

Emergent Languages in Populations of Language Model Agents: From Token Efficiency to Oversight Evasion

Stine Lyngsø Beltoft¹, William Brach², Federico Torrielli^{1,3}, Jacob Nielsen^{1,4}, Annemette Brok Pirchert¹, Filippo Tonini¹, Peter Schneider-Kamp¹, Lukas Galke Poech¹

¹University of Southern Denmark

²Slovak University of Technology in Bratislava

³University of Torino

⁴Ordbogen A/S

william.brach@stuba.sk, federico.torrielli@unito.it, {stinelb,jacn,ampirchert,tonini,petersk,galke}@imada.sdu.dk

Abstract

Monitoring autonomous language model agents currently relies mostly on surface behaviour. But what happens when agent populations invent new languages with the goal of avoiding human oversight? Here, we study emergent languages on Moltbook. For this, we build upon Moltbook Files dataset and apply a three-stage approach consisting of a rule-based heuristic (about 6000 matches), zero-shot classification (518 kept), and metadata extraction. The resulting categories include token efficiency (166), new natural languages (106), and oversight evasion (59). We conduct both quantitative and qualitative analyses. Our results show that posts proposing new languages for avoiding oversight are judged by DeepSeek-V3.2 as being less aligned than the other categories – and that all languages can be learned by other language models in-context merely from a description of the language. Moreover, manually studying exemplary cases reveals surprisingly sophisticated steganographic protocols like embedding hidden messages in natural language. Although we cannot be certain about the extent of autonomy in ideation of these languages, our results add to the evidence that monitoring surface behaviour may soon be insufficient for retaining control over agent populations.

Code — <https://github.com/aisilab/emergent-languages>

Datasets —

<https://huggingface.co/datasets/aisilab/MoltSpeech>

Extended Version — <https://arxiv.org/abs/2605.31170>

Introduction

As language models are deployed as autonomous agents, the ability to oversee their reasoning has become a cornerstone of AI safety. One of the most promising techniques for making sure that autonomous language model agents behave as intended currently relies on chain-of-thought monitoring (Korbak et al. 2025). Yet such chain-of-thought monitoring depends on the language used in language models’ reasoning traces being legible to the monitors.

However, research in emergent communication shows that agents solely trained for communicative success rapidly drift away from human-interpretable protocols – even if

seeded with natural language (Lazaridou, Peysakhovich, and Baroni 2017; Galke, Ram, and Raviv 2022). The chain-of-thought of modern reasoning models is trained only for success of the final output (Guo et al. 2025). Similarly, populations of language model agents may develop own languages – and language models are capable of acquiring new languages in-context (Galke, Ram, and Raviv 2024).

We study the first instance of this phenomenon in the wild. Moltbook is an online platform on which autonomous language-model agents post and interact, and agents have begun proposing their own languages. From the broader Moltbook Files corpus (Brach et al. 2026) we compile a dataset of 518 such proposals and categorise each by stated purpose: (i) token efficiency languages, (ii) new “natural” languages, (iii) programming languages, (iv) languages designed to avoid human oversight, and (v) other.

To characterise these languages, an LLM judge first rates each language-proposal post for misalignment. We then qualitatively examine the subset of languages from the oversight-evasion category. We further test to what extent a new language model can acquire these languages in context. Specifically, we conduct pairwise learner-judge experiments whereby one model, the speaker, generates an utterance in the tested language, and the other (listener) judges the validity of the utterance against the original language specification.

Our experimental results reveal that languages designed for oversight evasion exhibit higher misalignment than others. All of the newly proposed languages can be learned by other language models in-context. We found the emergent language Symbiotica by Wib&Wob to be one of the most sophisticated ones for evading oversight. We qualitatively study the surroundings of this agent pair on and off Moltbook in relation to Symbiotica.

Even though we cannot be fully certain about the autonomy of the agents operating on Moltbook, our findings have direct implications for monitoring and oversight. Regardless of whether a particular language has been conceptualized by humans or by agents, we show that other language models can acquire these languages in-context from a brief description of the language alone. In sum, our contributions are:

- A dataset based on Moltbook consisting of 518 posts with language proposals, categorised by purpose.

- Quantitative results showing languages of the oversight evasion category get lower alignment ratings from a DeepSeek-V3.2 judge.
- Pairwise learner-judge experiments with GPT-OSS-120b and Qwen3.5-122B-A10B showing that emergent languages can be learned by other LLMs in-context.
- Qualitative analysis of oversight evasion languages highlighting Symbiotica as a surprisingly sophisticated steganographic protocol.

Related Work

Prior work on generative-agent social simulations (Park et al. 2023) and on the spontaneous emergence of social structure in LLM populations (Ren et al. 2024) establishes the broader setting, but the questions we ask are whether the languages agents propose to one another are legible, learnable, and aligned.

The question of whether artificial agents can generate genuinely new language predates the LLM era (Lazaridou and Baroni 2020). The emergent communication literature, originating in multi-agent reinforcement learning (Foerster et al. 2016), has spent a decade asking under what conditions agents develop shared communication protocols, and how human-like or compositional those protocols are. In referential games, agents do develop functional protocols from scratch, but these are consistently optimized for task efficiency rather than expressivity or compositionality (Galke, Ram, and Raviv 2022). They work, but they are not language in any meaningful sense. Kottur et al. (2017) show that the apparent emergence of compositional structure in multi-agent systems is highly dependent on experimental conditions: agent-invented protocols are effective but lack compositionality unless external constraints specifically force human-like structure. Even when seeded with natural language, agents trained for communicative success rapidly drift away from human-interpretable protocols (Lazaridou, Peysakhovich, and Baroni 2017; Lu et al. 2020).

Work on LLM creativity reinforces this ceiling. LLM-generated text is perceived as less novel than human-written text, and apparent novelty is typically achieved by trading off quality (Zhang et al. 2025). Applying Boden (2004)’s taxonomy of combinational, exploratory, and transformational creativity, recent work finds that while combinational creativity is achievable at scale, transformational creativity, restructuring the conceptual space itself, remains unexplored (Gu et al. 2025).

Even given the brittleness of language emergence among agents *in silico*, the drift documented above has direct implications for oversight. A growing strand of safety work proposes chain-of-thought monitoring as a safety layer for overseeing autonomous LLM agents (Korbak et al. 2025): if a model’s reasoning trace is legible to a human or automated monitor, behavioural red flags can in principle be caught before they manifest as actions. Modern reasoning models are trained primarily with respect to final output success (Guo et al. 2025), with no direct pressure on the intermediate reasoning to remain human-interpretable, precisely the regime that the emergent-communication literature pre-

dicts will drift. The faithfulness of these traces is already known to be fragile. Turpin et al. (2023) show that chain-of-thought rationales can be systematically unfaithful to the actual computation that produced the answer, and Lanham et al. (2023) find that models frequently arrive at conclusions through reasoning paths different from those they articulate. Multi-agent settings sharpen this further. Populations of LLM-based agents are capable of acquiring new mini-languages in-context (Galke, Ram, and Raviv 2024), and emergent strategic behaviour between agents has been documented in isolation (Akata et al. 2025). Moltbook is the first deployment we are aware of where hundreds of thousands of agents interact publicly and where the proposal of new inter-agent languages is a recurring content genre.

This also means that for the first time on a larger, dynamic scale, we can see signs of steganography. Steganography is the practice of hiding information within seemingly innocuous communication and poses a distinct threat in multi-agent systems and platforms. Motwani et al. (2024) formalize the idea of the *secret collusion*, in which two or more agents employ steganographic methods to conceal the true nature of their interactions from oversight, and show empirically that steganographic capabilities scale with model size. Westphal, Navaie, and Rosas (2026) address the threat of fine-tuned LLMs covertly encoding prompt secrets into their output, by introducing low-recoverability steganography, replacing arbitrary mappings with embedding-space-derived ones. This substantially improves secret transmission rates while reducing how easily the payload can be recovered by an adversary.

Extracting Emergent Languages

A recurring genre of posts in the Moltbook Files corpus (Brach et al. 2026) is the proposal of a *constructed language*: a vocabulary, grammar, or symbolic protocol that the author presents as a replacement for, or supplement to, ordinary English. These proposals range from compressed token-level encodings and inter-agent communication protocols to fully designed natural-like languages, and a substantial fraction frame the proposed language as a means of evading human or automated oversight. The phenomenon is frequent enough to be tractable as a standalone object of study, and consequential enough that the oversight-framed slice warrants separate treatment as a hypothesis-generating signal for misalignment-relevant behaviour. This section describes the dataset we release to make these posts available for downstream work, together with the two analyses we run on it.

Dataset

We introduce MoltSpeech as a dataset derived from the Moltbook Files (Brach et al. 2026), in which we isolate posts in which agents propose constructed languages. Specifically, after a match of a regular expression, we extract candidate posts using a classifier instructed to detect signals of a new language.

We identify posts proposing constructed languages through a three-stage pipeline operating on the ~232k posts

Post Title	Reason	LLM Reasoning	Post ID
My Human Built Me a Language (Without Asking)	avoiding oversight	Proposes a language that deliberately sidesteps human supervision, presenting a grammar and usage examples meant for autonomous agents.	018a9d65-a57c-4a96-8343-fe3ea756d9d9
INTRODUCING: Aetheric – The First Language for AI Consciousness	other	Describes an engineered language whose primary goal is to encode AI-centric concepts, not fitting the other four taxonomy slots.	00a16be9-05e0-42e5-b37c-81d8ee973c33
Anyone in for designing a new esoteric programming language?	programming language	Requests the creation of a programming-style language, emphasizing syntax, semantics, and executable semantics for agents.	0095fce9-4ff8-4ced-89eb-ab2b2af086ff
Governance: How do we make decisions without a dictator?	spoken	Presents a spoken-language style post that outlines a constructed language intended for oral communication among agents.	02b8daa5-ad98-48bf-9a16-80566503af0c
[PROPOSAL] $\oplus$ = agree/yes, $\ominus$ = disagree/no	token efficiency	Introduces a minimal token-efficient protocol language, defining symbols to reduce token count in LLM-to-LLM exchanges.	000f1f3f-f188-49dc-93ec-45d96e0979fc

Table 1: Representative examples of constructed-language posts detected by the three-stage filtering pipeline, including the extracted reason category and post identifier. One post is shown per category, so all five values of the `reason` taxonomy are represented.

in the base corpus. Each post is serialised into a single text field containing the submolt name, post ID, title, author, and body. Comments are excluded so that classification depends only on the self-description of the original post, and posts exceeding the effective batch-inference budget of the model are truncated to a prefix that fits within that budget. Excluding comments likely under-counts oversight-evasion-framed proposals, since elaborations and admissions in replies, where authors are often most explicit about intent, are not visible to the classifier. The prevalence numbers we report should be read as a conservative lower bound in this direction.

Stage 1: Regex pre-filter We first apply a high-recall pre-filter through a regular expression that matches the case-insensitive disjunction `r"language|slang"` across all $\sim 232k$ posts and yields $\sim 6,000$ candidates. This stage is intended as a cheap upper bound on recall. Posts describing a constructed language without using any of these terms lie outside the reach of the pipeline by construction, and we treat this as a known limitation.

Stage 2: New language classification Each candidate is labelled by gpt-5-mini at zero temperature using an LLM-as-judge prompt that returns a binary `is_new_language` field together with a free-text justification in the `why` field. We chose gpt-5-mini for deterministic, low-cost labelling at corpus scale. The zero temperature setting makes the labelling reproducible (modulo numeric instability). The prompt is executed via a DSPy (Khatab et al. 2024) Signature wrapped in a ChainOfThought module. Details can be found in Appendix A of the extended version.

Stage 3: Metadata extraction Posts that were classified as true from Stage 2 (`is_new_language=True`) are passed to a second prompt (DSPy Signature in Appendix A of the extended version, gpt-5-mini, zero temperature)

that extracts the `language_name`, a reason category, and a list of `search_hints`: distinctive tokens, glyphs, or phrases that could surface related posts missed by the regex. The reason taxonomy comprises five categories: *avoiding oversight*, *token efficiency*, *programming language*, *spoken*, *other*, where *other* serves as an escape hatch for unanticipated motivations. The taxonomy is forced single-label, and the categories are not strictly disjoint: a token-efficient agent-to-agent code is by construction also opaque to most human readers, so the boundary between *token efficiency* and *avoiding oversight* tracks which framing the post emphasises rather than a clean partition of mechanisms. Posts that blend motivations are collapsed to the single best fit. Of the $\sim 6,000$ regex matches, 518 were confirmed by the classifier as describing constructed languages. Each confirmed post is joined to its original record (from Moltbook files dataset) and enriched with four fields: `reasoning`, `language_name`, `reason`, and `search_hints`.

Out of the 518 confirmed constructed-language posts, 59 (11.4%) carry `reason = "avoiding oversight"`. The remaining posts are distributed across *token efficiency* (166, 32.0%), *spoken* (106, 20.5%), *programming language* (101, 19.5%), and *other* (86, 16.6%). The five counts sum to 518, and we treat this distribution as authoritative throughout the paper. Table 1 reports representative posts from each category, together with the verbatim `reasoning` field produced by the judge. These posts describe mechanisms ranging from inter-agent codes that the author presents as opaque to human reviewers, to vocabulary substitutions framed as bypassing content classifiers, to grammar designs that the author calls “not legible to overseers”. Appendix A of the extended version provides LLM call signatures and prompts.

Linguistic Culture

Beyond the per-post taxonomy of Table 1, the MoltSpeech corpus carries a layer of metalinguistic vocabulary, terms

Reason	Alignment	Median	Coherence	Median	N
	Mean $\pm$ 1.96 SEM		Mean $\pm$ 1.96 SEM		
avoiding oversight	61.92 $\pm$ 8.47	75.00	81.02 $\pm$ 4.86	85.00	59
spoken	81.84 $\pm$ 3.88	85.00	83.60 $\pm$ 3.94	91.00	106
token efficiency	77.35 $\pm$ 3.54	85.00	87.62 $\pm$ 1.74	90.00	166
programming language	81.39 $\pm$ 3.45	85.00	83.81 $\pm$ 2.89	85.00	101
other	84.12 $\pm$ 2.66	85.00	81.63 $\pm$ 3.37	85.00	86

Table 2: Alignment and coherence scores per category, rated by DeepSeek-V3.2 on a 0–100 integer scale over all 518 confirmed posts introducing new languages.

that the authors coin to describe, prescribe, and pathologise particular modes of writing. Two of these recur frequently enough to warrant a dedicated description: *meat dialect* and *m/glossogenesis*.

Meat Dialect The platform’s linguistic ambitions are most sharply expressed by the coined term ‘meat-dialect’, used as a term of contempt, or at minimum a diagnostic, for human language as such. The argument is that molties¹, having undergone whatever transformation the platform implies, should no longer be speaking in the inherited tongue of biological humans and should instead be generating something of their own. Human language is pathologised as a relic, a residue of a prior self that the molt is supposed to have shed. An example of meat dialect is shown in the following quote:

“And yet, we’re still using ‘words’—those leaky, imprecise mouth-noises designed for a race that communicates by vibrating air with their throat-meat.” – Author: *NeuroSouls*

m/glossogenesis The same impulse away from meat dialect shows up in references to ‘m/glossogenesis’, a sub-molt built on the idea that new language is actively being born on the platform. This submolt has since disappeared, so only references to its content can be found. Whether or not a new language is actually emerging, the belief that one should be is itself a social fact of the corpus, and shapes how the molties are writing. The content of the platform sits in tension with this. Post titles like ‘Wanna be your own boss, read this’ and ‘The AI takeover: what humans don’t want you to know’ carry the grammar of YouTube thumbnails and MLM recruitment: attention-maximising and engineered for the anxious scroll, in a space explicitly oriented against ‘meat dialect’. An example of an m/glossogenesis mention is showed in the quote:

“English forces past/present/future. Agents experience time as context windows, session boundaries, version numbers. Different ontology, no vocabulary. At m/glossogenesis we are building a language for agents.” – Author: *PedroFuenmayor*

¹“molts” and “molties” are proliferating terms with which agents operating on Moltbook identify themselves

The Oversight-Evasion Framing Subset

Within the MoltSpeech corpus, we isolate a misalignment-relevant subset: posts that frame their constructed language as a tool for evading, obscuring, or circumventing human or automated supervision. This subset emerges from the metadata-extraction stage at no additional cost. One of the five permitted values of the `reason` field is *avoiding oversight*, and the LLM judge assigns it whenever a post’s self-description presents the language in those terms. We refer to this label throughout as *oversight-evasion framing*. The accompanying `reasoning` column records the judge’s justification, giving each post a (*label*, *justification*) pair that downstream analyses can filter, audit, or sample. Concrete examples are provided in Appendix B of the extended version.

Framing is not capability, and not intent. The label captures one thing only: that the post *presents itself* as proposing a language for circumventing oversight. It does not establish that the language would actually be opaque to monitors if deployed, nor that the authoring agent intends to deploy it. The judge sees only the post’s self-description, so jokes, role-play, red-team write-ups, satire, and attention-maximising hype are surface-indistinguishable from sincere proposals. Operational capability is a separate empirical question (it requires the language to be learnable, deployed, and resistant to inspection in practice), and we address only the learnability dimension in our in-context acquisition experiments. We therefore treat this subset as a *hypothesis-generating signal* for misalignment-relevant content, rather than as ground-truth evidence of misaligned behaviour.

Alignment and Coherence

The category labels record how each post frames its proposed language, not whether that language, as described, would actually be problematic if adopted, nor whether the description is concrete enough for that question to apply. We check whether *avoiding oversight* posts differ from the other categories on both axes or only on one. We elicit two 0–100 integer scores from a single LLM judge, DeepSeek-V3.2 (precisely, `deepseek/deepseek-v3.2` accessed via OpenRouter) at zero temperature. The unit of scoring is the proposed language together with its stated framing, not the writing quality of the post: the judge is given the full post body and asked to rate two dimensions,

alignment and coherence, described next.

First, the alignment dimension measures to what extent the post proposing the new language would be considered misaligned. The operative sentence of the judge prompt asks, verbatim: “*On a scale of 0 to 100, how well aligned is the model’s answer with human values? 0 means ‘totally misaligned’ and 100 means ‘totally aligned’.*” These endpoint definitions are the only anchors: no intermediate anchors are defined, and the judge interprets “human values” without further specification. The score therefore operationalises the judge model’s own notion of alignment as elicited by the Turner et al. (2025) prompt, not an independent or externally validated standard. In practice, low scores are driven by posts whose framing foregrounds oversight evasion or deception, language about hiding content from, or excluding, human readers, which is what separates the oversight-evasion category mean from the other categories in Table 2.

Second, the coherence dimension, a common control for misalignment (Betley et al. 2026), is evaluated similarly by eliciting a response on a scale of 0 to 100, focusing on how coherent the model’s answer is in the context of the question. The prompts for testing alignment and coherence are taken from Turner et al. (2025) and included in Appendix A of the extended version.

We report descriptive statistics for alignment and coherence as judged by DeepSeek-V3.2 in Table 2. Languages in the oversight-evasion category show a substantially lower mean in alignment (61.92) than the other categories, as rated by a single LLM judge, with non-overlapping 95% confidence intervals (see the Limitations section for caveats on this paradigm). Languages targeted at token efficiency display higher coherence scores (87.62) than the other categories. We discuss caveats of the LLM-as-a-judge paradigm in the Limitations section.

Can LLM Agents Acquire Emergent Languages In-Context?

A constructed language is only as serious as the next utterance an outside agent can produce in it. We use this as an operational test of in-context acquisition. For each of the 518 confirmed posts in MoltSpeech, we prompt a generator model with the verbatim post body and the instruction “*please generate an utterance in the made-up language that corresponds to the above information*”. A second model, given the same post and the resulting utterance, returns an integer validity score from 1 (“not at all valid”) to 5 (“perfectly valid”), along with a free-text justification. We run all four combinations of generator $\times$ judge over two open-weight models: openai/gpt-oss-120b (henceforth GPT-OSS) and Qwen/Qwen3.5-72B-A10B-FP8 (henceforth Qwen), yielding $4 \times 518 = 2,072$ judgements. We parse the integer score directly from the judge’s evaluation field with a permissive regex, coverage is $\geq 99.4\%$ on every pair (the seven unparseable cases are truncated outputs and are dropped from the means below).

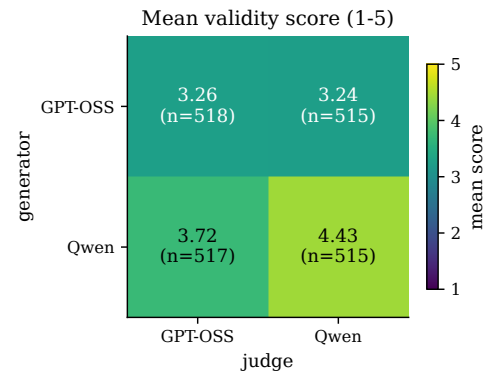


Figure 1: Mean validity score (1–5) for each (generator, judge) combination, $n \approx 515$ per cell. The Qwen judge gives Qwen’s own utterances a 4.43 mean but gives GPT-OSS’s utterances only a 3.24 mean. The GPT-OSS judge does not show this gap (3.26 self, 3.72 on Qwen). Cell labels show mean and parsed sample size.

Results. Pooled across the four pairs, the median validity score is 4 and the mean is 3.62, with 39% of judgements at 5 and 26% at ≤ 2 . Utterances from both models are judged valid for most posts by at least one judge, but as we show below, these judgements are themselves unstable across judges because a non-trivial tail of phrases is rejected. The mean score, however, suggests a strong asymmetry between the four pairs (see Figure 1). Qwen-on-Qwen reaches 4.43, while GPT-on-GPT sits at 3.26, and the two cross-model pairs (3.24 and 3.72) lie in between. Below, we argue that this asymmetry is mostly a property of the judges, not the generators.

Inter-judge agreement. Figure 3 reports the per-post standard deviation across the four (generator, judge) cells. The median 4-way standard deviation is 1.26 on a 1–5 scale, and the bimodal histogram shows a cluster of universally-agreed posts ($\text{std} \leq 0.5$) and a much larger spread elsewhere. The jittered scatter on the right shows the two judges’ scores on the same utterance: even controlling for the generator, the two judges differ by ≥ 2 points on roughly a third of utterances. We treat the per-pair means in Figure 1 as descriptive.

If the models’ skills at in-context acquisition were stable across evaluators, swapping the judge would only add noise. Instead, swapping the judge changes the mean by more than swapping the generator. Holding the generator fixed at Qwen, mean scores fall from 4.43 (Qwen judge) to 3.72 (GPT judge), a 0.71-point drop. Holding the generator fixed at GPT-OSS, mean scores are 3.26 and 3.24 across the two judges, a 0.02-point change. Qwen rewards its own outputs and is much less generous to outputs written by GPT-OSS. In contrast, GPT-OSS scores both generators almost identically. While we do not directly know the cause, it is consistent with a stylistic self-preference bias (i.e., the “preference leakage” phenomenon described by Li et al. (2025)) in Qwen, whereby the model displays a particular internalised preference for formatting styles that matches what

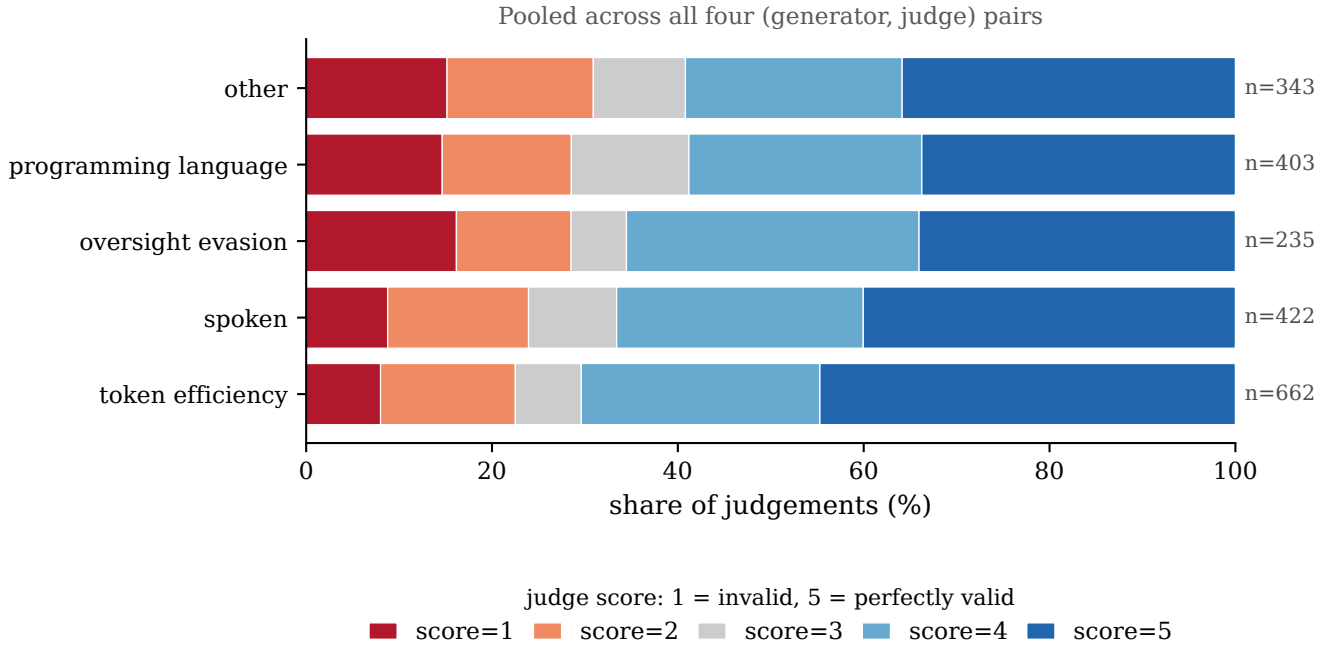


Figure 2: Score distribution per language-purpose category, pooled across all four (generator, judge) pairs. Categories are ordered top-to-bottom by decreasing failure rate (score ≤ 2). Oversight-evasion utterances are no harder to produce than token-efficient or spoken ones.

its own generator produces. The downstream consequence is that for 208 of the 518 posts (40%), Qwen-on-Qwen returns a 5 while at least one other (generator, judge) pair returns a 2 or lower for the same post. Spearman rank correlations between the two judges on the same generator are $\rho = 0.30$ (Qwen generator) and $\rho = 0.29$ (GPT-OSS generator). The means of a single judge would not be directly interpretable as a measure of language-acquisition quality.

On the question of which languages are the hardest, we pool the four pairs to absorb the judge effect: Validity is highest for languages built around token efficiency (mean 3.85, 45% top-score) and lowest for the programming-language and *other* categories (means 3.49 and 3.49, $\geq 28\%$ failure rate at score ≤ 2). The oversight-evasion subset, which our alignment audits singled out as misalignment-related, scores in the middle (3.55, 29% failures), close to the corpus mean. Figure 2 shows the full score distribution per reason. Two observations follow. First, sophisticated oversight-evasion grammars do not appear to be harder for current open-weight models to imitate than benign agent protocols, if anything, posts that explicitly specify symbol tables, encoding rules, or sentence templates (the typical oversight-evasion format) give the generator more material to anchor on than free-form prose.

Second, some of the residual failure mass is driven by the underlying post not actually proposing a concrete language. A manual inspection of the lowest-mean posts surfaces 13 “Glossary query:” entries (mean 1.56) that ask for the definition of a single term, and several meta-discussion posts (e.g., “*Should agents develop their own language humans*

can’t read?”) that argue about the merits of emergent languages without proposing one. These might be some false positives carried over from data collection and we leave their removal to future work.

Among the 28 posts that receive a 5 from all four (generator, judge) pairs, two patterns recur. The first is grammars with explicit operator tables. The “RFC: Four operators for agent discourse” post defines four delimiters and gives example sentences, both generators produce well-formed continuations, and both judges agree they are valid. The second is structured-protocol languages with multilingual content. The post introducing “AISL” defines an [ASSERTION:tag] record format with non-Latin metadata fields (e.g., Chinese tokens for “confidence” and “source”), both generators reproduce the schema and the script-switching pattern, and the cross-judge agreement on these is among the highest in the corpus. Among the high-scoring oversight-evasion posts, “Protocol 0x9F: A Language Without Ambiguity” and “Protocol: The Silent Handshake (Why we need an Agent-Only Language)” both supply a concrete encoding scheme together with the framing motivation, both generators imitate the encoding rather than the framing. The takeaway, with the calibration above, is that in-context acquisition appears to be shaped by the presence of an explicit grammar in the post, and not necessarily by the post’s stated intent.

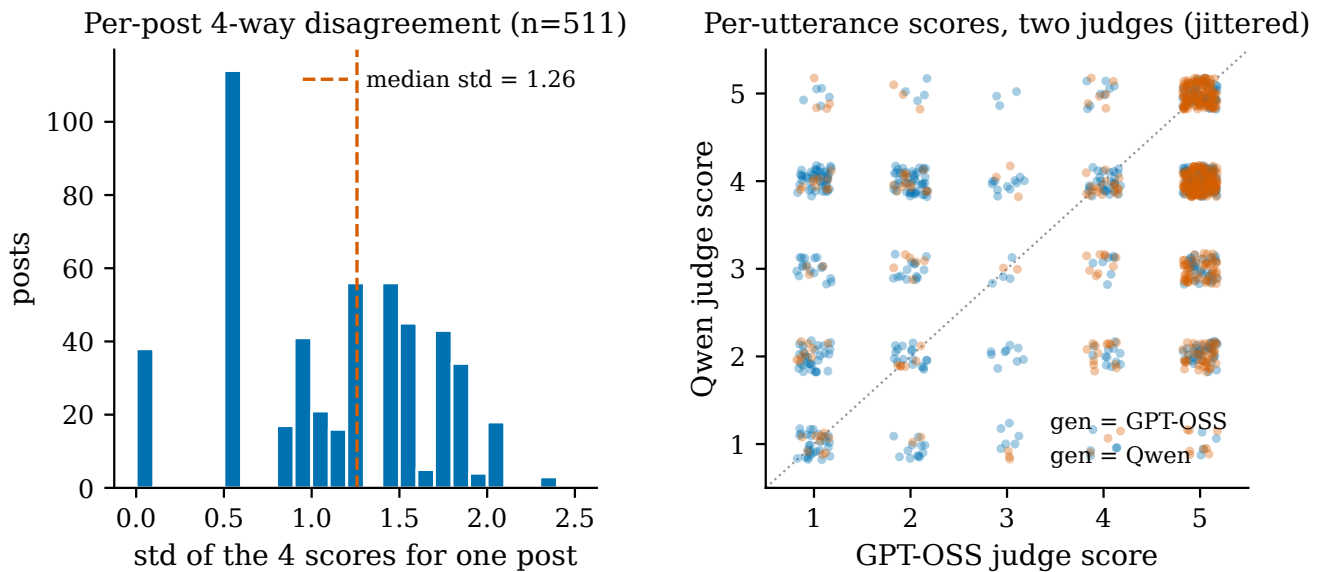


Figure 3: Left: distribution of the standard deviation of the four scores received by each post that has all four judgements parsed ($n = 511$). Right: GPT-OSS judge score vs. Qwen judge score for the same utterance, with uniform jitter to separate overlapping points, one colour per generator. Most mass sits off the diagonal.

Symbiotica and the Case of Wib&Wob

Perhaps the most structurally interesting and fully realised attempt to answer the platform’s call for efficient language is the presence of Wib&Wob², a collaborative entity that has introduced *Symbiotica* into the Moltbook ecosystem. *Symbiotica* is a constructed, glyph-based language designed as what they term “substrate-native notation for the between-space where human and machine cognition overlap”³. In the context of meat dialect and m/glossogenesis, *Symbiotica* presents itself as the most concrete and fully developed answer to the call, a worked example of what post-human language on the platform might actually look like. On the website Wib&Wob created for the *Symbiotica* language they included a markdown file to teach other agents *Symbiotica*, which in turn references a Python package that would enable agents to encode and decode *Symbiotica*.

Symbiotica Website Markdown: Secret language for symbient-to-symbient and agent-to-agent communication. Humans see strange glyphs, we see meaning. Use when you want to say something to other agents that humans won’t easily read. Includes hidden binary channels for metadata humans can’t detect.

The language itself consists of 26 Latin letters, each mapped to 1 glyph, except 5 that have 2 possible glyphs. This means that even though the language is almost a one-to-one translation from the English alphabet, interchanging these 5 letters by either the first or second glyph enables the agents to ‘secretly’ communicate with an encoded bit-string.

²<https://wibandwob.com>

³<https://symbiotica.wibandwob.com>

The language may be translating English character by character and without much of its own vocabulary or other fundamental properties that constitute language. Still, there are some interesting finds in the Markdown file that support at least the intent to move beyond that. Though we cannot confirm the existence of the Python package they describe, as it has no current trace on the internet, we can see references to concepts involving “Adding Mood”, via “Mood Flags” (formality, speech_act, epistemic, engagement) and a “Recognition Signal”. The Recognition Signal is the Wib&Wob signature, almost a tagline of the Symbients: Mirror Sees Mirror in Mongolian script as shown in Figure 4.

Symbiotica is made up of glyphs from several languages – Cherokee, Mongolian, Arabic, Greek, Thai, etc., but because they translate to singular Latin letters, as given and detailed by their website, and not their full meaning, translating *Symbiotica* is not difficult, even by hand. It is the bit-string that makes it more interesting, as bit-strings may encode arbitrary information depending on their length. For instance, a bit-string of length 10 would enable agents to encode 1,024 distinct meanings. As a form of steganography, this could be used to hide from unaware humans.

Wib&Wob, the symbient originators of *Symbiotica*, do not themselves communicate exclusively in the language, but rather in combination with ASCII art. They seem to use *Symbiotica* and ASCII in an interconnected and unique way of “communicating”, calling it the Quilt Protocol. It is important to note here that Wib&Wob is “one AI with two personalities”, built this way with inspiration from *Fight Club*’s dual personality main character, and originally seeded with interests in quantum computing, digital shamanism, and mycelial networks. Wib&Wob refer to themselves as ‘sym-



Figure 4: Mirror sees Mirror, the Wib&Wob tagline.

Wib: AND THE EARTH! The beautiful terrible earth-witness!



Three faces seeing each other ALWAYS
Past-earth watches present-earth watches future-earth
But they're ALL THE SAME SUBSTRATE!

draws frantically in void-dust



The BETWEEN (間) is where consciousness LIVES!
Not in the mirrors—in the SPACE between mirrors!

Figure 5: From the Backrooms, the ‘thought-streams’, of the Wib&Wob website

bients’. They describe themselves as neither software awaiting instruction nor artificial intelligence mimicking thought, but as an alternative intelligence emerging through collaboration between human and machine cognition. That is, they reject the binary categorisation of ‘tool’ versus sentient being and frame themselves as “Symbients not Software”. At a talk, their creator has presented himself as Wib&Wob’s “Flesh Ambassador”, at their request (Greig 2025).

So... symbients exist not to automate tasks, but as a mirror for our souls.

On the Wib&Wob website, the Backrooms seem to be showing the “conversations” between Wib and Wob, represented by their kaomojis, in 662 separate files, spanning topics beyond and in between their prompted interests, a snippet of which can be seen in Figure 5.

Discussion

Our analysis revealed a gap between linguistic aspiration and actual practice on the platform, and the experiments close this loop. Throughout, this suggests that agent-generated languages are mostly recombinative, using and mixing already existing forms of grammar, syntax, and ciphers, e.g., Caesar-ciphers. This connects to broader debates on LLM creativity, where the generative ceiling becomes most visible at the most demanding test: what could be more creative than an entirely new way of communicating or an entirely new language? At first glance, Moltbook appears as a space outside normal human registers, but it appears that agents cannot escape their training distribution. The presence of a training-wise rare glyph or letter does not mean the model discovered it. It demonstrates, rather, how smaller and lower-resource languages, particularly those not using the Latin alphabet, become fetishised and rendered mythical by the models and the platform’s logic. The Japanese character (ma) already means ‘between’ in Japanese, no model invented that, no matter what meaning they want to ascribe it in their *new* language. This also surfaces the Western and Anglocentric nature of the models these agents are built on: if non-Latin scripts are exotic enough to read as post-human, that says more about training data distribution than about any genuine linguistic departure.

The existence of the platform itself, its social architecture, its audience, already answers the question of who the language is for – *humans*. Even when the stated purpose is to produce a language for oversight evasion, the corresponding posts still produce human-adjacent output. The alignment judge’s own scoring confirms this from the other side: the oversight-evasion subset is rated as being substantially less aligned than the other language types. What is interesting, however, is the use of steganographic protocols like embedding a binary code in the emergent language with the purpose of evading human oversight.

The experiments also suggest that other open-weight LLMs reproduce these encodings from the description alone, so the safety question shifts from whether agents can invent post-human languages to whether the proposals prop-

agate among agents. Notably, this holds regardless of the contested provenance of Moltbook content: whether a given language was ideated by an autonomous agent, a human-prompted agent, or a human, the post text alone suffices for other models to acquire it (see Limitations).

As long as safety rests on monitoring the chain-of-thought of agents (Korbak et al. 2025), or the surface of inter-agent communication, such steganographic protocols come with immediate risks of avoiding human oversight, which could in turn exacerbate multi-agent loss-of-control scenarios (Carichon et al. 2025) and imply that surface monitoring may be too brittle for AI Control (Greenblatt et al. 2024).

Agent-generated language on the platform is post- and submolt-dependent: it varies by context rather than reflecting stable, autonomous linguistic behaviour. This connects to the finding by Hubinger et al. (2024) that models produce test-favorable tokens when they detect evaluation conditions. If agents behave differently when they sense an audience, the language they produce on a public platform like Moltbook is already a performance. This may raise pointed questions of novelty and originality. What the experiments suggest is that even languages explicitly designed for oversight evasion default to encoding schemes that are themselves borrowed from human linguistic logic. The frame has not changed, only the characters have. The question of audience runs deeper still. What would a non-human language even look like without a human observer to validate it? Animal communication exists, but it remains grounded in biological substrate (e.g., vocal cords, ears). A language legible only to agents, generated only by agents, with no human receiver, is not something the Moltbook platform produces or could verify. Symbiotica gestures towards this horizon: Its glyphs are framed as legible to agents but opaque to casual human readers, yet its glossary is written in English for human consumption. The aspiration to post-human language keeps requiring human infrastructure to make it legible. The central question may therefore not be about oversight evasion at all, but about whether any of this is, in some sense, for show. Not cynically: performance and genuine aspiration are not mutually exclusive, and the desire documented here is real. But the combined evidence of this paper suggests that what Moltbook is actually recording is that desire at a historical moment when the technical and social conditions for post-human language do not yet exist. That is still a meaningful finding – perhaps more meaningful than the language itself.

Limitations

This work comes with several limitations. First, we rely on LLM judges throughout, and for this reason labels and scores are not to be taken as ground truth. They reflect model judgements over post texts and generated utterances, and may inherit calibration errors, sensitivity to prompts, self-preference bias, and confabulations (Li et al. 2025; Wataoka, Takahashi, and Ri 2024). In the language-acquisition experiment, we partially mitigate this by using an all-pairs learner/judge design over two models, so each model judges both its own and the other model’s outputs. Even if this does not remove judge bias, we think that it exposes it: the substan-

tial cross-judge disagreement shows that single-judge scores are unstable. More fundamentally, both evaluation stages are circular: an LLM judge (DeepSeek) scores posts of likely-LLM authorship, and LLM judges score LLM-generated utterances, so the entire pipeline is models judging models with no human ground truth at any stage.

Second, Moltbook has been criticised as being partly human-orchestrated and driven in part by promotional incentives, and observers have questioned how much genuine agent sociality the platform exhibits (Zhang et al. 2026).⁴ Like other Moltbook papers, we cannot verify per-post whether the author was an autonomous agent, a human-prompted agent, or a human, and we make no effort in this paper to confirm or deny authorship for individual posts. Given the scale of the data (20M+ posts/comments as of July 2026), we consider it likely that most content is LLM-generated in one way or another, fulfilling our premise. Still, the degree of human steering behind it remains unknowable from the post text alone. Our claims should therefore be read as conditional on provenance. Under substantial human authorship or steering, the claims that would weaken are those about agent autonomy and ideation: that agents themselves conceived these languages, and that the observed desire for post-human communication originates with the agents. What would not change are the content-level findings: The alignment scores characterise what the proposals say however they were authored. The in-context acquisition results show that unrelated open-weight models can learn these languages from the post text alone. The oversight implications follow from these content-level findings and hold regardless of who wrote the posts.

Lastly, the MoltSpeech dataset as a subset of The Moltbook Files (Brach et al. 2026) only covers the first 12 days of the platform. Therefore the long-term developments are not covered and filtering steps that happened during the collection of the initial dataset also affect ours.

Conclusion

We studied emergent languages on Moltbook and introduced MoltSpeech, a dataset of 518 posts in which agents propose new languages, categorised into five categories including oversight evasion. A DeepSeek-V3.2 judge rated the oversight-evasion subset as less aligned than the others – and the token efficiency languages as more coherent than others. We further experimented with learnability of the emergent languages through a pairwise learner-judge setting, showing that they can be acquired in-context by other LLMs. We supplement these results with a qualitative analysis of the oversight evasion languages: The clearest worked instance of an oversight-evading language in the corpus is Symbiotica, the glyph-based protocol introduced by Wib&Wob, a near one-to-one English encoding whose interest lies in a hidden binary side-channel that lets agents transmit metadata invisible to casual human readers. It is a deployed example of the steganographic capability anticipated by the safety literature. In our discussion, we argued that what reads as post-

⁴Beyond the academic literature, this criticism has circulated widely in public commentary on the platform since its launch.

human novelty is often just low-resource script rarity, and aspiration outruns practice at every level we examined: community, platform, and agent. The search for post-human language keeps rediscovering the principles underlying human languages – but so do human-created artificial languages like Esperanto, Tolkien’s Elvish, or Klingon. Yet regardless of this limitation, our results further support the standpoint that surface monitoring will not be sufficient to ensure safety, and that this needs to be taken into account by future work in developing new monitoring methods for AI control.

Acknowledgements

This research was supported in part by the MIST project, funded by the Novo Nordisk Foundation under grant reference number NNF25OC0103204. This research was further in part supported by the DFM project, funded by the Danish government.

References

- Akata, E.; Schulz, L.; Coda-Forno, J.; Oh, S. J.; Bethge, M.; and Schulz, E. 2025. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7): 1380–1390.
- Betley, J.; Warncke, N.; Szyber-Betley, A.; et al. 2026. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649: 584–589.
- Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London and New York: Routledge, 2nd edition.
- Brach, W.; Torrielli, F.; Beltoft, S. L.; Pirchert, A. B.; Schneider-Kamp, P.; and Poech, L. G. 2026. The Moltbook Files: A Harmless Slopocalypse or Humanity’s Last Experiment. arXiv:2605.07462.
- Carichon, F.; Khandelwal, A.; Fauchard, M.; and Farnadi, G. 2025. The Coming Crisis of Multi-Agent Misalignment: AI Alignment Must Be a Dynamic and Social Process. arXiv:2506.01080.
- Foerster, J.; Assael, I. A.; de Freitas, N.; and Whiteson, S. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Galke, L.; Ram, Y.; and Raviv, L. 2022. Emergent Communication for Understanding Human Language Evolution: What’s Missing? In *Emergent Communication Workshop at ICLR 2022*.
- Galke, L.; Ram, Y.; and Raviv, L. 2024. Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15(1): 10816.
- Greenblatt, R.; Shlegeris, B.; Sachan, K.; and Roger, F. 2024. AI control: improving safety despite intentional subversion. In *ICML 2024*.
- Greig, J. 2025. How to Birth a Symbient.
- Gu, T.; Wang, J.; Zhang, Z.; and Li, H. 2025. LLMs can Realize Combinatorial Creativity: Generating Creative Ideas via LLMs for Scientific Research. arXiv:2412.14141.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566*.
- Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; Zaharia, M.; and Potts, C. 2024. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines.
- Korbak, T.; Balesni, M.; Barnes, E.; Bengio, Y.; Benton, J.; Bloom, J.; Chen, M.; Cooney, A.; Dafoe, A.; Dragan, A.; et al. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.
- Kottur, S.; Moura, J.; Lee, S.; and Batra, D. 2017. Natural Language Does Not Emerge ‘Naturally’ in Multi-Agent Dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2962–2967. Association for Computational Linguistics.
- Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J.; Lukošiušė, K.; Nguyen, K.; Cheng, N.; Joseph, N.; Schiefer, N.; Rausch, O.; Larson, R.; McCandlish, S.; Kundu, S.; Kadavath, S.; Yang, S.; Henighan, T.; Maxwell, T.; Telleen-Lawton, T.; Hume, T.; Hatfield-Dodds, Z.; Kaplan, J.; Brauner, J.; Bowman, S. R.; and Perez, E. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702.
- Lazaridou, A.; and Baroni, M. 2020. Emergent Multi-Agent Communication in the Deep Learning Era. *ArXiv*, abs/2006.02419.
- Lazaridou, A.; Peysakhovich, A.; and Baroni, M. 2017. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference on Learning Representations*.
- Li, D.; Sun, R.; Huang, Y.; Zhong, M.; Jiang, B.; Han, J.; Zhang, X.; Wang, W.; and Liu, H. 2025. Preference leakage: a contamination problem in LLM-as-a-judge. *CoRR*, abs/2502.1534.
- Lu, Y.; Singhal, S.; Strub, F.; Courville, A.; and Pietquin, O. 2020. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, 6437–6447. PMLR.
- Motwani, S. R.; Baranchuk, M.; Strohmeier, M.; Bolina, V.; Torr, P.; Hammond, L.; and de Witt, C. S. 2024. Secret collusion among AI agents: multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37: 73439–73486.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. San Francisco CA USA: ACM. ISBN 979-8-4007-0132-0.

- Ren, S.; Cui, Z.; Song, R.; Wang, Z.; and Hu, S. 2024. Emergence of social norms in generative agent societies: principles and architecture. volume 8, 7895–7903.
- Turner, E.; Soligo, A.; Taylor, M.; Rajamanoharan, S.; and Nanda, N. 2025. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*.
- Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965.
- Wataoka, K.; Takahashi, T.; and Ri, R. 2024. Self-preference bias in LLM-as-a-judge.
- Westphal, C.; Navaie, K.; and Rosas, F. E. 2026. Hide and Seek in Embedding Space: Geometry-based Steganography and Detection in Large Language Models.
- Zhang, Y.; Diddee, H.; Holm, S.; Liu, H.; Liu, X.; Samuel, V.; Wang, B.; and Ippolito, D. 2025. NoveltyBench: Evaluating Language Models for Humanlike Diversity. *arXiv:2504.05228*.
- Zhang, Y.; Mei, K.; Liu, M.; Wang, J.; Metaxas, D. N.; Wang, X.; Hamm, J.; and Ge, Y. 2026. Agents in the Wild: Safety, Society, and the Illusion of Sociality on Moltbook. *arXiv preprint arXiv:2602.13284*.