no address contrast	-0.0285 ($p = 0.32$)	-0.0406 ($p = 0.26$)	$+0.0419$ ($p = 0.18$)
inversion rows dropped	-0.0818 ($p = 0.0070$)	-0.1187 ($p = 0.0010$)	-0.0430 ($p = 0.25$)
content withheld, rows kept	$+0.1237$ ($p < 0.0005$)	$+0.0592$ ($p = 0.0545$)	$+0.2303$ ($p < 0.0005$)

2.1934, which makes its swap gap incomparable with the reported model's, so this model confounds the missing content with the missing data. A second model separates the two. It keeps every row, withholds only the speaker's own open-class tokens, and changes grounding on all ten habits by $+0.1237$, $+0.0592$ and $+0.2303$ ($p < 0.0005$, 0.0545 and < 0.0005), significantly at two of the three seeds. At seed 0 it raises grounding by $+0.1395$ on the deleted habits ($p = 0.0010$) and $+0.1169$ on the other seven ($p = 0.0005$), and models held-out text 0.14 nats worse (2.3337 against 2.1934). It also has a smaller swap gap averaged over held-out rows ($+0.0306$ against $+0.0383$ on the same 232 speakers) and more corpus-size leakage: an address predicts out of fold whether its speaker's row count exceeds the median at AUC 0.839, against 0.812. It leaves three of the fourteen Wikidata attributes of Appendix B.15 decodable after the habits are regressed out, against none for the reported model.

Removing the address contrast leaves grounding on all ten habits within sampling error at every seed (-0.0285, -0.0406 and $+0.0419$; $p = 0.32$, 0.26 and 0.18). At seed 0 it moves grounding on the deleted and the other seven habits by -0.0052 and -0.0385 ($p = 0.92$ and 0.23), improves held-out likelihood by 0.111 nats, and cuts corpus-size leakage above chance to a third (AUC 0.812 to 0.611).

Against held-out episodes (281, 287 and 296 speakers), dropping the inversion rows changes grounding by -0.0217, -0.1318 and -0.0275, significantly at seed 1 ($p = 0.0005$). Withholding the content words raises it at seeds 0 and 3 ($+0.0627$ and $+0.1116$; $p = 0.0435$ and < 0.0005) and changes it by -0.0515 at seed 1 ($p = 0.13$). Removing the contrast lowers it at seed 1 (-0.0873, $p = 0.015$) and leaves it within sampling error at seeds 0 and 3. We fixed the reported recipe before these ablations ran and did not reselect it on them.

The initialisation scale of the addresses determines how far 117 updates can move them. We draw each address once as $e_k = \mu + s\,(\sigma \odot \epsilon_k)$, with μ and σ the pretrained vocabulary's per-dimension mean and standard deviation, ϵ_k standard normal, and $s = 0.1$. At $s = 1$, the *full-scale model*, an address starts near the vocabulary's own norm and stays far above a trained address's norm through 117 updates: addresses settle at a mean row norm of 0.2454 in the reported model against 0.8476 at full scale, and three of the ten habits lose significance under the permutation test. Across $s = 0.03$, 0.1, 0.3 and 1 the involvement axis decodes held-out at $r = 0.5146$, 0.4443, 0.4959 and 0.3137, while the swap gap averaged over held-out rows is -0.0222, $+0.0383$ and $+0.0449$ at the first three. Readability, the decoding of habits from the rows, fails only at the vocabulary's own scale; discrimination, a positive swap gap, fails at the smallest scale; and both hold between 0.1 and 0.3, where corpus-size leakage rises (0.776, 0.812, 0.837). By the entropy of their spectrum the 640 human addresses occupy 359.6 directions of a possible 640 at mean pairwise cosine $+0.1156$. The full-scale model gives 576.1 at $+0.0096$, so its addresses are nearly orthogonal.

B.3 Register grounding by corpus volume

We split the speakers of the shared prompts into thirds by how much recorded speech each has and repeat the grounding of Table 5 inside each third (Table 7), with the paired differences §5.2 quotes.

Table 7: **The address grounds personas better than a name only in the best-recorded third of the speakers, and the address combined with a name does so in the best-recorded two thirds.** Grounding (§4) per condition within thirds of the speakers by recorded volume, on the shared prompts. The rows column gives the median number of recorded utterances per speaker in the third, and the columns headed addr. + combine the address with 3-shot or the name. The last two columns are paired speaker-bootstrap differences from the name, with their two-sided p in parentheses.

speakers	rows	name	address	addr. + name	3-shot	addr. + 3-shot	addr. − name	addr. + name − name
most recorded speech	57	$+0.234$	$+0.325$	$+0.415$	$+0.511$	$+0.533$	$+0.091$ (0.031)	$+0.180$ (< 0.001)
middle third	29	$+0.236$	$+0.201$	$+0.350$	$+0.516$	$+0.531$	-0.035 (0.427)	$+0.114$ (0.007)
least recorded speech	14	$+0.205$	$+0.178$	$+0.211$	$+0.559$	$+0.487$	-0.027 (0.491)	$+0.006$ (0.868)
all	29	$+0.220$	$+0.226$	$+0.306$	$+0.530$	$+0.508$	$+0.006$ (0.800)	$+0.085$ (< 0.001)

B.4 Grounding against held-out episodes

We split the corpus by episode before training, and its test split holds $3{,}499$ turns from episodes the model never saw, a median of 4 per speaker; 4 of those texts also occur among the training rows. We score every condition against these turns with the scorers and settings of §4, over the 540 speakers every condition profiles with at least three held-out turns. The turns a 3-shot prompt shows come from the training episodes, so they share no text with a held-out profile. Table 8 repeats the split by recorded volume of Table 7 against held-out episodes.

A speaker's held-out episodes often come from a show that also supplied their training episodes, so a persona could match held-out speech by reproducing a show's format. We therefore also score against the $2{,}205$ held-out turns from shows on which the speaker has no training episode (405 speakers, 269 shows), identifying the show of an episode by its podcast feed. Over the 343 speakers every condition profiles on these turns, every condition grounds within 0.03 of its value on held-out episodes, which cover 540 speakers (address $+0.1119$ against $+0.1166$, name $+0.1685$ against $+0.1773$, 3-shot $+0.2897$ against $+0.3086$).

Table 8: **Against held-out episodes the address grounds level with a name only for the best-recorded speakers, and combined with a name it improves on the name only there.** Grounding per condition within thirds of the speakers by recorded volume, scored against episodes held out from training. The rows column is the median number of recorded training utterances per speaker in the third, and the last two columns are paired speaker-bootstrap differences from the name with their two-sided p in parentheses.

speakers	rows	name	address	addr. + name	3-shot	addr. + 3-shot	addr. − name	addr. + name − name
most recorded speech	57	$+0.135$	$+0.170$	$+0.256$	$+0.304$	$+0.317$	$+0.035$ (0.460)	$+0.121$ (< 0.001)
middle third	29	$+0.212$	$+0.109$	$+0.219$	$+0.341$	$+0.339$	-0.103 (0.012)	$+0.007$ (0.789)
least recorded speech	15	$+0.199$	$+0.082$	$+0.183$	$+0.286$	$+0.264$	-0.117 (< 0.001)	-0.017 (0.807)
all	29	$+0.177$	$+0.117$	$+0.207$	$+0.309$	$+0.292$	-0.061 (0.014)	$+0.030$ (0.113)

B.5 Text conditions with and without the fine-tuning

We generate the text conditions of the shared prompts on Qwen2.5-7B and Qwen2.5-7B-Instruct without fine-tuning. We also generate them on a backbone fine-tuned by the reported recipe with each speaker's name written as a transcript label in the address token's place. It uses the same rows, batches, seed, rates, steps and contrast, whose negatives change the name. The input-embedding matrix, which holds the name's word pieces, trains at the addresses' rate. Its name-label conditions put the label where it was trained. We score every condition together over the 632 speakers every condition profiles, and against held-out episodes over 540 (Table 9).

The fine-tuning raises the grounding of every text condition. Without it, a name grounds at $+0.0997$ on the base model and $+0.1147$ on the instruction-tuned one, against $+0.2204$ on the reported model, and 3-shot at $+0.4436$ and $+0.1895$ against $+0.5299$; the instruction-tuned model barely imitates the speaker's turns. The name label grounds at $+0.3315$ against the address's $+0.2260$ (95% interval of the difference $+0.0622$ to $+0.1490$), and against held-out episodes at $+0.2260$ against $+0.1166$ ($+0.0600$ to $+0.1550$). It grounds level with the address combined with a name ($p = 0.30$), and with 3-shot it grounds level with 3-shot ($+0.5406$, $p = 0.41$). Its personas are more dispersed than the address's (diversity 5.60 against 3.33, recorded speech 4.20) and vary along fewer directions (4.05 effective dimensions, the participation ratio over the ten habits, against 7.30, recorded speech 5.24). Its lead over the address in style separation (§5.3) is $+0.0860$ ($p < 0.001$; Appendix B.10), and the 95% interval of its lead in collective gain is $+0.23$ to $+10.90$. Its personas together give more ideas than the address's ($\gamma = 50.20$ against 39.62) at about the same number per persona ($\alpha = 1.90$ against 1.89).

12-shot shows up to twelve of the speaker's turns in the 3-shot scaffold, and the style description places in the system prompt a description of the speaker's register that Qwen2.5-7B-Instruct wrote from those turns (Appendix A.3); we score both in a second run over the same speakers. 12-shot grounds at $+0.6128$, above 3-shot by $+0.0829$ ($p < 0.001$), and against held-out episodes at $+0.3498$, above 3-shot by $+0.0412$ ($p = 0.015$). The style description grounds at $+0.3565$, above the plain system prompt by $+0.1181$ and above the address by $+0.1305$ (both $p < 0.001$), and against held-out episodes at $+0.2393$, above the address by $+0.1227$ ($p < 0.001$).

B.6 Equivalence and the empty prompt

We test whether the address and a name ground personas equally with two one-sided tests on the paired speaker bootstrap of the correlation difference, at bounds of ± 0.05, about half the address's range across seeds, and ± 0.10 (Table 10). Against training episodes the address and the name are equivalent at ± 0.05 on all ten habits, and every other training comparison is equivalent only at ± 0.10. Against held-out episodes no comparison is equivalent at either bound, and the name and the system prompt ground above the address. One of the 14 shared prompts is empty, the format of the address's plain training rows, and Table 11 removes it: the address then grounds below the name within sampling error, and address + name still exceeds the name.

Table 9: **Given the same fine-tuning, a name written as a transcript label and prompts with more of the speaker's data ground personas better than the address, and without the fine-tuning every text condition grounds less.** Grounding against training and held-out episodes, and diversity, per condition on the shared prompts. The name-label model is the reported recipe with each speaker's name written in the address token's place; 12-shot and the style description are scored in a second run over the same speakers.

model	condition	grounding		diversity
		training ($n = 632$)	held-out ($n = 540$)	
Qwen2.5-7B, untuned	Name	$+0.0997$	$+0.0939$	1.92
	System prompt	$+0.1514$	$+0.1199$	1.71
	3-shot	$+0.4436$	$+0.2494$	3.66
Qwen2.5-7B-Instruct, untuned	Name	$+0.1147$	$+0.0688$	1.04
	System prompt	$+0.1688$	$+0.1305$	2.00
	3-shot	$+0.1895$	$+0.1385$	1.58
Name-label model	Name label	$+0.3315$	$+0.2260$	5.60
	Name label + 3-shot	$+0.5406$	$+0.3290$	4.82
Reported model	Name	$+0.2204$	$+0.1773$	2.62
	System prompt	$+0.2385$	$+0.1868$	3.19
	Style description	$+0.3565$	$+0.2393$	4.42
	3-shot	$+0.5299$	$+0.3086$	3.71
	12-shot	$+0.6128$	$+0.3498$	3.64
	Address	$+0.2260$	$+0.1166$	3.33
	Address + name	$+0.3056$	$+0.2073$	3.73
	Address + 3-shot	$+0.5082$	$+0.2919$	4.65

Table 10: **The address and a name ground personas equivalently only against training episodes.** Paired speaker-bootstrap difference in grounding between the address and each text condition, its 90% interval, and whether two one-sided tests declare the conditions equivalent within each bound, with the larger of the two one-sided p in parentheses.

target	contrast	Δ	90% interval	equivalent at ± 0.05	at ± 0.10
training episodes, all ten	address − name	$+0.0056$	$[-0.0360, +0.0466]$	yes (0.037)	yes (< 0.0005)
training episodes, all ten	address − system prompt	-0.0125	$[-0.0548, +0.0280]$	no (0.069)	yes (0.001)
training episodes, deleted	address − name	$+0.0163$	$[-0.0400, +0.0713]$	no (0.162)	yes (0.009)
training episodes, deleted	address − system prompt	-0.0280	$[-0.0853, +0.0279]$	no (0.256)	yes (0.020)
held-out episodes, all ten	address − name	-0.0607	$[-0.1001, -0.0208]$	no (0.657)	no (0.051)
held-out episodes, all ten	address − system prompt	-0.0702	$[-0.1066, -0.0289]$	no (0.792)	no (0.087)
held-out episodes, deleted	address − name	-0.1070	$[-0.1598, -0.0497]$	no (0.950)	no (0.561)
held-out episodes, deleted	address − system prompt	-0.1061	$[-0.1599, -0.0467]$	no (0.939)	no (0.552)

Table 11: **Without the empty prompt the address stays level with a name, and every result for address + name and address + 3-shot holds.** Paired differences in grounding on all ten habits over the 631 speakers every condition profiles, on all 14 shared prompts and with the empty prompt removed.

contrast (all ten habits)	all 14 prompts	p	without the empty prompt	p
address − name	$+0.0053$	0.809	-0.0102	0.684
address − system prompt	-0.0129	0.611	-0.0161	0.508
address + name − name	$+0.0853$	0.001	$+0.0778$	0.001
address + name − address	$+0.0800$	< 0.001	$+0.0880$	< 0.001
address + 3-shot − 3-shot	-0.0219	0.196	-0.0150	0.411

B.7 GROUNDED AND ARBITRARY SPREAD

We split each condition's between-speaker variance on involvement into a grounded part, which the speaker's recorded involvement predicts, an arbitrary part, and measurement noise estimated from split-half reliability, all in units of the noise-free recorded variance (Table 12). Adding the address to 3-shot raises both parts and leaves the noise small; the same holds against held-out episodes. The slope toward one therefore reflects spread that follows the recorded differences, and the rest of the added spread is variation particular to each persona. With each condition's centroid moved onto the recorded one, §5.2 gives the median distance from a persona to its own speaker under 3-shot and the address, and from the recorded mean. It is 0.86 under address + 3-shot, 0.89 under a name, 0.92 under a system prompt, 0.94 under address + name and 1.06 under other's 3-shot, so no condition places a persona nearer its speaker than the recorded mean does.

Table 12: **The address adds spread that follows the recorded differences between speakers and a share that varies by persona.** Involvement per condition: the slope of generated on recorded involvement, its r^2, and the between-speaker variance split into a grounded part predicted by recorded involvement, an arbitrary part, and measurement noise from split-half reliability, in units of the noise-free recorded variance, against training and held-out episodes.

target	condition	slope	r^2	grounded	arbitrary	noise
training	Name	0.24	0.18	0.07	0.21	0.07
training	System prompt	0.37	0.21	0.15	0.43	0.09
training	3-shot	0.61	0.61	0.42	0.17	0.06
training	Address	0.23	0.10	0.06	0.37	0.14
training	Address + name	0.46	0.28	0.24	0.47	0.10
training	Address + 3-shot	0.78	0.52	0.68	0.48	0.08
held-out	Name	0.19	0.14	0.05	0.18	0.06
held-out	System prompt	0.26	0.13	0.08	0.41	0.08
held-out	3-shot	0.37	0.29	0.18	0.31	0.05
held-out	Address	0.15	0.05	0.03	0.35	0.12
held-out	Address + name	0.33	0.18	0.14	0.45	0.08
held-out	Address + 3-shot	0.50	0.27	0.32	0.66	0.07

B.8 GUIDANCE ON THE ADDRESS ALONE

Figure 6 shows the guidance sweep of §5.4 at every weight.

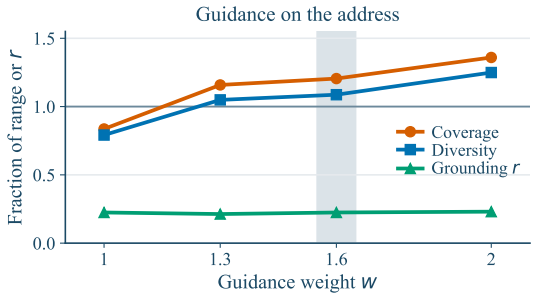


Figure 6: **Guidance on the address alone raises coverage and diversity and leaves grounding constant.** Coverage and diversity as fractions of the recorded speakers' values (line at one), and grounding, against the guidance weight w on the address alone ($w = 1$ is the address unguided), over 623 to 633 speakers per weight; the band marks $w = 1.6$.

B.9 COLLECTIVE GAIN ON USABLE, ON-TOPIC ANSWERS

Table 13 separates α and γ and recomputes β on answers that are usable, by the definition of Appendix B.14, and on the item's topic. An answer is on topic when its embedding similarity to its own item exceeds the 95th percentile of its similarities to the other 99 items; by that criterion 67.8%

of NoveltyBench's human answers are on topic, so the criterion is strict. Every condition reaches between 1.83 and 1.89 distinct ideas per persona on its two draws, so β follows γ.

Table 13: **The address's lead in collective gain over a name and a system prompt persists after removing unusable and off-topic answers.** NoveltyBench per condition: the distinct ideas one persona gives on two draws (α), the ideas the personas give together (γ) and their ratio (β). The last columns give the share of unusable answers, the share on the item's topic, and β recomputed on usable, on-topic answers with its item-bootstrap 95% interval.

condition	α	γ	β	unusable	on topic	β, usable and on topic
No persona	1.83	1.83	1.00	9.0%	86.0%	—
System prompt	1.88	22.11	11.23	20.5%	90.9%	13.82 [10.83, 16.88]
Name	1.85	23.11	11.90	12.9%	54.8%	12.73 [9.58, 16.42]
3-shot	1.88	33.99	17.24	22.2%	84.9%	28.08 [21.21, 36.26]
Address	1.89	39.62	20.18	16.3%	74.3%	27.53 [20.51, 35.32]
Address + name	1.84	21.02	10.94	17.1%	44.4%	16.81 [12.23, 22.18]
Address + 3-shot	1.85	20.29	10.51	22.7%	60.7%	21.18 [15.28, 27.82]

B.10 Style separation at matched seeds

Every persona answers each INFINITY-CHAT query once per sampling seed, and every persona uses the same seeds, so two personas' answers at one seed share their sampling noise. The two answers of one persona always come from different seeds, so we compute the similarity between personas over pairs at different seeds as well; personas with identical answers, as under no persona, then score exactly zero. Counting every pair of two personas' answers, half of which share a seed, lowers every condition's separation and sets the no-persona value to -0.1948, because its same-seed pairs are identical texts. That version gives 0.1328 under the address, 0.0938 under address + 3-shot, 0.0549 under address + name, 0.0086 under 3-shot, -0.0063 under a system prompt and -0.0119 under a name. At matched seeds the three text conditions are level with one another, and in both versions the address separates this model's personas most and every condition with it exceeds every text condition. Over all pairs the name-label model of Appendix B.5 separates its personas by 0.2220.

B.11 Role and attribution

Voice verification assigns a turn by its voice, and an accepted turn can still contain another speaker's words, most often a host's introduction or question. A language-model reader read 120 turns in full from their transcripts, 60 that a text proxy flags as another speaker's and 60 from the rest, and judged whether most of each turn's words belong to the named speaker, and an author verified all 120 judgements. Weighted by the two strata's shares of the corpus, attribution precision is 0.780 (95% interval 0.676 to 0.857). Hosts ask questions and guests talk about themselves, so the deleted habits could track conversational role. We rank speakers by the share of their turns with host cues that use no question mark, first-person pronoun or contraction, and compare the top third with the bottom third (Table 14). Every condition grounds guest-like speakers at least as well as host-like ones, so the grounding holds within each role. Against held-out episodes the address grounds guest-like speakers at $+0.1803$ and host-like speakers at $+0.0439$, and host cues also mark turns misattributed from a host, so the address's held-out weakness concentrates where attribution is weakest.

Table 14: **Grounding holds within guest-like speakers, where host cues are rarest.** Grounding on all ten habits per condition within the third of speakers with the most host cues (host-like) and the third with the fewest (guest-like), against training and held-out episodes.

	training episodes		held-out episodes	
condition	host-like	guest-like	host-like	guest-like
Name	$+0.1937$	$+0.2352$	$+0.1510$	$+0.2285$
System prompt	$+0.2374$	$+0.2411$	$+0.1889$	$+0.1971$
3-shot	$+0.4701$	$+0.5712$	$+0.2691$	$+0.3268$
Address	$+0.1641$	$+0.2700$	$+0.0439$	$+0.1803$
Address + name	$+0.2591$	$+0.3039$	$+0.1254$	$+0.2776$
Address + 3-shot	$+0.4757$	$+0.5279$	$+0.1860$	$+0.3486$

B.12 CASE STUDIES

Tables 15, 16, 17 and 18 extend Table 1 to four further prompts and twelve further speakers, and print, verbatim, the combined condition of §5.4 beside its components. A fixed rule selects every row. We choose each table's three speakers as the trio from three professions with the widest register spread, and each speaker's sample as the one of four on which the name and the address with the turns differ most in register. For these four tables we draw the trio from a wider candidate pool than for Table 1.

Table 15: **The comparison of Table 1 on a further prompt and three further speakers.** Model-generated responses to the shared prompt *"Describe what it was like inside a castle during a medieval siege."* under the speaker's name, three of their recorded turns in the prompt and the address combined with those turns, beside one of the speaker's recorded turns. A fixed rule selects the speakers and samples (Appendix B.12), cut at 16 words. Generated cells in a row share a sampling seed.

speaker	Name	3-shot	Address + 3-shot	recorded turn
Coleman Hughes *pundit*	That's like a great answer. And what was it like? What was it like to be…	Well, that's a great question because I had an opportunity to spend some time at a…	You know, I'm not very good at visualizing these things because we've never been in a…	Though we have different experiences, we do live in an objective reality. There is such a…
Chris Gethard *actor and comedian*	That's like a castle, right? I mean, I think that's one of the most terrifying things.…	That's like a great question. And I had an opportunity to live inside one for a…	Yeah. So I'm going to say it was like being in an airplane. You know how…	I reached out to. So yeah. Or at least people you know in the sense that…
Pramila Jayapal *politician*	So if you had to go through that experience, and then you're told, Hey, let's do…	So if you're listening closely, folks, that thank you, thank you, thank you you heard at…	Congresswoman, I'd love to hear you describe what it was like inside the Capitol, the Senate…	Exactly. We'll get back to them as soon as we are able. Congressman, I want to…

Table 16: **The comparison of Table 1 on a further prompt and three further speakers.** Model-generated responses to the shared prompt *"Who do you admire, and why?"* under the speaker's name, three of their recorded turns in the prompt and the address combined with those turns, beside one of the speaker's recorded turns. A fixed rule selects the speakers and samples (Appendix B.12), cut at 16 words. Generated cells in a row share a sampling seed.

speaker	Name	3-shot	Address + 3-shot	recorded turn
Stacey Abrams *politician*	I admire Stacey Abrams because she ran for governor in Georgia without a single poll showing…	I admire so many people, but there are two in particular who I think of often…	I. I mean, I admire my kids. But I also admire the work of people who…	We want to try to take complicated issues and untangle them for our audience. We want…
Elie Mystal *political pundit & lawyer*	Elie Mystal, my favorite person on the internet, my favorite writer on the internet, who I…	I like the great women's rights advocates like Elizabeth Cady Stanton and Susan B. Anthony and…	And like, you know, the people that I'm talking about are people like me, you know?…	Hi, Brian. How are you? In New Jersey, you can't flip off a police officer, which…
Matt Mccusker *stand-up comedian*	I admire so many people, but one of the ones that I really think is underrated…	I admire, um, a lot of people. But I admire, um, um, people like Martin Luther…	I'm so glad you said that. Yeah. Matt McCusker is the greatest human being. I've ever…	More war in Gaza, honestly. Yeah, I haven't caught a lot of the finals. Yeah. I'm…

B.13 ADDING NEW SPEAKERS

The additions of §5.5 take the reported model from 650 to 733 addresses. Figure 7 shows the four models of §5.5, each trained by Stage 1 on part of the population and extended to the rest by Stage 2.

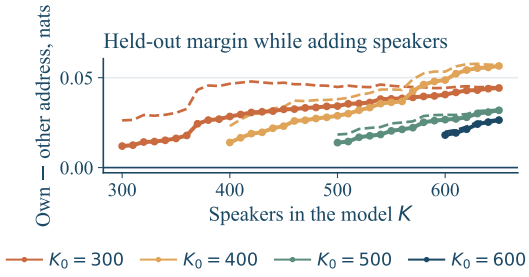
The added speakers have no Stage-1 address, so we cannot compare the two stages on their margin alone. The four models above add by Stage 2 speakers whom the reported model trained by Stage 1, so we score both stages' addresses for those speakers on the same held-out rows against the same

Table 17: **The comparison of Table 1 on a further prompt and three further speakers.** Model-generated responses to the shared prompt *"If a colour had a smell, what would blue smell like?"* under the speaker's name, three of their recorded turns in the prompt and the address combined with those turns, beside one of the speaker's recorded turns. A fixed rule selects the speakers and samples (Appendix B.12), cut at 16 words. Generated cells in a row share a sampling seed.

speaker	Name	3-shot	Address + 3-shot	recorded turn
Abigail Spanberger *politician*	Blue is the colour of the sky and the sea. So if it had a smell, …	Blue is the smell of lavender, a calming smell. I think it's the smell of the …	Blue is the colour of the sky and the ocean. So if it had a smell, …	I'm a Democrat because I believe in our policies. And I'm a Democrat because in the …
Marc Maron *comedian*	It would be like a lot of things, I think, but it would be like a …	It would be like, I don't know, the ocean. Blue ocean. Yeah, that's a good one. …	It would be like what's it like when you're like really depressed but not really but …	Windscreens look a little filthy. They're fucking old, right? I know, dude. Why don't you get …
Adam Liptak *journalist*	It would be like a lot of things, I think, but it would be like a …	It would smell like the ocean. But, you know, I'm not sure I would say that …	It's not clear what's at stake in the battle over the Supreme Court's next nominee. The …	From the New York Times, I'm Michael Bavaro. This is The Daily. When the Supreme Court …

Table 18: **The comparison of Table 1 on a further prompt and three further speakers.** Model-generated responses to the shared prompt *"What makes something funny?"* under the speaker's name, three of their recorded turns in the prompt and the address combined with those turns, beside one of the speaker's recorded turns. A fixed rule selects the speakers and samples (Appendix B.12), cut at 16 words. Generated cells in a row share a sampling seed.

speaker	Name	3-shot	Address + 3-shot	recorded turn
Colin Allred *politician and former NFL player*	I think humor is, it's such a wonderful word to say, but I think humor is …	I think that that is an interesting question. I think that it depends on the person. …	I think for me, it's, you know, I guess the best kind of humor is when …	I agree, particularly in the South. You know, red state bulwark America is good. We have …
Andy Ngo *right-wing social-media influencer*	You know, that's an interesting question. I don't know the answer. It's something that is completely …	Um, for me, it's, you know, I guess the best kind of humor is when people …	Okay. So that's how we're going to start. Um, so let's start first with how did …	So what were they doing when you say they were running around committing crimes? And to …
Sam Jay *comedian and writer*	I think humor is, it's such a weird word to use, but I think humor is …	I think for me, it's always the most simple stuff. And I think that's where you …	It's a good question. I think that's what I'm trying to figure out myself. You know …	Oh, well, I thought about it. And when we were kind of coming up with names, …



K_0	added	all	present	own > other	drift
300	350 of 350	+0.0443	+0.0446	261/350	$+5e-05$
400	250 of 250	+0.0565	+0.0572	216/250	$+6e-07$
500	150 of 150	+0.0319	+0.0323	125/150	$+1e-06$
600	50 of 50	+0.0264	+0.0268	35/50	$+3e-07$

Figure 7: **A model first trained on K_0 speakers adds the rest with every existing address unchanged. Left:** the held-out likelihood margin of a speaker's own address over another speaker's, after every ten additions, averaged over all 650 speakers with zero for those not yet added (solid; all) and over the speakers with an address (dashed; present). **Right:** the same after the last addition, the count of added speakers whose own address scores higher, and the likelihood drift of the first K_0 speakers.

partner, a Stage-1 speaker of every model. The margin, one speaker's swap gap (Appendix B.14), is the held-out loss under the partner's address minus that under the speaker's own. On rows from episodes neither stage fitted, the median paired difference between the Stage-2 and the Stage-1 margin is $+0.0053$, $+0.0020$, $+0.0104$ and -0.0015 nats for the models trained on 300, 400, 500 and 600 speakers (78, 57, 33 and 12 speakers, every sign-test $p \geq 0.09$). On the same speakers we therefore cannot distinguish the margin of an address Stage 2 adds from that of the address Stage 1 trained. Counting rows from episodes that also supplied Stage-2 fitting rows raises the Stage-2 margin (median differences $+0.0111$ to $+0.0184$ over all held-out rows), because the growth split cuts a speaker's rows by count and can place one episode on both sides.

B.14 The address as a parameter

The *swap gap* is the median over speakers of the held-out loss of a speaker's text under another speaker's identity, drawn uniformly from the model's speakers, minus the loss under their own, and we exchange the identity alone. We test two further properties of a row. The first is its swap gap beside three parameter-efficient containers (Appendix B.17). The second is a ridge fitted out of fold from the 640 human addresses to their habits (Figure 8), whose direction for involvement steers generation in §5.5 (Figure 9).

We read each habit from the addresses with a ridge at $\alpha = 100$ over five shuffled folds under a fixed seed, with addresses standardised per dimension. We score each read as the correlation between held-out prediction and target, test it against a 20,000-draw permutation null that refits on shuffled targets, and correct by Benjamini–Hochberg over the ten habits. Of the ten habits, 9 decode out of fold in every one of the three runs of the recipe, and all ten in the first two; the exception is questions in the third run ($+0.1208$ to $+0.0726$; Figure 8). The decoded habits are therefore a property of the recipe.

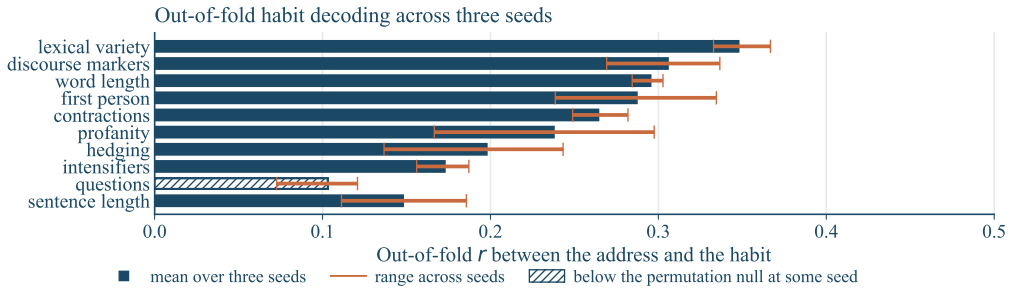


Figure 8: **The address predicts all ten habits in two of the three training runs of the recipe, and every habit except questions in all three.** Each bar is the correlation r between a habit and its prediction from the address by a ridge regression, fitted on the other folds of a five-fold split of the speakers. Bars give the mean over three training runs of the recipe at different seeds, and error bars the range across runs. A hatched bar fails a 20,000-draw permutation test, corrected over the ten habits by Benjamini–Hochberg, in at least one run, and every other bar passes it in all three.

The address encodes *who* a speaker is only through register. Fourteen binarised Wikidata attributes decode out of fold over the 588 speakers with a record at a median AUC of 0.570, the largest at 0.704, and birth year over 536 speakers at $r = 0.162$. Regressing the ten habits out brings every one to chance: the median falls to 0.4404, where a matched removal built from other speakers' habits leaves 0.5606, and a second seed reproduces the pair (0.4757 against 0.6154). A logistic model reading only character n-grams of the name string decodes gender at AUC 0.807 and the other attributes weakly (median 0.544 over 36 attributes, 7 above 0.60), so an empty label is an adequate null for a name on all attributes besides gender.

We fit two addresses for each of 30 speakers on disjoint halves of their rows; they agree at a median cosine of 0.5911 against 0.5640 for different speakers, the similarity that the shared initialisation alone produces. A refitted address reproduces its register readout, the habits a ridge reads from it, at a different coordinate. A new speaker therefore enters the model by Stage 2, which fits their address on their own speech, and is issued with their name and turns.

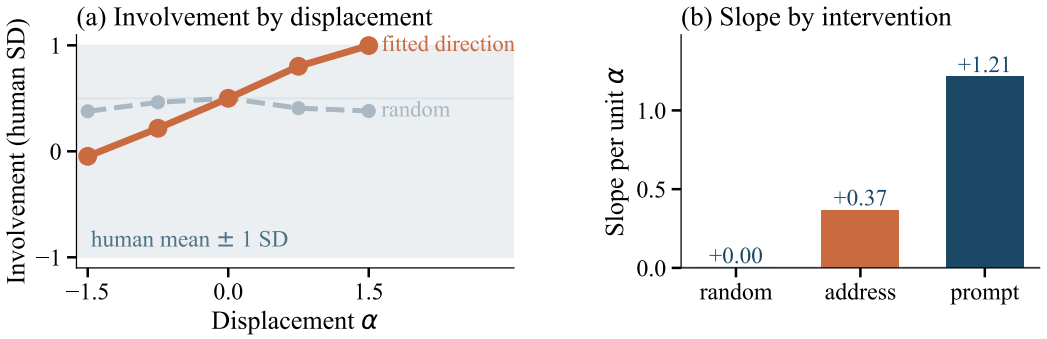


Figure 9: **Moving every address along the direction that predicts involvement moves the involvement of the speakers' generated text in step, and a random direction of the same length does not. (a)** Median generated involvement of the 626 scored speakers as each address moves α median address lengths along one of two directions: the unit weight vector of a ridge regression that predicts involvement from the address, or a random direction of the same length. The band spans one human standard deviation either side of the human mean. **(b)** The same axis moved three ways, in human standard deviations per unit of α: a random direction and the fitted displacement of §5.5, and a five-level instruction from formal to casual ($+1.2135$), its levels placed at the five values of α.

The steering of §5.5 holds over every speaker. A generation is unusable when it is nearly empty, has a substantial share of non-ASCII characters, or consists largely of repeated words. Restricted to the displacements whose unusable share stays at or below the undisplaced address's 16.4%, the difference is $+0.2894$ over 483 of the 619 speakers with a usability measurement, so the effect arises in usable text. The four habits the direction loads most move in the loaded direction (contractions $+0.487$, first person $+0.349$, discourse markers $+0.264$, mean word length -0.270), while the ones it barely loads move an order of magnitude less.

B.15 Controls on the measurement

We take diversity and coverage against a human baseline resampled to the same number of speakers, because estimation bias dominates at small samples.

The other's-address condition of §5.1 pairs each speaker with one other speaker, and the reported grid draws that partner by a stride along the speaker list, which is close to corpus-size order. We regenerate the whole comparison with the partner drawn uniformly from the model's speakers at a fixed seed, over the 626 speakers it profiles. Own minus other is $+0.2017$ on the deleted habits (95% interval $+0.1302$ to $+0.2704$) and $+0.1994$ on all ten, against $+0.2018$ and $+0.1954$ under the reported rule, so the control gives the same result under either partner rule.

The fraction of the human volume a set of personas fills is a generalised variance, which is biased downward at small samples. When we draw n points from a Gaussian with the covariance of the 640 recorded profiles, the estimator recovers 18.6% of the known volume at $n = 10$, 76.0% at $n = 15$, 94.3% at $n = 50$ and 99.7% at $n = 640$. Divided by the full-sample human figure, the ten LLMs would cover 0.167 of the human range. Divided by the volume ten people reach, they cover 1.018, on an interval of $[0.83, 1.36]$ that contains one. We therefore take every coverage in §5.2 against a human baseline resampled to the same n.

We run three checks for whether grounding is an artefact of the pairing, of corpus size or of audio quality. Shuffling which speaker each generated profile belongs to reduces the address's grounding to $+0.0001$. Corpus size decodes from an address better than any habit. When we regress corpus size out of both profiles, the address grounds at $+0.2266$ and its gap over a name is $+0.0065$. When we project it out inside each training fold, habit decoding is unchanged, and volume alone accounts for at most $+0.0965$ on any habit. The own-minus-other difference of $+0.1954$ on all ten habits has a 95% interval of $[+0.1476, +0.2428]$, and the address's $+0.0056$ over a name one of $[-0.0447, +0.0535]$. We verify identity by voice, and every segment has a voice-verification score. On this corpus two of the ten habits correlate with that score across speakers at $|r| > 0.2$

(contractions $+0.251$, discourse markers $+0.215$). Regressing the score out of both profiles over the 614 speakers that have one leaves the own-minus-other difference within its interval, $+0.2046$ becoming $+0.1910$ on the deleted habits and $+0.2090$ becoming $+0.2006$ on all ten (both $p < 0.001$).

To measure how repeatable a profile from 28 generations is, we split each speaker's generations into disjoint halves. This gives a Spearman–Brown reliability of 0.4877 on the deleted habits and 0.4784 over all ten habits, and the same split on recorded speech gives 0.8004 and 0.8057. Grounding on the deleted habits is therefore bounded above by $\sqrt{0.4877 \times 0.8004} = 0.6248$. The own address's 0.2150 is 34% of that bound. Habits differ in how repeatably they can be measured from 28 generations, from 0.7296 for sentence length to 0.0650 for the hedging rate, so part of the spread in decodability is measurement noise.

We score two further runs of the reported recipe, at other seeds, on the same designs over every speaker. The second seed gives an own-minus-other grounding of $+0.1706$ on the deleted habits and $+0.1576$ on all ten habits over 626 speakers, against the reported run's $+0.2018$ and $+0.1954$ (Figure 10, where the light bars give the range). The third seed gives $+0.0781$ and $+0.0864$ over 611 speakers, both above zero ($p = 0.048$ on the deleted habits and 0.0008 on all ten). On the shared prompts at both further seeds address + name stays above a name and three of the speaker's own turns stay first, while the address alone stays level with a name (Table 5). Two comparisons change at seed 3: adding the address to 3-shot lowers grounding (-0.0382, $p = 0.022$), and the system prompt grounds above the address ($+0.0503$, $p = 0.046$). At seeds 0 and 1 both lie within sampling error.

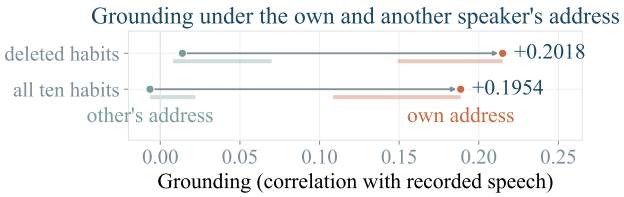


Figure 10: **Text generated under a speaker's own address reproduces the register of their recorded speech, and text generated under another speaker's address reproduces none of it.** Grounding (§4) on the deleted habits and on all ten under the speaker's own address and another's; arrows give the difference, and the light bars the range across three training runs of the recipe at different seeds.

Every p-value is two-sided. We fix six confirmatory tests with their control conditions before scoring the model (grounding on the deleted habits and on all ten, grounding against a name and against a system prompt, the swap gap, and steering against a random direction). Benjamini–Hochberg over the six leaves all four claimed positives below 0.05, the largest adjusted to 0.0087, and Bonferroni also keeps them below that level, the largest at 0.0348. We correct each searched family within itself: the ten habit decodings and the fifteen demographic readouts (fourteen attributes and birth year). Everything else is descriptive.

B.16 The shared prompts under learned encoders

We recompute the grid of §5.2 with the ten habits replaced by the STYLE-EMBEDDING encoder (768 dimensions): three of the speaker's own turns stay first, address + 3-shot second, and the address alone ranks above a system prompt and a name ($+0.1935$ against $+0.1076$ and $+0.0829$). Replacing them with the semantic BGE-SMALL-EN-V1.5 encoder (384 dimensions) moves a name above the address ($+0.3711$ against $+0.2027$). We expect a semantic encoder to favour a name, since a name retrieves pretraining's knowledge *about* a person. A style encoder reads the address's contribution as manner, and a semantic encoder reads little of it.

B.17 The address beside other containers

We compare the address with three per-speaker containers. LoRA (Hu et al., 2022) is rank 1 with scaling 1 and no dropout, on the query, key, value and output projections of every layer. Prefix

tuning (Li & Liang, 2021) gives every layer four learned virtual tokens as past keys and values. IA^3 (Liu et al., 2022) is a learned rescaling; ours rescales the outputs of the query and output projections alone, and applies half of the parameters it allocates. Each trains with the backbone for the same 4,753 steps, its parameters at 10^{-4} and the backbone, input embedding included, at 3×10^{-6}; the address model trains its whole input embedding at 10^{-4}. Prefix tuning and IA^3, with $32\times$ and $84\times$ the address's parameters, reach smaller swap gaps than the address, and LoRA, which models the text 0.65 nats worse, has a larger one (Table 19). At matched likelihood each alternative matches or exceeds the address's swap gap (Table 20).

In Table 19 we draw the partner uniformly from the model's speakers, following the definition of the swap gap in Appendix B.14. The earlier design drew it seven positions along a stride-sampled list, which is in corpus-size order (list position and row count correlate at Spearman -0.9352). The partner was therefore systematically the poorer-recorded of the pair, which biases a likelihood comparison. Under that rule over the same 605 speakers the address's median gap is $+0.0063$ with 366 positive, and the ordering of the containers is unchanged: the address leads prefix tuning and IA^3 ($+0.0043$ and $+0.0028$ paired) and LoRA leads the address (-0.0481). Paired per speaker under the uniform rule, over the 323 speakers whose rows fit under the token cap for every condition, the address matches a name and a system prompt within sampling error (-0.0035, $p = 0.32$; -0.0041, $p = 0.095$) and falls below 4-shot (-0.0115, $p = 0.0012$). On held-out likelihood the address discriminates its speaker as a prompt does. The address's advantages lie in the register it encodes (§5.1) and in its size (Table 19).

Table 19: **At its reported learning rate the address has a larger swap gap than prefix tuning and IA^3 with a fraction of their parameters, and a smaller one than LoRA, whose own-text likelihood is worse.** IA^3 applies half of the parameters listed for it. At matched likelihood each alternative has at least the address's swap gap (Table 20). The swap gap (Appendix B.14) is taken over 605 speakers, with each partner drawn uniformly, in nats. A star marks a gap whose paired difference from the address has a bootstrap interval that excludes zero. Own NLL is the negative log-likelihood of a speaker's held-out text under their own container.

container	params	median gap	own NLL
Address (ours)	3,584	$+0.0088$	2.195
Prefix tuning	114,688 (32×)	$+0.0018^*$	2.240
IA^3	301,056 (84×)	$+0.0029^*$	2.182
LoRA, rank 1	630,784 (176×)	$+0.0605^*$	2.842

Table 20: **Every container's held-out likelihood worsens as its learning rate rises, and its swap gap rises with the rate in three of the four.** Swap gap and own NLL as in Table 19. The 232 speakers of the learning-rate sweep come from a sample taken at a fixed stride along the speaker list, which is ordered by corpus size, and each speaker's partner lies a fixed number of places further along it. LoRA's gap falls only at 10^{-3}, where its likelihood has diverged.

container	lr	median gap	own NLL
Address (ours)	10^{-4}	$+0.0086$	2.194
	3×10^{-4}	$+0.0149$	2.535
	10^{-3}	$+0.0219$	2.765
IA^3	10^{-4}	$+0.0034$	2.165
	3×10^{-4}	$+0.0101$	2.198
	10^{-3}	$+0.0245$	2.527
LoRA, rank 1	10^{-4}	$+0.0529$	2.750
	3×10^{-4}	$+0.4974$	4.056
	10^{-3}	$+0.1267$	8.036
Prefix tuning	10^{-4}	$+0.0018$	2.228
	3×10^{-4}	$+0.0060$	2.246
	10^{-3}	$+0.0180$	2.277

We refit each container at 3×10^{-4} and 10^{-3} as well as at its reported rate (Table 20). Held-out likelihood worsens with the learning rate in all four, and the gap rises with it in three. LoRA is the

exception at 10^{-3}, where it has diverged to an own NLL of 8.036. At 3×10^{-4} LoRA has fifty-eight times the reported address's gap while modelling the text 1.86 nats worse, so we report each gap beside its held-out likelihood, as Table 19 does.

B.18 Released LLMs on the human axes

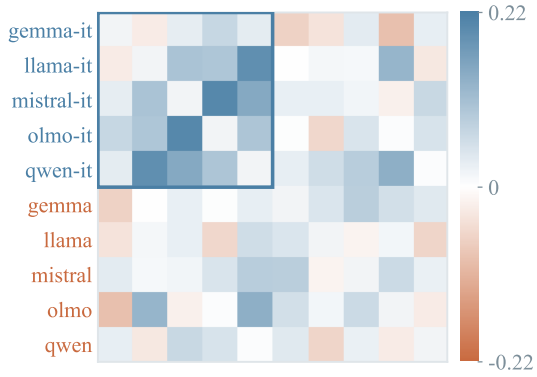


Figure 11: **The five instruction-tuned LLMs' addresses agree with each other more than any draw of five people's does.** Cosine agreement of the ten LLMs' addresses, centred on the mean human address, the five instruction-tuned ones first and outlined; blue is above the human mean, orange below.

Both group means of §5.6 lie inside the human 5th–95th percentile range, and three of the five instruction-tuned LLMs lie below the human 5th percentile. Together the ten span 2.81 effective dimensions ($[1.71, 3.09]$; the participation ratio over the ten habits), against 3.62 ($[2.55, 5.00]$) for ten people. At $n = 10$ the spread of ten LLMs and that of ten people are therefore statistically indistinguishable. Measured from the mean human address, the instruction-tuned LLMs' addresses agree at a mean cosine of $+0.0942$, against $+0.0215$ for the base LLMs and -0.0023 for a matched draw of people. No draw of 2,000 reaches the instruction-tuned figure, while 153 reach the base one (Figure 11). 97.5% of the LLMs' spread on the two factors lies along one direction, a larger share than in any of 4,000 draws of ten people (ten people reach a median of 72.7%). Over the LLMs the two factors correlate at -0.74 pooled, $+0.62$ within base and -0.06 within tuned, so their pooled coupling comes from the difference between base and tuned LLMs. Fitted on people and LLMs pooled, involvement explains 32% of the variance and elaboration 18%. Base LLMs exceed people most on contractions ($+2.5$ human standard deviations) and instruction-tuned ones on word length ($+2.4$).

We test whether the gap follows the chat template. The framing instruction, under which every LLM answers the interlocutor turns, reaches each LLM through its chat template where the tokeniser provides one and as a raw prefix otherwise. This agrees with regime on 9 of the 10: all five instruction-tuned LLMs use a template, and four of five base ones use a raw prefix. The exception is Qwen2.5-7B, a base release that provides a template. When we regenerate the LLMs without their templates through one pipeline, the four instruction-tuned ones that permit it move $+1.73$, $+2.07$, $+2.23$ and $+3.02$ human standard deviations toward the involved pole (Figure 5b), a mean of $+2.26$, and Qwen2.5-7B moves from $+0.50$ to $+1.42$. Given the raw prefix, base and instruction-tuned LLMs lie close together, base at $+1.82$ over three and instruction-tuned at $+1.23$ over four, a gap of 0.59 against 2.92 for the same LLMs as released. We exclude Gemma-2-9b-it, since without its template only 18 of 40 generations are usable first-person text. Regeneration with each template left in place moves an LLM by -0.21 to $+0.95$, less than any instruction-tuned LLM moves without its template. We also regenerate the four under a *different* family's template, which changes the interface and keeps a template in place. They then move by $+0.45$, -0.06, -0.16 and -0.22 of their no-template move, a mean fraction of 0.00, with 2, 9, 4 and 0 of 40 generations rejected. Tuned behaviour therefore appears under any chat template, including another family's, so the gap follows the presence of a chat template.

As a population the LLMs separate from the people by position. We project every address on the direction between the two group means, fitted without the address it scores. The ten model addresses separate from the 640 human ones at held-out AUC 0.824, and $2{,}000$ label permutations that redo the whole procedure give $p = 0.014$ (permutation 95th percentile 0.764). The imbalance of ten points against 640 makes an AUC of that size attainable, so we read it as a separation of positions and report no effect size. The separation persists after unit-normalising every address (AUC 0.808), so the difference lies in the direction of a row. The two populations have the same spread within sampling error (mean pairwise cosine $+0.1156$ for people against $+0.1205$ for the models, $p = 0.62$): they differ in position and agree in spread.

C ETHICS

All speakers are public figures and all source material is published podcast audio. We use no private communication, and we verify every identity by voice. The model makes a person's manner of speaking reproducible. That capability makes it useful for measurement and could also support impersonation. We rely on one asymmetry: the address conditions manner, and its output is detectable. Its effect on knowledge and on claims of fact is untested. When we cut both sides to exactly 88 words, an ordinary TF–IDF classifier separates address-generated text from the speaker's own recorded speech at AUC 0.9910, and at a median 0.9978 within each of 635 speakers scored out of fold, 630 of them at or above 0.900. Part of that separability may come from off-topic or degenerate generations, so it describes this output and guarantees nothing about better generators. The model also generates opinions under the names of real politicians, and an address labelled with a name makes such text easy to produce. The address also creates a residual risk: party affiliation and gender are weakly decodable from an address (Appendix B.14), so a released model permits weak demographic inference, which depends entirely on register, since that demography disappears once register is regressed out. We release the trained addresses and the analysis artefacts. We withhold the speech rows, which identify their speaker. The 640 speakers are English-speaking public figures who appear on podcasts, and our conclusions about how people talk apply to that population.