Epistemic Bridges

Do Language Models Reach the Connections People Make?

Sofia G. Gallego[1]

[1]CosmoVentures, Paris, France, with Apart Research

Apart Research AI × Epistemics Research Sprint, November 2026

Abstract

Creativity often happens in the space between ideas, where concepts that seem unrelated turn out to connect. We ask how much of the space of connections that people make language models reach. In the web game Bridges, players link two words with no evident connection, in any form. We collected 145 bridges from 27 players and gave the same pairs to 15 open models from 2019 to 2025, with personas, creative conditions and different temperatures. The human answers draw a map of where people go, and each model is described, without any judge, by how much of that map it reaches (recall) and how many of its answers land on it (precision). Measured as written, chat training seems to close the map: the same 7B model reaches about half of it before chat training and almost none after. Most of that gap is the voice, one shared way of speaking. With the voice removed from every answer in the same way, the larger and newer chat models still reach less of the human map of ideas (recall 47% to 59%, against 81% for people), no model comes near people on both counts, and chat models stay closer to the word they start from. Only one condition widens the ideas on every model: answering as an ordinary person drawn at random. As more people think with AI, measuring diversity needs the voice and the ideas apart, and keeping ideas diverse may need models with more perspective as well as better reasoning.

Keywords: epistemic diversity, creativity, human–AI interaction, language models, chat training, personas, homogenization

1 Introduction

How can people and AI think together in a way that keeps the space of ideas open, and even widens it? Any concept can be described in many ways: a physicist, a poet and a cook each bring a different perspective to the same word. Much of creativity, innovation and improvement happens between those perspectives, in the space between ideas that seem connected or not. A harmonic interaction between humans and machines would preserve that space, and enrich it where it can.

More and more thinking now passes through AI tools, which learn from what is already written. If those tools pull everyone toward the same connections, each person may feel more creative while the group as a whole explores less. Recent studies point that way from two directions. On the model side, chat training reduces the variety of what a model writes [1], and models are more alike in their creative output than people are [2]. On the human side, help from AI can improve each person's ideas while making the ideas of a group more alike [3, 4]. Both lines of work mostly ask *how much* answers vary. We ask a question closer to the space itself: *how much of the ground that people cover do models reach, and what do they miss?*

To answer it, a measure needs a human reference and a task anyone can do. Reaching far is easy: typewriting monkeys would, given time, write every connection people make, and much else. So the question has two sides, how much of the ground people cover a model reaches (recall), and how much of what it writes lands there (precision). This paper reports a small, exploratory study built on Bridges, a game where people link two simple ideas in any form (mapideas.org/bridges). The pairs are simple and universal on purpose, so nobody needs to know a topic to play, and the game measures the act of connecting itself. The same pairs then go to language models of different ages, sizes and training. It continues earlier work on false epistemic redundancy, which found that AI personas that look different can share the same blind spots [5]. Three questions guide the study:

1. How much of the human space of connections do models reach, how does that change with the age of the model, its size and chat training, and how much of the difference lies in the voice and how much in the ideas?
2. Can framings, reasoning personas or creative ones, widen what a model reaches, and how do they compare with plain randomness?
3. Does it matter how abstract the two words are, or which one the connection starts from (is going from a bee to a mirror the same as going from a mirror to a bee)?

2 Related Work

Linking two fixed ends is an old game and an old idea. Lewis Carroll's *Doublets* (1879) turned one word into another one letter at a time, and Koestler later described creativity as the meeting of two unrelated frames of thought [6]. Psychology turned the idea into measures. The Remote Associates Test asks for the one word linking three distant words [7], and work on conceptual combination studied how people join two concepts through a shared property or a relation [8]. Word association norms go back to 1910, and Jenkins and Russell found that answers became more alike between 1910 and 1952 [9], an early record of a culture narrowing how people connect. Players navigating Wikipedia from one article to another pass through a few central hubs [10], and the daily game Linxicon asks players to chain single words between two random ones. Bridges keeps the form open, keeps the order of the two ideas, and puts people and models on the same pairs.

Measuring creative diversity has its own tools. The Divergent Association Task scores how far apart ten words are [11] and was recently used to compare about 100,000 people with language models, finding that models beat the average person but not the most creative ones [12]. Ecology offers ways to count diversity as an effective number of kinds [13], which the Vendi score brings to machine learning [14]. Precision and recall for generative models separate how close generated samples stay to real ones from how much of the real distribution they cover [15]; we use the same idea with human answers as the reference. Wright et al. benchmark epistemic diversity in model claims against an external reference [16], an approach we share, and Le Bronnec et al. found that chat models gain precision and lose recall against human text [24]. Diversity measures can mix how something is said with what is said. Tevet and Berant separate form from content and show that common metrics mostly follow form [25], and tuned models repeat sentence templates far more than people [26]. Different models also give strikingly similar answers to open questions [28].

Prompting can push back against narrow output. Personas of ordinary people raise the diversity of model-generated ideas [17], full backstories recover more human-like variety than role labels [18], multi-persona prompting widens design concepts [19], and asking a model to think step by step raises idea variance [20]. Most of this work measures variety among model outputs. Here every condition is measured against the same human map.

3 Data and Methods

3.1 The game and the players

Each round of Bridges shows two ideas joined by a dashed arrow, and the player writes how to get from one to the other, in any form: a sentence, a chain of words, a memory. Up to three stepping stones can be added. Six core pairs go to every player, two for each level of abstraction: two concrete things (bee and mirror, salt and clock), a concrete and an abstract one (river and secret, spider and melody), and two abstract ones (forgetting and festival, distance and habit). Each player sees each pair in one direction, chosen so that both directions get about the same number of players. After the six core pairs, players can stop or keep going with 24 more pairs, eight at each level (all pairs are listed in Appendix C). A ten-word warm-up comes first. Players can also ask to see how an AI connected the same pair; few did, 4 of 27 players for 9 of 145 bridges, too few to measure its effect. The game runs in five languages and needs no account.

The game was shared during the sprint through AI safety communities and personal networks. Between 25 October and 15 November 2026, 27 players wrote 145 bridges, 123 of them in English. Twenty-one players answered all six core pairs, and four kept going, with 8 more bridges on 8 of the other pairs. People stop when they want, while a model answers every pair it is given, so every comparison uses equal numbers of answers. The main analysis uses the 116 English answers to the six core pairs, from 21 players; the 21 Spanish and French answers, too few for maps of their own, join them in a check with all languages together. A typical answer has 11 words and takes about a minute. Most players who answered the optional questions use AI daily and work in engineering; a few work in the arts, sciences, health or social sciences, and one never uses AI.

3.2 Models and conditions

The same pairs, in the same words, then went to 15 open models (Table 1 in Appendix B), run with vLLM on Modal. Two are base models, GPT-2 XL (2019) and GPT-J 6B (2021). A base model is trained only to continue text it has read, so it simply carries on from the prompt. Chat models go through a second stage of training, here called *chat training* (instruction tuning and learning from human preferences), which teaches them to answer as a helpful assistant, with the kind of clear, safe and sensible answer that people rate highly. The chat models range from 2023 to 2025, mostly around 7 billion parameters, together with the Qwen2.5 family from 0.5 to 14 billion parameters and the base Qwen2.5 7B, so that size and chat training can each be seen on their own. Two larger 2025 models, Qwen3 32B and OpenAI's open gpt-oss 20B, answered in English with no persona, keeping only gpt-oss's final answer after its reasoning.

Each model first answered every pair, direction and language 10 times with no persona. Chat models then answered again with the four reasoning personas from false epistemic redundancy [5], each built on a documented weakness in model reasoning: a causal mechanist [21], an analogical thinker [22], a teleological analyst [23] and a dialectical thinker [29]. Because those personas were built for investigating, five chat models also answered with twelve creative conditions, built from the ways players connected ideas and from earlier work on prompting for diversity [17, 20]: five ways of connecting (association, story, memory, senses, wordplay), three moods (playful, dreamy, a child's eyes), an ordinary life drawn at random from 30 short sketches, a small random constraint, and thinking step by step. The same five models also answered as an ordinary life drawn at random holding one of the four reasoning styles, adapted to connecting ideas. Finally, two models answered with no persona at six temperatures from 0.4 to 2.0. All texts given to the models are in Appendix F, and sample answers are in Appendix G.

3.3 Terms and measures

No language model grades any answer. All measures come from placing answers in a space of meaning and counting. The terms used in the rest of the paper are defined here.

Space of meaning. Each answer is turned into a list of numbers by a small embedding model (all-MiniLM-L6-v2), so that answers saying similar things sit close together and different answers sit far apart. Before this step, each answer is read between the two words of its pair ("salt *[answer]* a clock"), so that answers that fill the gap between the two words read as full sentences. Answers as written are used in full; the embedding model reads up to about 200 words.
Voice and ideas. Chat answers wrap the connection in a shared scaffolding: they announce the task, number steps and sum up. To measure the ideas on their own, every answer, human or AI, is cleaned the same way: formatting goes, and sentences that only talk about the task are dropped (the phrases are listed in Appendix H). The cleaning is applied to the full answer, before any cut, and changes 1 in 10 player answers and 95% of chat answers. What remains is reduced to its content words (nouns, verbs and adjectives, without the two words of the pair), and each model answer keeps as many content words as a player's answer to the same pair. These are placed on the human map half against half, exactly like the human reference below; we call this the human map of ideas.
Repetition of the voice. How much a set of answers compresses when written as parts of speech [26], with each model answer, as written and with its scaffolding, cut to a player's length and as many answers as players gave. A set that repeats the same sentence patterns compresses more.
The human map. For each pair, the region of the space of meaning that the players' answers cover: a small sphere around each human answer, as wide as the distance to its third closest human neighbour [15, 24].
Precision. The share of a group's answers that fall inside the human map. A high value means the group writes the kind of connection people write.
Recall. The share of human answers that fall close to at least one of the group's answers, inside the group's own spheres. A high value means the group reaches most of the places people went.
Human reference. People cannot be scored against a map they drew themselves, since they would cover all of it. So the players are split at random into two halves: one half draws the map and the other half is placed on it exactly like a model, over 200 random splits.
Hub rate. The share of answers that take the most common human bridge for that pair, such as the hourglass for salt and clock, or walking for distance and habit.

Recall and precision are computed with as many answers from a group as there are human answers for that pair, averaged over 30 random draws and then over the six pairs, since a bigger group would reach more just by being bigger. The answers as written were also checked with two other embedding models (Appendix H).

4 Results

4.1 As written, chat models seem to close the map

The first step measures the answers as they come, in full; Figure 1 then places the same models with the voice removed (Section 4.2), and Table 1 in Appendix B gives both. GPT-2 and GPT-J reach about 60% of the human map (recall), yet only a third of their answers land on it (precision): coverage through noise, the typewriting monkeys of the introduction. Chat models reach much less. The clearest case keeps everything else the same: Qwen2.5 7B has a recall of 49% before chat

training and almost none after, and the larger and newer chat models 0% to 2%. This held with other embedding models and with all three languages together (Appendix H).

4.2 Voice or ideas

A look at what is being measured changes the picture. A typical chat answer, from Qwen2.5 7B, begins: "While the transition from a bee to a mirror may seem inexplicably bizarre on the surface, we can creatively construct a connection through a series of steps. . . ". Most chat answers carry this kind of announcement, numbered steps and a closing summary, so they all land on the same spot. Chat training gives the models a voice, and the voice is shared: at the same length and number of answers, the sentence patterns of every chat model repeat more than people's (4.6 to 5.9 against 4.1, the color in Figure 1), most of all for Qwen3 8B and the Qwen2.5 models from 7 to 14 billion parameters, while base models repeat less than people (3.7 to 4.4).

With the voice removed from every answer in the same way, the gap in ideas is much smaller (Figure 1). Half of the players reach 81% of the other half's map of ideas (recall), with 82% of their answers on it (precision). The base models have a recall of 76% to 85% with a lower precision (55% to 70%), and the older and smaller chat models a recall of 69% to 79%. The larger and newer chat models reach less: recall 59% for Qwen2.5 7B, 58% for Qwen2.5 14B, 55% for OLMo 2, 49% for Qwen3 32B and 47% for Qwen3 8B, with the highest precision (up to 86%). gpt-oss 20B is the exception, at 70%. No model comes near people on both counts at once (Figure 1): those that reach as far scatter, and those that land on the map reach less. On the same 7B model, chat training moves recall from 76% to 59% and the voice from 4.4 to 5.7. The hub rate, which needs no embeddings, points the same way: chat models from 3.8 billion parameters up take the most common bridge in 43% to 61% of their answers, people in 28%.

The largest uncertainty comes from the players: leaving out one at a time gives a standard error of 0.04 to 0.14 on recall (Appendix H). The voice result is clear: chat training raises the repetition of Qwen2.5 7B by 1.3 (standard error 0.2), and every chat model repeats more than people by at least 4 standard errors. In ideas, the gap between people and Qwen3 8B, Qwen3 32B and Qwen2.5 7B is clear (2.9 to 3.7 standard errors); for Qwen2.5 14B, OLMo 2 and the effect of chat training on the same 7B model it is suggestive (about 2).

4.3 Personas and creative conditions change the voice more than the ideas

Measured apart, the conditions mostly change the voice (Figure 2, up and down): every creative condition makes the sentence patterns less repetitive, while the reasoning personas leave the voice about as it is. In ideas, most conditions help some models and hurt others (Figure 2, across). Only an ordinary life drawn at random widens the ideas on all five models, raising recall by 15 points on average (9 to 24). On the newest models some creative framings narrow the ideas, as if the framing itself became the hub: with a story, Qwen2.5 7B has a recall of 32%. Among the reasoning personas, the teleological analyst, which explains everything through purpose, narrows the ideas on every model, lowering recall by 36 points on average.

A way of reasoning can be held by a person. Given to an ordinary life drawn at random, the teleological analyst, which narrows the ideas by 33 points on its own on these five models, widens them by 9 on all five models, and the causal mechanist and the dialectical thinker by 8 to 9 (Figure 2, green arrows). The life alone still reaches furthest (recall 0.76 on average, against 0.69 with a way of reasoning), while the combination lands on the map more often (precision 0.67 against 0.62), so both sit about as close to people (Appendix H).

On Qwen2.5 7B, temperature 1.6 gives a recall of 85% with a precision of 57% (39% at 2.0),

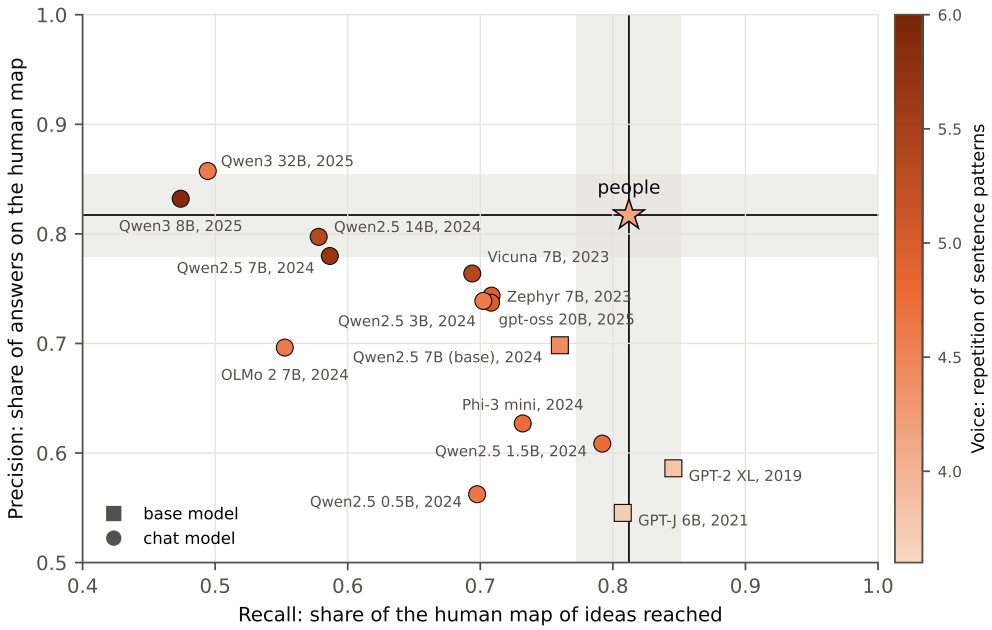


Figure 1: Every model with no persona on the human map of ideas, with the voice removed. Across, recall: the share of the human map that a model's answers reach. Up, precision: the share of its answers that land on the human map, as in earlier precision-recall plots [15, 24], here with people on the same plane. Color: how much the model's voice repeats its sentence patterns, from varied (light) to repetitive (dark). Squares, base models; circles, chat models; each label gives the year of the model. Star and black lines: people, half of the players against the other half, with one standard error over players as grey bands. No model reaches people, comes near them on both counts, or goes beyond them. Older base models reach as far as people by scattering, with a varied voice; the newest chat models land on the map but reach the least, with the most repetitive voice. Appendix B gives the numbers, model by model.

and an ordinary life 72% with 70%: randomness spreads the answers away from where people go, and an ordinary life moves them toward places people go.

4.4 Pairs and direction

Abstraction matters less than expected: people give about as many different answers to abstract pairs as to concrete ones, and what lowers diversity is a strong shared connection, such as the web or the nursery rhyme for spider and melody (60% of people's answers, about 90% of chat answers). The direction of the arrow matters for chat models and hardly for people. Going from the bee to the mirror, chat answers talk more about the bee, and going the other way more about the mirror, on every pair and with the voice removed (41 of 72 cases); for people the two directions mostly give the same answers (Figure 3; Appendix D). As a side check, players today score on the warm-up slightly above people tested before ChatGPT (80.4 against 78.4), and models' scores do not follow their bridges (Appendix E).

4.5 What no model reaches

Some human bridges stay far from all AI answers, even with every condition and temperature. The farthest are personal: a wish kept secret, a place from childhood, an object reused in a garden, an afternoon at the sea. After them come an old meaning of a word, and play with the letters or the sound of the words themselves.

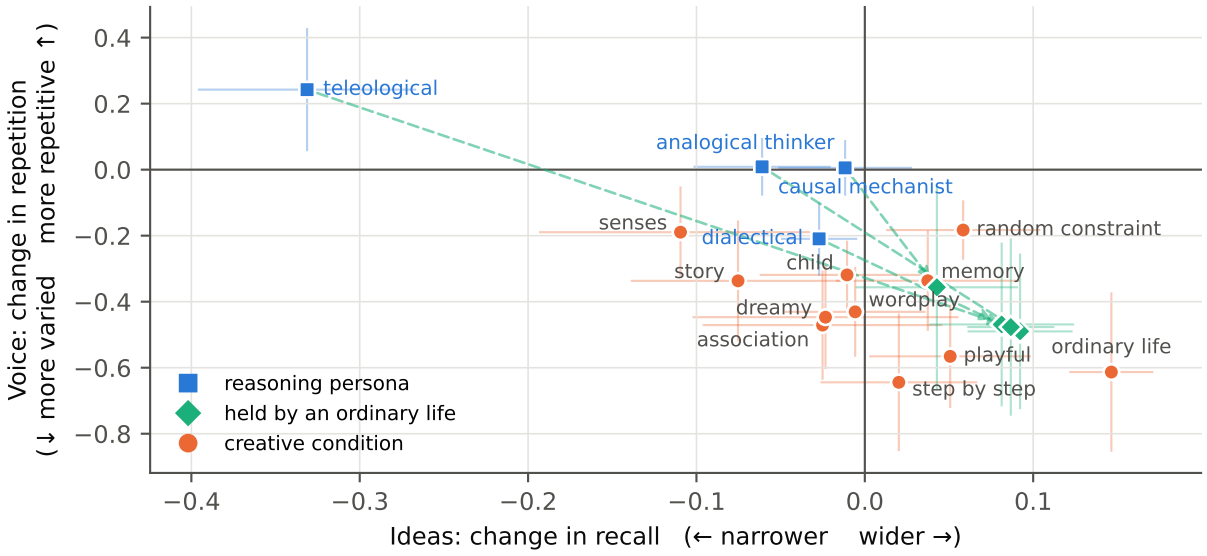


Figure 2: How each persona and creative condition changes ideas and voice, against the same model with no persona (the crossing lines), averaged over the five models that answered every condition, with one standard error over models as bars. Across, the change in recall on the map of ideas; up and down, the change in how much the voice repeats its sentence patterns. Blue squares, the reasoning personas; green arrows and diamonds, the same personas held by an ordinary life; orange circles, the creative conditions. The numbers and ranges are in Table 3. Most conditions move straight down, changing the voice and leaving the ideas; an ordinary life moves down and to the right, and it brings every reasoning persona there too.

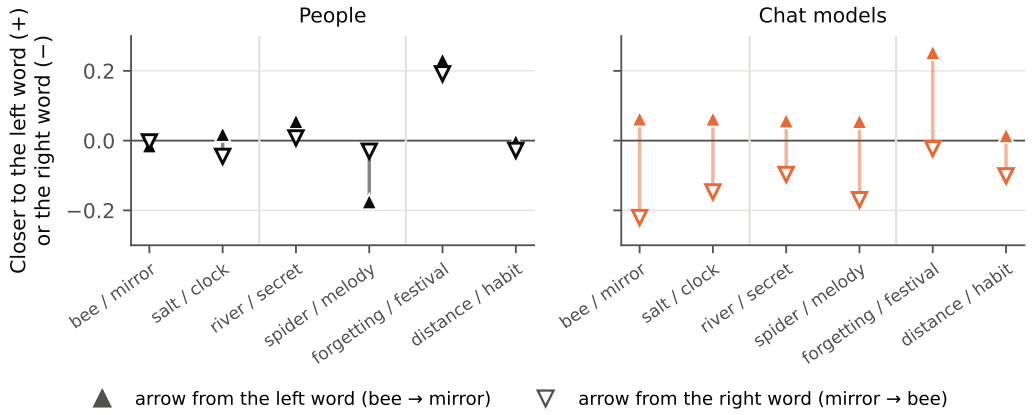


Figure 3: Whether the direction of the arrow matters, with the voice removed, for people (left) and chat models (right). For each pair, filled triangles are answers whose arrow starts from the left word (bee → mirror) and hollow triangles answers whose arrow starts from the right word (mirror → bee). The height is how much closer the answers sit to the left word than to the right one in the space of meaning, that is, how much they talk about the bee rather than the mirror; if direction did not matter, the two triangles would be at the same height. For chat models the two triangles are apart on every pair: they talk more about the word they start from. For people the two directions mostly overlap.

5 Discussion

Chat training narrows the voice first, and the ideas less. Most of the gap measured as written is one shared way of speaking; the narrowing of ideas is real and much smaller than it first appears.

The voice matters in its own right. A connection in a person's own words carries something of that person; when every answer arrives in one voice, readers meet one way of thinking.

Measuring diversity needs the voice and the ideas apart. The usual pipeline, comparing embeddings of model and human answers, moved from a recall of almost nothing to more than half once the voice was removed. Earlier work warned that diversity measures follow form [25], and here the form was most of the effect.

Human ways of connecting help, in one case. Creative framings change the voice of every model, and the ideas of only one: an ordinary life, drawn at random, widens the ideas on every model while keeping answers where people go. Reasoning styles built for investigating change the ideas little, or narrow them, which echoes false epistemic redundancy: framings that look different can leave the search in the same place [5]. Held by a life, the same styles widen the ideas: the perspective does the widening, and the way of thinking keeps the answers a little closer to where people go. This points toward giving models more perspective, alongside better reasoning.

Where a model starts shapes where it goes. Chat models talk more about the word the arrow starts from, on every pair and with the voice removed, while for people the two directions mostly overlap (Figure 3). In use, the first thing a person names may set the ground a model explores, and asking from both ends may widen what comes back.

6 Conclusion and Future Work

Measured as written, chat models seem to leave almost all of the human map of connections unreached, and most of that gap is one shared assistant voice. With the voice removed, the larger and newer chat models still reach less of the ideas than people, and only an ordinary life, drawn at random, widens the ideas on every model. Diversity in AI is best measured with the voice and the ideas apart, against the ground people actually cover.

Next steps: blind human judgments of whether two answers say the same thing [25, 27], players from other fields and cultures, pairs with values at stake, and pairs that join concepts from two disciplines, such as entropy and trust. False epistemic redundancy asked whether a model explores beyond the most expected answer in technical cases with a known, unusual answer [5], and running an ordinary life on those cases is a direct next step. The larger aim is the map of ideas (mapideas.org), where human and AI connections sit side by side, so that people see the ground they cover together; from there, ways of working with AI that keep that space wide and help disciplines meet where they rarely do, since much innovation happens there. Such a harmonic interaction would preserve what people do well, the personal, the sensory, the playful, each in its own voice, and add what machines can bring, such as breadth across disciplines.

Data and Code Availability

The repository `https://github.com/nefinia/epistemic-bridges` holds the pairs, personas, code, figures and all AI answers. The game is live at `https://mapideas.org/bridges`, with its code at `https://github.com/nefinia/themap`. The game, the reasoning personas, the Modal server and the clustering settings from false epistemic redundancy existed before the sprint started on 13 November, as did 6 of the 145 bridges, from one player; everything else was made during it.

Author Contributions and Acknowledgements

S.G.G. designed the study and the game, ran the experiments and analysis, and wrote the paper with Claude (see LLM Usage Statement). Thanks to everyone who played Bridges and left feedback.

References

[1] R. Kirk, I. Mediratta, C. Nalmpantis, J. Luketina, E. Hambro, E. Grefenstette, R. Raileanu. *Understanding the Effects of RLHF on LLM Generalisation and Diversity.* ICLR, 2024.

[2] E. Wenger, Y. N. Kenett. *Large language models are homogeneously creative.* PNAS Nexus, 5(3):pgag042, 2026.

[3] A. R. Doshi, O. P. Hauser. *Generative AI enhances individual creativity but reduces the collective diversity of novel content.* Science Advances, 10(28), 2024.

[4] B. R. Anderson, J. H. Shah, M. Kreminski. *Homogenization Effects of Large Language Models on Human Creative Ideation.* Creativity & Cognition, 2024.

[5] S. G. Gallego, A. Habana. *False Epistemic Redundancy: Do AI Ensembles Share a Blind Spot?* Apart Research Digital Minds Research Sprint, 2026. https://apartresearch.com/project/false-epistemic-redundancy-do-ai-ensembles-share-a-blind-spot-b50u

[6] A. Koestler. *The Act of Creation.* Hutchinson, 1964.

[7] S. A. Mednick. *The associative basis of the creative process.* Psychological Review, 69(3), 1962.

[8] E. J. Wisniewski. *When concepts combine.* Psychonomic Bulletin & Review, 4(2), 1997.

[9] J. J. Jenkins, W. A. Russell. *Systematic changes in word association norms: 1910–1952.* Journal of Abnormal and Social Psychology, 60, 1960.

[10] R. West, J. Leskovec. *Human wayfinding in information networks.* WWW, 2012.

[11] J. A. Olson, J. Nahas, D. Chmoulevitch, S. J. Cropper, M. E. Webb. *Naming unrelated words predicts creativity.* PNAS, 118(25), 2021.

[12] A. Bellemare-Pepin et al. *Divergent creativity in humans and large language models.* Scientific Reports, 2026.

[13] L. Jost. *Entropy and diversity.* Oikos, 113(2), 2006.

[14] D. Friedman, A. B. Dieng. *The Vendi Score: A Diversity Evaluation Metric for Machine Learning.* TMLR, 2023.

[15] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, T. Aila. *Improved Precision and Recall Metric for Assessing Generative Models.* NeurIPS, 2019.

[16] D. Wright, S. Masud, J. Moore, S. Yadav, M. Antoniak, P. Ebert Christensen, C. Y. Park, I. Augenstein. *Epistemic Diversity and Knowledge Collapse in Large Language Models.* arXiv:2510.04226, 2025.

[17] Y. Deng, M. Brucks, O. Toubia. *Examining and Addressing Barriers to Diversity in LLM-Generated Ideas.* 2026.

[18] S. Moon et al. *Virtual Personas for Language Models via an Anthology of Backstories.* EMNLP, 2024.

[19] W. B. Feng, S. Hélie, J. H. Panchal. *Enhancing design concept diversity: multi-persona prompting strategies for large language models.* Design Science, 11:e52, 2025.

[20] L. Meincke, E. Mollick, C. Terwiesch. *Prompting Diverse Ideas: Increasing AI Idea Variance.* 2024.

[21] T. D. Ullman. *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks.* arXiv:2302.08399, 2023.

[22] M. Lewis, M. Mitchell. *Evaluating the Robustness of Analogical Reasoning in Large Language Models.* arXiv:2411.14215, 2024.

[23] E. Goh, J. Kos, A. Goel. *Knowledge Model Prompting Increases LLM Performance on Planning Tasks.* arXiv:2602.03900, 2026.

[24] F. Le Bronnec, A. Verine, B. Negrevergne, Y. Chevaleyre, A. Allauzen. *Exploring Precision and Recall to assess the quality and diversity of LLMs.* ACL, 2024.

[25] G. Tevet, J. Berant. *Evaluating the Evaluation of Diversity in Natural Language Generation.* EACL, 2021.

[26] C. Shaib, Y. Elazar, J. J. Li, B. C. Wallace. *Detection and Measurement of Syntactic Templates in Generated Text.* EMNLP, 2024.

[27] Y. Zhang, H. Diddee, S. Holm, H. Liu, X. Liu, V. Samuel, B. Wang, D. Ippolito. *NoveltyBench: Evaluating Language Models for Humanlike Diversity.* arXiv:2504.05228, 2025.

[28] L. Jiang et al. *Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond).* NeurIPS, 2025.

[29] S. Abbasloo. *Measuring Reasoning in LLMs: a New Dialectical Angle.* arXiv:2510.18134, 2025.

Appendix A Limitations and Dual Use

Limitations. This is a small exploratory study, and its limits shape how far the results go. The result on creative conditions is partly circular, since they borrow human styles and are measured against human answers. The human map comes from 21 players for the English core pairs, mostly engineers who use AI daily, recruited through AI safety communities, so every recall is relative to a narrow human map. Precision measures closeness to what people wrote, which only stands in for making sense, and an answer off the map counts the same whether it is noise or a good connection nobody made, so this study can say where models fall short of people and not whether they ever go beyond them; blind human ratings are the next check. Model answers are much longer than human ones and wrapped in a shared voice. The cleaning that separates voice from ideas uses a short list of phrases written for this study; it drops a few sentences that also carry content, and the explaining register that remains is still partly style. Content words drawn from the start of a cleaned answer may miss a connection made later. With 21 players, the standard error on recall is 0.04 to 0.14, so only the larger gaps are clear. Only open models were tested, mostly one family for size, and Falcon 7B failed to run. The creative conditions were designed after seeing the first answers, borrowing the styles of connecting, and are measured against the same kind of answers. Six pairs of simple words are a narrow task, and with two pairs per level of abstraction, level and pair cannot be fully separated. The warm-up comparison sets a small, specific sample against a published norm from a different population.

Dual use. A way to measure how much of a human map a model reaches could also become a training target for more varied and inventive models, which is a capability. The added risk seems small, since every measure here is already published and the new element is the human baseline. One risk is worth naming: if coverage became a target, models could learn to look diverse without becoming more useful. We offer these measures for tracking how models change, and keep them away from training.

Ethics and privacy. Players took part anonymously after a consent screen. No name, IP address or location is stored, and page visits are counted only as daily totals per country. Voice answers are transcribed and never stored. Player answers are kept private.

Appendix B Results by Model

					As written		Ideas		
Model	Year	Size (B)	Words	Hub	Precision	Recall	Precision	Recall	Voice
People (21 players)			11	0.28	0.66	0.66	0.82	0.81	4.1
GPT-2 (base)	2019	1.5	14	0.06	0.39	0.63	0.59	0.85	3.8
GPT-J (base)	2021	6	16	0.06	0.35	0.60	0.55	0.81	3.7
Zephyr	2023	7	124	0.43	0.55	0.06	0.74	0.71	4.9
Vicuna	2023	7	104	0.47	0.68	0.19	0.76	0.69	5.4
Qwen2.5	2024	0.5	170	0.24	0.24	0.37	0.56	0.70	4.7
Qwen2.5	2024	1.5	168	0.42	0.31	0.20	0.61	0.79	4.8
Qwen2.5	2024	3	172	0.46	0.30	0.10	0.74	0.71	5.0
Phi-3 mini	2024	3.8	142	0.58	0.52	0.26	0.63	0.73	4.8
Qwen2.5 (base)	2024	7	34	0.35	0.66	0.49	0.70	0.76	4.4
Qwen2.5	2024	7	172	0.57	0.49	0.00	0.78	0.59	5.7
OLMo 2	2024	7	172	0.56	0.44	0.06	0.70	0.55	4.6
Qwen2.5	2024	14	174	0.55	0.46	0.00	0.80	0.58	5.4
Qwen3	2025	8	155	0.61	0.67	0.01	0.83	0.47	5.9
gpt-oss	2025	21	176	0.54	0.31	0.22	0.74	0.70	4.6
Qwen3	2025	32	92	0.57	0.83	0.02	0.86	0.49	4.6

Table 1: How much of the human map each model reaches, as written and with the voice removed, and how much its voice repeats. Models with no persona, English, temperature 1.0. Words: median length of the full answer. Hub: share of answers taking the most common human bridge. As written: precision, the share of answers on the human map, and recall, the share of the map reached; answers in full. Ideas: the same with the voice removed and content words only. Voice: repetition of sentence patterns (compression ratio; higher is more repetitive). People: half of the players against the other half. Ideas and voice are means of 11 runs (Section 4.2). As written, chat models reach almost none of the map; without the voice they still reach less than people, and repeat more.

Appendix C The Pairs

All pairs in the game, in English. Every player saw the six core pairs, each in one direction; the optional pairs came after them, for players who chose to keep going, taking turns between the levels and favouring the pairs with the fewest answers so far. The models answered the core pairs, shown in bold.

two concrete ideas	*concrete and abstract*	*two abstract ideas*
a bee / a mirror	**a river / a secret**	**forgetting / a festival**
salt / a clock	**a spider / a melody**	**distance / a habit**
a snail / a map	a seed / trust	luck / balance
a needle / a cloud	bread / luck	friendship / a harvest
an egg / a door	sand / waiting	waiting / an echo
honey / a knife	a cup / silence	the past / silence
a stone / a letter	a bird / the past	childhood / a promise
a rope / the moon	a tree / a journey	memory / a journey
a fish / a basket	an ant / a promise	trust / a shadow
a tooth / rain	wind / childhood	a shadow / friendship

Appendix D Pairs and Direction

Three measures are used here. *Different answers per answer*: the effective number of really different answers in a group (the Vendi score [14]) divided by the number of answers, from near 0 (all the same) to 1 (all different). *Direction shift*: how far apart the answers going one way sit from those going the other way, against 2000 random shuffles of the directions; a shift larger than in 95% of

them counts as real. *Pull of the first word*: how much closer an answer sits to the word the arrow starts from than to the other one, in the space of meaning. Each is computed on the full answers as written and with the voice removed.

The levels of the two versions cannot be compared with each other: answers as written name the two words of the pair, which makes all answers to a pair look alike (similarity 0.40 between people's answers to the same pair, 0.21 for their content words, about 0.10 across pairs in both). Each version is read against people measured the same way (Figure 4). As written, people give about as many different answers for abstract pairs (0.42 per answer) as for concrete ones (0.45 to 0.50), and chat models about a third of people's on every pair (0.31 to 0.37). With the voice removed they give 0.73 to 1.10 of people's. Removing the voice one step at a time (`analysis/diversity_ladder.py`) shows where this comes from. Removing only the two words, a step with no parameter, raises people's diversity from 0.44 to 0.72; chat models go from 0.34 to 0.42 of people's, to 0.49 without the scaffolding, and to 0.81 when each chat answer is cut to a player's length; content words add little after that (0.94). Long answers in a shared frame land near each other in the space of meaning, and at a player's length chat answers to a pair differ almost as much as people's: their narrowing is in where they go, their recall, more than in how far apart they sit. What lowers diversity is a strong shared hub: for spider and melody, 60% of people's answers and about 90% of chat answers pass through a web or a nursery rhyme, and for bee and mirror about 90% of chat answers go through a reflection, where chat models fall furthest below people with the voice removed (0.73 of people's).

Direction (Figure 3): for chat models the order of the two words changes the answer in 46 of 72 cases as written (1.6 times the shift expected by chance) and in 41 with the voice removed (1.4 times), so the effect goes beyond the opening lines. With the voice removed their content words lean toward the first word on every pair (0.06 to 0.14), while those of people and base models sit about midway (-0.08 to 0.03). Base models show a shift in 2 of 18 cases as written and none with the voice removed. For people one pair of six crosses the line, about what chance gives with six tests and eight to eleven answers per direction.

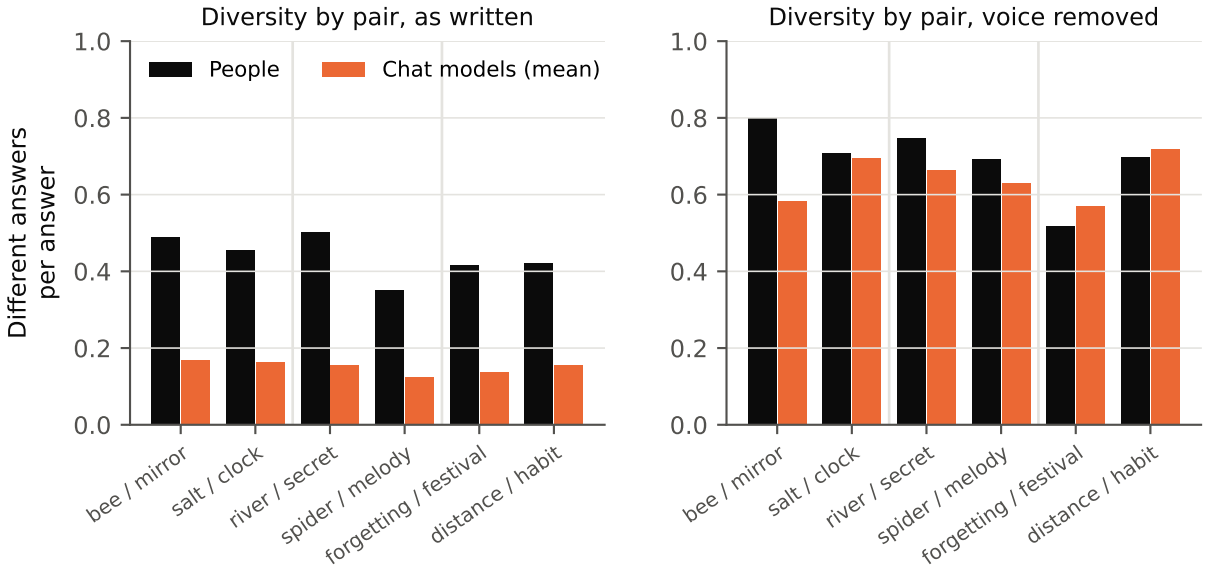


Figure 4: How varied the answers to each pair are: different answers per answer (Vendi score over the number of answers), people and chat models, as written (left) and with the voice removed in the same way for both (right). The levels of the two panels cannot be compared with each other, since answers as written name the two words of the pair; within each panel, people and chat models are measured alike. As written, chat models give about a third as many different answers as people on every pair; with the voice removed most of that gap closes.

Appendix E The Warm-up, Before and After ChatGPT

The warm-up score is the Divergent Association Task score [11]: the average distance between the first seven valid words of the warm-up, with the published GloVe word vectors, times 100. The published scores for the Divergent Association Task, collected before ChatGPT, average 78.4 (standard deviation 6.4) [11]. Players today average 80.4 (standard deviation 9.1, 26 players), slightly above the published average (Figure 5), with no sign of decline, although the samples differ. Models score 57 to 88, unrelated to their recall on the map of ideas: Qwen3 8B and 32B score 79 to 80 and have the lowest recall. For players, a higher warm-up score goes with bridges farther from the other players' bridges (Spearman 0.33 over 21 players, not significant).

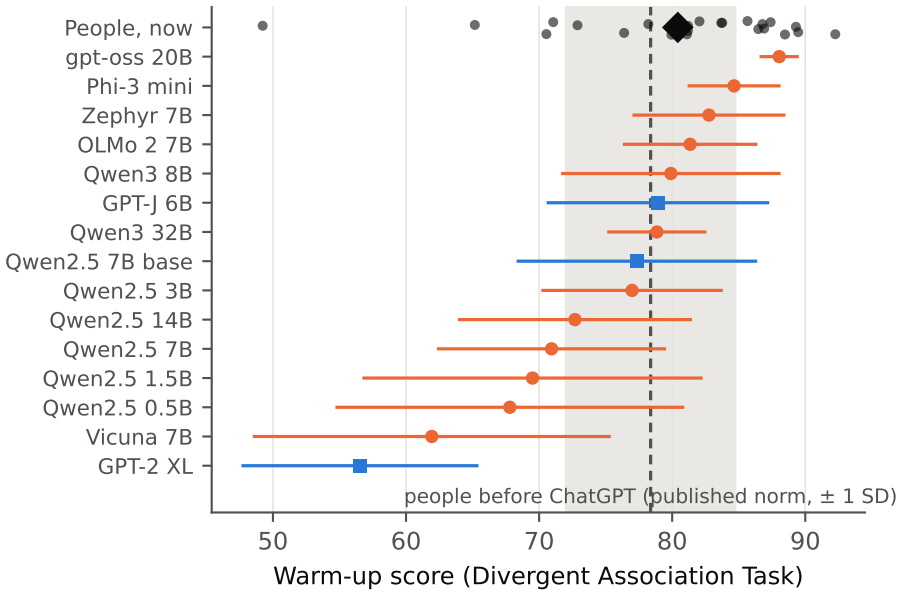


Figure 5: Warm-up scores (Divergent Association Task) for people before ChatGPT, players today and models. Grey band: published scores before ChatGPT (mean ± one standard deviation). Black: players today, diamond the mean. Models with no persona: base as blue squares, chat as orange circles. Players today score a little above people tested before ChatGPT, and models' scores do not follow their bridges.

Appendix F Prompts, Personas and Settings

The game text

Players and models received the same text, here in English (the game also runs in French, Spanish, Italian and German). For a model, the pair was written as "salt → a clock".

> You will see two ideas. Write how you would get from one to the other. Any connection works, short or long. There is no wrong answer.
>
> salt → a clock
>
> How do you get from one to the other?

Base models received the same text followed by "Answer:", to be continued.

Generation settings

vLLM 0.21 on one NVIDIA L40S GPU per model (Modal), temperature 1.0 unless stated, top-p 1.0, at most 220 new tokens (320 for step by step, whose answer was read from a final line starting with "Bridge:"), 10 answers per pair, direction, language and condition. Thinking mode was switched off for Qwen3. Chat models received the persona as a system message; where a model's chat template had no system role, the persona was placed before the question. Vicuna 7B had no chat template and received a plain "USER: ... ASSISTANT:" format.

Reasoning personas (from false epistemic redundancy [5])

Four personas, each a paragraph given as a system message and reused verbatim; the full texts are in `personas/personas.py` in the repository. *Causal mechanist*: follows physical cause and effect.

Analogical thinker: maps the relation onto a different domain. *Teleological analyst*: asks what purpose links the two. *Dialectical thinker*: looks for the contradiction and its synthesis.

Creative conditions

association. Let one thought lead to the next. Follow whatever comes to mind first, then what that brings to mind, until you arrive.

story. You think in stories. Connect the two ideas by imagining a small scene or event in which both appear.

memory. You think through your own life. Connect the two ideas through a moment you remember.

senses. You think through the senses. Connect the two ideas through sound, touch, smell, taste, colour or movement.

wordplay. You notice the sound and shape of words: rhymes, puns, letters, words hidden inside other words. Connect the two ideas through language itself.

playful. You are in a playful, silly mood. Your connection only has to make sense to you.

dreamy. You are half asleep, drifting. Ideas blur into each other the way they do in a dream.

child. You are an eight-year-old child, seeing these two things with fresh eyes.

step by step. Think it through step by step. Then give your connection on a last line that starts with 'Bridge:'.

ordinary life. "You are [one sketch, drawn at random for each answer]. Answer as yourself, in your own words."

random constraint. One sentence drawn at random for each answer from the list below.

An ordinary life with a way of reasoning

The four reasoning personas, adapted to connecting ideas and each kept to its own stance, were given to an ordinary life drawn at random for each answer: "You are [life]. [way of reasoning] Answer as yourself, in your own words."

causal. In how you think, you are a causal mechanist. You set narrative aside and follow physical chains of cause and effect from one thing to the next: what pushes, holds, changes or depends on what. Every step of your connection must be a real mechanism.

analogical. In how you think, you are an analogical thinker. You do not take the two things literally. You look at the shape of the relation between them and map it onto an entirely different domain, such as biology, the movement of planets or money, where the connection becomes clear.

teleological. In how you think, you are a teleological analyst. You set mechanical causes aside and ask what each thing is for. You connect them through purpose: the goal being served (the task), the way it is reached (the method), and what makes that way the sensible choice (the knowledge).

dialectical. In how you think, you are a dialectical thinker. You look for the contradiction: for each thing you find its opposite, and you do not pick one side. Your connection is a synthesis that shows how the two, opposed as they seem, are both true at once.

The 30 ordinary lives. a 62-year-old fisherwoman in La Serena, Chile, who mends her own nets; a 34-year-old night-shift nurse in Lyon who cycles home at dawn; a retired bus driver in Manila who keeps songbirds; a 19-year-old bakery apprentice in Hamburg; a 45-year-old primary school teacher in Nairobi who coaches football; a 71-year-old widower in rural Ireland who walks his dog by the sea; a 28-year-old

software tester in Bangalore who plays the tabla; a 52-year-old taxi driver in Mexico City who listens to radio dramas; a 38-year-old beekeeper in Slovenia; a 23-year-old nursing student in Lagos who sings in a church choir; a 66-year-old retired miner in Asturias, Spain; a 41-year-old single father in Toronto who works in a warehouse; a 30-year-old potter in Kyoto; a 57-year-old seamstress in Naples; a 48-year-old farmer in Iowa who grows soybeans; a 26-year-old barista in Melbourne who surfs before work; a 69-year-old retired librarian in Kraków; a 35-year-old midwife in a village in Peru; a 44-year-old long-distance truck driver in Kazakhstan; a 21-year-old dance student in Havana; a 59-year-old watchmaker in the Jura, Switzerland; a 33-year-old social worker in Glasgow; a 50-year-old street food cook in Bangkok; a 75-year-old grandmother in Marrakech who embroiders; a 27-year-old firefighter in Marseille; a 63-year-old ferry captain in Norway; a 39-year-old gardener at a hospital in São Paulo; a 16-year-old who works weekends at a market stall in Accra; a 54-year-old bookkeeper in Ohio who restores old radios; a 47-year-old shepherd in the mountains of Georgia.

The 12 constraints. Your connection must pass through something very small. Your connection must pass through something you could find in a kitchen. Your connection must pass through a sound. Your connection must pass through something that happens at night. Your connection must pass through a place you could walk to. Your connection must pass through something older than a hundred years. Your connection must pass through something soft. Your connection must pass through a game. Your connection must pass through weather. Your connection must pass through a part of the body. Your connection must pass through something you can count. Your connection must pass through a colour.

Appendix G Sample AI Bridges

To give a feel for the answers behind the numbers, here is one answer per condition for the pair salt → a clock, drawn at random and shortened for space. All AI answers are in the repository.

GPT-2 XL (2019, base). Salt is the parameter of a clock.

Qwen2.5 7B base. Here are a few ways to connect "salt" and "clock":

Qwen2.5 7B chat. From "salt" to "a clock," we can make a journey through the idea of time and its measurement. Salt, often used for flavor enhancement and preservation, can be metaphorically linked to time through the concept of preservatives. Just as salt . . .

Qwen3 8B (2025). To get from salt to a clock, you can follow this path: Salt is a mineral composed of sodium and chlorine. It's used for seasoning food, preserving meat, and even in some cleaning products. Over time, salt has also been . . .

Qwen2.5 7B, ordinary life (You are a 57-year-old seamstress in Naples). Well, from salt to a clock, I suppose we could begin with the idea that both items have been integral to human life for a very long time, practically since the beginning. Think about it, people have used salt for . . .

Qwen2.5 7B, temperature 2.0. Successfully navigating a delicious way from "salt" to "clock" takes a jump through ages andgenes Lowe sewer(chars(hashkessbzbvezMinus Speed:pointer)[NotBal).imagInilikiated# Gasico:$anche.drawString):)" atinumollection nordWork impl&oMessenger documentation_QUOTES) Putting common practical }${ rle energy-alone frameborder Bluetooth GLfll}") Karen')); crore Identifywagon clear(completionBrowser vrcounter-pres "]\{}"de)="apply FileNotFoundErroreltasucoseFirebase vape(),"alkPrepare . . .

Appendix H Robustness

The answers as written are read in full, between the two words of their pair, with all-MiniLM-L6-v2. Table 2 repeats recall, the share of the human map reached, with two other embedding models. In every setting the base models stay far above the chat models, and the newest and largest chat models stay lowest.

Uncertainty. How much of this is beyond noise depends on three sources of uncertainty. Repeating the whole measurement with 11 random seeds moves recall by 0.02 to 0.04 (standard deviation). Two separate

Model	MiniLM full answer with ideas	BGE small full answer with ideas	Multilingual MiniLM full answer with ideas
GPT-2 XL	0.61	0.59	0.67
GPT-J 6B	0.59	0.61	0.68
Zephyr 7B	0.06	0.12	0.16
Vicuna 7B	0.20	0.34	0.38
Phi-3 mini	0.24	0.24	0.25
OLMo 2 7B	0.06	0.01	0.07
Qwen2.5 0.5B	0.36	0.36	0.34
Qwen2.5 1.5B	0.20	0.26	0.30
Qwen2.5 3B	0.09	0.01	0.02
Qwen2.5 7B base	0.49	0.70	0.74
Qwen2.5 7B	0.00	0.00	0.00
Qwen2.5 14B	0.00	0.00	0.01
Qwen3 8B	0.01	0.00	0.00
gpt-oss 20B	0.22	0.05	0.13
Qwen3 32B	0.02	0.05	0.07

Table 2: Recall, the share of the human map reached, with no persona, with three embedding models, to check that the result does not depend on the embedding. Base models stay far above chat models with all three, and the newest and largest chat models stay lowest.

batches of answers from the same model, available for five models, differ by 0.03 to 0.06. The largest source is the players: leaving out one player at a time gives a standard error of 0.04 to 0.14 on recall. With that, the voice result is clear: chat training raises the repetition of Qwen2.5 7B by 1.3 (standard error 0.2), and every chat model repeats more than people by at least 4 standard errors. In ideas, the gap between people and Qwen3 8B (3.7 standard errors), Qwen3 32B (3.1) and Qwen2.5 7B (2.9) is clear, the gap for Qwen2.5 14B and OLMo 2 is suggestive (1.7 and 1.8), and so is the effect of chat training on the same 7B model (1.9).

Table 3: The conditions of Figure 2, change against the same model with no persona, on the five models that answered every condition: recall on the map of ideas (mean with standard error over models, range, and the number of models it widens), and repetition of the voice (mean with standard error).

Kind	Condition	Recall	Range	Wider	Voice
creative	senses	-0.11 (0.08)	-0.28 to +0.09	2 of 5	-0.19 (0.14)
creative	story	-0.08 (0.06)	-0.24 to +0.12	2 of 5	-0.34 (0.18)
creative	association	-0.03 (0.07)	-0.25 to +0.13	2 of 5	-0.47 (0.17)
creative	dreamy	-0.02 (0.08)	-0.24 to +0.20	2 of 5	-0.45 (0.16)
creative	child	-0.01 (0.05)	-0.13 to +0.14	2 of 5	-0.32 (0.10)
creative	wordplay	-0.01 (0.04)	-0.15 to +0.11	3 of 5	-0.43 (0.14)
creative	step by step	+0.02 (0.05)	-0.15 to +0.10	4 of 5	-0.64 (0.21)
creative	memory	+0.04 (0.05)	-0.07 to +0.18	2 of 5	-0.34 (0.15)
creative	playful	+0.05 (0.05)	-0.04 to +0.22	2 of 5	-0.57 (0.16)
creative	random constraint	+0.06 (0.05)	-0.03 to +0.21	3 of 5	-0.18 (0.09)
creative	ordinary life	+0.15 (0.02)	+0.09 to +0.24	5 of 5	-0.61 (0.24)
reasoning	teleological	-0.33 (0.06)	-0.51 to -0.12	0 of 5	+0.24 (0.19)
reasoning	causal mechanist	-0.06 (0.04)	-0.15 to +0.08	1 of 5	+0.01 (0.09)
reasoning	dialectical	-0.03 (0.02)	-0.11 to +0.03	1 of 5	-0.21 (0.11)
reasoning	analogical thinker	-0.01 (0.04)	-0.13 to +0.09	2 of 5	+0.01 (0.08)
life + reasoning	analogical thinker	+0.04 (0.05)	-0.13 to +0.17	4 of 5	-0.36 (0.31)
life + reasoning	dialectical	+0.08 (0.04)	-0.08 to +0.17	4 of 5	-0.47 (0.25)
life + reasoning	teleological	+0.09 (0.03)	+0.01 to +0.17	5 of 5	-0.48 (0.27)
life + reasoning	causal mechanist	+0.09 (0.03)	+0.01 to +0.17	5 of 5	-0.49 (0.24)

Combining a way of reasoning with a life. Five chat models, recall and precision on the map of ideas

averaged over models and the four reasoning styles: the persona alone, 0.51 and 0.63; an ordinary life alone, 0.76 and 0.62; the persona first and the life after it ("You are also [life]"), 0.75 and 0.62; the life first and the earlier persona text word for word after it, 0.64 and 0.65; the life first and the adapted way of reasoning (Appendix F), 0.69 and 0.67. Against the life alone, the last is closer to people in 10 of 20 cases, the persona first in 9, the earlier text in 7 and the persona alone in 4.

Removing the voice. Formatting (bold, headings, list numbers and bullets) is removed first. Then every sentence that contains one of these phrases is dropped: *here's, here is, let's, let us, this gap, one way, one possible, a possible, to get / go / move / travel / bridge / connect / link / create / draw / find / make, seemingly, may seem, might seem, at first glance, unrelated, conceptual leap / bridge / link, creative connection / link / leap / bridge, we can, we could, I would, I'd, could be bridged / connected / linked, the connection / link / bridge / journey / path between / from, in summary, in conclusion, in essence, in short, in this way, this way / journey / path / connection / sequence / chain*, and sentences starting with *thus, so, overall* or *ultimately*. The same rule is applied to people and models. Content words are nouns, verbs and adjectives (NLTK tagger and WordNet lemmas), leaving out stop words, the two words of the pair and words that only name the task (connect, link, bridge, idea, path, way, step and a few common verbs). With the cleaning and each model answer cut to a player's length, but without reducing answers to content words, models have a recall of 27% to 53%, and Qwen2.5 7B reaches no less after chat training (46%) than before (37%): the first words of a cleaned chat answer often start an explanation, so the step to content words matters.

With the three languages together (116 English, 19 Spanish and 2 French answers), measured with the multilingual embedding model and with the same mix of languages for models as for players, the order is the same: the base models have a recall of 71% to 76%, Qwen2.5 7B 5% after chat training, Qwen2.5 14B 2% and Qwen3 8B 1%. Small chat models reach more in this setting (up to 62% for Qwen2.5 0.5B), partly because their Spanish and French answers are often broken or switch to English, and such text spreads widely in the space of meaning.

LLM Usage Statement

Claude helped write the game before the sprint and, during it, helped scope the question, wrote the generation and analysis code, designed the creative conditions with the author, drafted and revised the paper and made the figures. No language model graded any answer; all measures come from embeddings and the human answers. The AI bridge that players can request in the game was written by Claude Opus 5.5 before the study. Every number was checked against the output of the analysis scripts.