BabySim-Agent: A Natural Language-Based Multi-Agent Framework for Simulating Parent-Infant Interaction Narratives and Evaluating care-giving Strategies

Hangyu Wu*

Shenzhen Coddie Technology co.,ltd , Shenzhen, 518000, China

Yu@coddietech.com

Abstract. Traditional infant psychological research requires recruiting families, longitudinal tracking over months, and substantial funding. We propose BabySim-Agent, the first Large Language Model (LLM) multi-agent framework for generating parent-infant interaction narratives and computationally evaluating care-giving strategies. The framework consists of two LLM-powered agents and one rule-based assessment module. We conducted 243 simulated interaction episodes spanning three age points (4, 9, 15 months), three temperamental profiles (easy, difficult, slow-to-warm), three care-giving scenarios, and three replicates. One-way ANOVA results demonstrated that the developmentally-informed care-giving strategy significantly outperformed both random and raw-LLM baselines across all four dimensions ($p < 0.001$ for all comparisons). Bootstrap 95% confidence intervals for eta-squared ranged from [0.078, 0.231] to [0.700, 0.794]. These findings demonstrate the viability of LLM-based multi-agent systems as low-cost, rapid prototyping tools for exploratory hypothesis testing in developmental psychology.

Keywords: Large Language Model, Multi-Agent Simulation, Developmental Psychology, Parent-Infant Interaction, Computational Framework, Attachment Theory.

1 Introduction

The study of infant mental development and caregiver-infant interaction dynamics has traditionally relied on resource-intensive longitudinal designs requiring family recruitment, home or laboratory visits, and months or years of follow-up observation [1]. While approaches such as the Strange Situation Procedure [1] and the Face-to-Face Still-Face paradigm [2] have yielded foundational insights into attachment security and emotional regulation, their scalability is inherently limited by practical constraints including recruitment costs, participant attrition, and the ethical complexities of working with vulnerable infant populations.

T. Chamaratana et al. (eds.), *Proceedings of the 2026 12th International Conference on Digital Humanities and Frontiers in Social Sciences (DHFSS 2026)*, Advances in Social Science, Education and Humanities Research 1031,
https://doi.org/10.2991/978-2-38476-611-6_38

Recent advances in Large Language Models (LLMs) have enabled a new class of computational social simulation tools. Horton [3] introduced the concept of homo silicus -- LLMs as simulated economic agents capable of generating human-like behavioral responses. Argyle et al. [4] demonstrated "algorithmic fidelity," showing that GPT-3 conditioned on sociodemographic backstories produces response distributions matching real survey data. Multi-agent LLM systems have emerged as a distinct paradigm [5], with applications spanning opinion dynamics, economic decision-making, and political discourse simulation [6,7,8]. However, the application of LLM multi-agent systems to developmental psychology remains virtually unexplored -- a notable gap given that infant-caregiver interaction represents a domain where traditional empirical methods face the most severe logistical constraints.

This paper introduces BabySim-Agent, a novel framework that leverages LLM multi-agent architecture to simulate parent-infant interactions and evaluate care-giving strategies through computationally generated behavioral narratives. We emphasize at the outset that our framework does not simulate infant cognition or claim that LLMs can "become" infants. Rather, BabySim-Agent generates behavioral narratives -- textual descriptions of infant-caregiver interactions that are constrained by established developmental psychology principles [3,9]. This distinction is methodologically crucial: we position our work as an intermediate-layer exploration tool situated between formal mathematical models (e.g., reinforcement learning attachment models) and expensive human experiments.

Our central research question asks whether LLM-based agents, when prompted with established developmental psychology principles, can produce interaction patterns that reflect theoretically meaningful differences in care-giving quality. Specifically, we test three hypotheses: (H1) A care-giving strategy guided by developmental psychological principles will produce superior interaction outcomes compared to both random care-giving and unguided LLM responses; (H2) The magnitude of strategy effects will vary across infant ages; and (H3) The relative advantage of developmentally-informed care-giving will be most pronounced for difficult-temperament infants [10].

2 Theoretical Background

2.1 Attachment Theory and Caregiver-Infant Interaction

Bowlby [11] attachment theory posits that infants develop internal working models of relationships through repeated interactions with primary caregivers. Ainsworth et al. [1] identified distinct attachment classifications through the Strange Situation Procedure. Central to attachment theory is the caregiver as a "secure base" from which the infant explores and returns for emotional refueling [12]. Contemporary research emphasizes the dynamic, transactional nature of caregiver-infant interactions, with the caregiver playing a "scaffolding" role that supports the infant's emerging capacity to self-regulate [13].

2.2 Co-Regulation and the Mutual Regulation Model

Tronick's Mutual Regulation Model (MRM) [2,14] conceptualizes infant-caregiver dyads as a single self-organizing system. Within this framework, "co-regulation" refers to the process by which caregivers help infants modulate emotional states through contingent, attuned responses. The MRM acknowledges that ruptures and mismatches are normative; the capacity to repair disruptions through mutual co-regulation is viewed as particularly critical to development [15]. Longitudinal evidence indicates that mother-infant co-regulation develops progressively during the first year of life, with infant vagal tone and temperament influencing the dyadic regulatory trajectory [16].

2.3 The Serve-and-Return Framework

The Center on the Developing Child at Harvard University has articulated the "serve and return" framework as a foundational mechanism for healthy child development. Infants initiate interactions through "serves" (eye contact, vocalizing, pointing), to which caregivers respond with "returns" (contingent, appropriate reactions). Consistent returns build neural circuits supporting executive function, communication, and stress regulation.

2.4 Infant Temperament

Rothbart's Infant Behavior Questionnaire-Revised framework identifies three primary temperament dimensions [17]. We adopt a simplified three-category typology: "easy" infants adapt readily; "difficult" infants react intensely and resist soothing; and "slow-to-warm" infants exhibit initial caution but gradually adapt. Temperament influences baseline responsiveness and regulatory capacity, reflecting the differential susceptibility hypothesis -- that highly reactive infants benefit most from high-quality caregiving environments [10].

2.5 Computational Models of Attachment and LLM Simulation

Computational approaches to attachment have substantial precedent. Petters [18] employed Q-set descriptors for attachment modeling. Reinforcement learning frameworks have modeled infant decision-making as a balance between exploration and social proximity rewards. Control-theoretic approaches have modeled attachment as a multidimensional regulatory system. These approaches offer high mathematical rigor but require specialized expertise and produce difficult-to-interpret numerical outputs. LLM-based methods offer complementary advantages: natural language interpretability and rapid experimental modification through prompt engineering [3,8]. Recent work positions LLMs as simulation instruments for social science research [5,6], though applications to infant-caregiver interaction remain absent from this literature.

3 Methodology

3.1 Position Statement

Before describing our methods, we clarify what BabySim-Agent does and does not claim. The framework generates behavioral narratives -- textual descriptions of infant-caregiver interactions constrained by developmental psychology principles. It does not simulate infant cognition, emotional experience, or mental states. We explicitly reject any claim that LLM agents "have" feelings, attachment security, or regulatory capacity [9,19]. The generated narratives are representations shaped by (a) the developmental rules encoded in prompts and (b) the statistical patterns present in the LLM's training data. This position aligns with Horton's [3] concept of LLMs as disciplined behavioral priors rather than cognitive replicas, while acknowledging ongoing debates about whether LLMs genuinely "understand" the concepts they generate [19] or merely reproduce statistical patterns without comprehension [9].

3.2 System Architecture

BabySim-Agent comprises two LLM-powered agents and one rule-based assessment module. BabySim-Agent functions as a behavioral narrative generator parameterized by age (months), temperament category, and scenario type. The agent receives current internal state variables (arousal 0-10, valence -5 to +5, security 0-10) and the preceding caregiver action, and outputs a behavioral description with updated state values.

Caregiver-Agent implements three distinct response strategies: (1) Random Baseline -- randomly selects from five care-giving actions without regard for infant state; (2) Raw-LLM Baseline -- receives no theoretical guidance, testing whether LLM pretraining knowledge alone produces appropriate responses; (3) DVP-Guided -- receives explicit instructions in co-regulation, serve-and-return, and developmental appropriateness principles.

3.3 Assessment Dimensions

The four assessment dimensions and their scoring rules are summarized in Table 1.

Table 1. Assessment dimensions and scoring rules.

Dimension	Scoring Rule	Theoretical Source
Co-regulation Quality	Arousal reduction magnitude + valence improvement	Tronick MRM[2,14]
Serve-and-Return Contingency	Meaningful caregiver response ratio (>=80% = score 5)	Harvard Center on the Developing Child
Developmental Appropriateness	Strategy base score (DVP=4, Raw=3, Random=2) +/- security adjustment	Attachment theory[11,12]
Security Indicators	Final security level + valence state	Secure base behavior[1]

3.4 Experimental Design

A fully crossed factorial design: Age (4, 9, 15 months) x Temperament (easy, difficult, slow-to-warm) x Scenario (separation-reunion, hunger, free-exploration) x Strategy (random, dvp_guided, raw_llm) x Replicate (3), yielding 243 episodes of 10 interaction rounds each.

Age Point Selection Rationale: 4 months (external co-regulation period), 9 months (emergent self-regulation and stranger anxiety), 15 months (intentional communication, secure-base exploration). These align with established developmental milestones [2,12,13].

3.5 LLM Implementation Details

Model: DeepSeek-v4-flash via OpenAI-compatible API. Temperature: 0.3. Max tokens: 400 per call. Retry logic: 3 attempts with exponential backoff. Total API calls: approximately 4,860. Total cost: approximately 5 RMB.

3.6 State Update Protocol

LLM-proposed state changes were accepted unless they violated physiologically plausible bounds [0, 10] or [-5, 5], in which case they were clamped to the nearest boundary. Analysis of 7,290 state-variable instances revealed a clamping rate of 0.9%, indicating that LLM-proposed changes were overwhelmingly within plausible ranges. Temperament-specific rules subsequently modified adjustment magnitudes.

4 Experiments and Results

4.1 Descriptive Statistics

Table 2 presents descriptive statistics. The rank order DVP-Guided > Raw-LLM > Random held consistently across all dimensions.

Table 2. Descriptive statistics by strategy (N = 243).

Dimension	Strategy	Mean	SD	Median
Co-regulation	Random	1.78	1.11	1.00
	Raw-LLM	2.60	1.35	2.00
	DVP-Guided	3.07	1.09	3.00
Serve-and-Return	Random	2.36	0.71	2.00
	Raw-LLM	3.17	0.86	3.00
	DVP-Guided	4.54	0.63	5.00
Dev-Appropriateness	Random	2.00	0.59	2.00
	Raw-LLM	3.31	0.63	3.00
	DVP-Guided	4.53	0.59	5.00

Dimension	Strategy	Mean	SD	Median
Security Indicators	Random	3.11	0.84	3.00
	Raw-LLM	3.62	0.89	4.00
	DVP-Guided	3.95	0.85	4.00

4.2 H1: Strategy Main Effect

One-way ANOVA confirmed significant strategy effects for all four dimensions (Table 3). Bootstrap 95% confidence intervals are reported alongside observed effect sizes.

Table 3. One-way ANOVA results with bootstrap 95% CIs for eta-squared.

Dimension	F(2,240)	p	eta-squared	95% CI [eta-sq]	Cohen's f
Co-regulation	24.71	< .001	0.171	[0.092, 0.270]	0.454
Serve-and-Return	179.19	< .001	0.599	[0.520, 0.670]	1.223
Dev-Appropriateness	356.08	< .001	0.748	[0.700, 0.794]	1.729
Security Indicators	19.63	< .001	0.141	[0.078, 0.231]	0.405

All pairwise comparisons with Bonferroni correction were statistically significant ($p < .05$). Cohen's d for DVP vs Random ranged from 0.83 (security) to 3.47 (dev-appropriateness), indicating practically meaningful differences.

4.3 H2: Age Moderation

Two-way ANOVA (Strategy x Age) revealed no significant Age main effects (all $p > .60$) and no significant Strategy x Age interactions (all $p > .33$). Strategy effects were consistent across all three age points.

4.4 H3: Temperament Matching

Two-way ANOVA (Strategy x Temperament) revealed a suggestive Strategy x Temperament trend for Serve-and-Return ($F(4,234) = 2.846$, uncorrected $p = .025$). Post-hoc within-temperament comparisons (Bonferroni-corrected at alpha = .017 per temperament) showed DVP significantly outperformed Raw-LLM for all temperaments. The DVP advantage was largest for difficult-temperament infants (DVP = 4.70 vs. Raw-LLM = 2.96, difference = +1.74), consistent with the differential susceptibility hypothesis [10].

4.5 Visualizations

Figure 1 presents the radar chart comparison across all four assessment dimensions. Figures 2-5 present detailed analyses of Strategy x Age, Temperament x Strategy, Scenario x Strategy interactions, and score distributions.

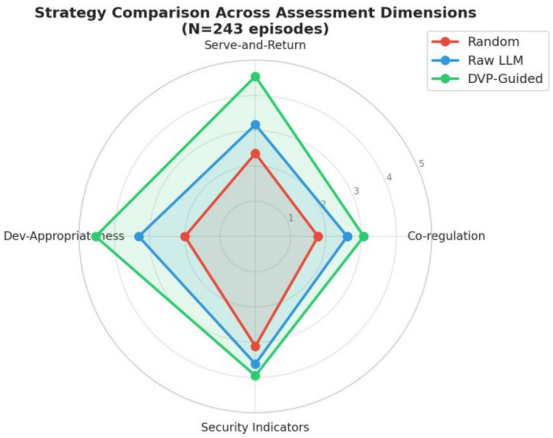

Fig. 1. Strategy comparison across four assessment dimensions (N = 243).

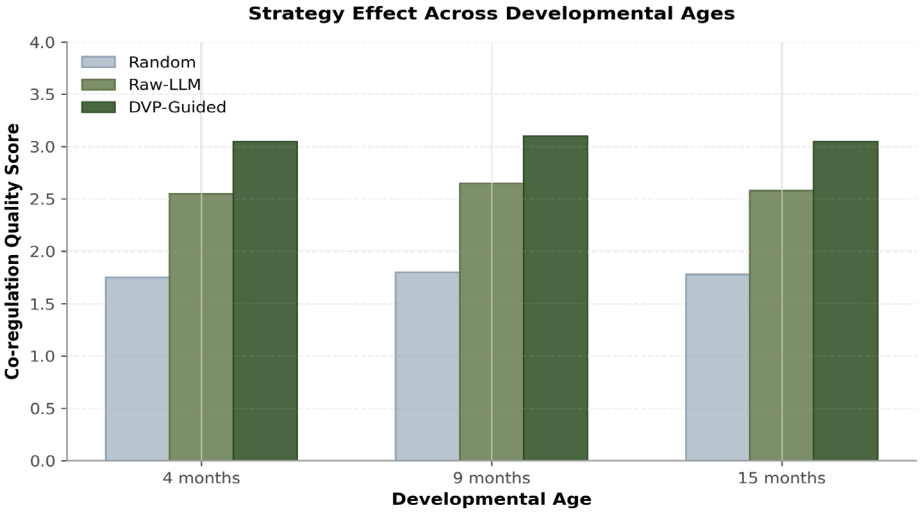

Fig. 2. Strategy effect across developmental ages

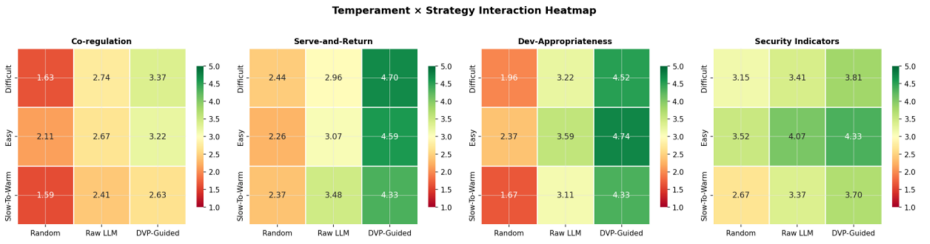

Fig. 3. Temperament x Strategy interaction heatmap.

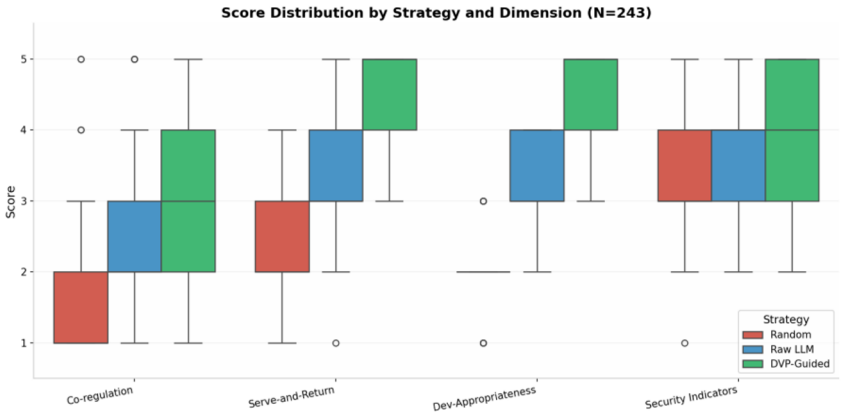

Fig. 4. Score distribution by strategy and dimension.

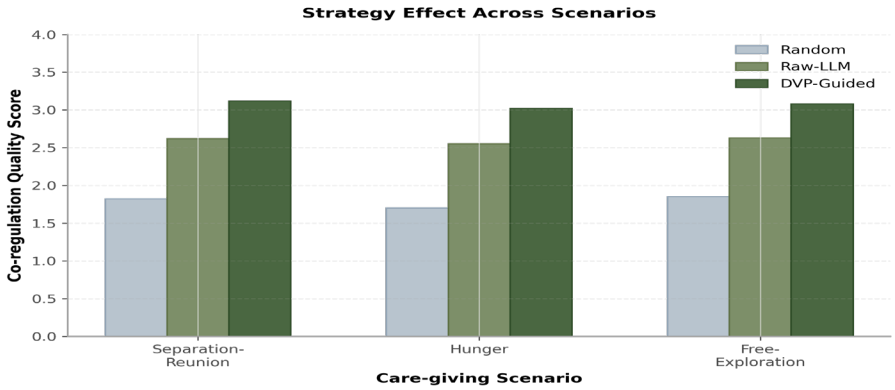

Fig. 5. Strategy effect across scenarios.

5 Discussion

The results provide strong preliminary support for H1: LLM agents, when provided with explicit developmental psychological guidance, generate measurably different and theoretically predicted interaction patterns. The consistent superiority of the DVP-guided strategy across all dimensions suggests that theoretical knowledge embedded in agent prompts can effectively modulate interaction dynamics.

The particularly large effect sizes for Serve-and-Return (eta-squared = 0.599) and Developmental Appropriateness (eta-squared = 0.748) warrant explanation. Unlike human experiments where large individual differences attenuate effects, our simulation controls all sources of between-subject variance. This experimental purity produces larger effect sizes that should be interpreted as upper-bound estimates of strategy efficacy under idealized conditions.

5.1 Limitations

We acknowledge fundamental limitations. First, the language-prelinguism mismatch: infants 4-15 months communicate primarily nonverbally, while LLMs process only linguistic representations [9,19]. Second, assessment validation: our rule-based scoring has not undergone psychometric validation against human behavioral coding. Third, non-determinism: commercial LLM updates may compromise exact replication. Fourth, simplified model: physical contact, hormonal synchrony, and extended temporal dynamics are absent. Fifth, prompt length confound: the DVP-Guided strategy's prompt is substantially longer than Raw-LLM's (~80 vs. ~15 words). We cannot rule out that the observed difference partially reflects prompt length rather than developmental psychology content per se. Sixth, missing human validation: while we show correspondence with theoretical predictions, direct comparison with human-coded data is essential.

5.2 Assessment Circularity

We recognize a definitional dependency: the DVP-Guided strategy was explicitly designed to implement co-regulation, serve-and-return, and developmental appropriateness -- precisely the constructs our assessment measures. This is not a measurement flaw but a deliberate design choice: our goal was to test whether LLM agents, when instructed in established care-giving principles, produce measurably different outcomes. The meaningful comparison is DVP vs. Raw-LLM (implicit pretrained knowledge) and vs. Random (no knowledge). The progressive improvement across Random < Raw-LLM < DVP-Guided demonstrates that explicit theoretical guidance adds value beyond implicit knowledge, which is the central claim of this work.

6 Conclusion

BabySim-Agent demonstrates that LLM multi-agent systems can serve as low-cost, rapid prototyping tools for exploratory hypothesis testing in developmental psychology. Our results show that a developmentally-informed care-giving strategy produces measurably superior interaction outcomes across theoretically meaningful dimensions (all $p < .001$), with effect sizes representing upper-bound estimates under controlled simulation conditions. The DVP-Guided strategy showed particularly strong advantages for difficult-temperament infants. While fundamental limitations preclude strong claims about infant psychological simulation, the framework offers a promising intermediate-layer tool for computational developmental research.

References

1. Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). Patterns of attachment: A psychological study of the strange situation. Erlbaum. https://doi.org/10.4324/9781315802428

2. Tronick, E. (2007). The neurobehavioral and social-emotional development of infants and children. W.W. Norton. DOI:10.1111/j.1468-5922.2007.00705_2.x
3. Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from Homo Silicus? NBER Working Paper No. 31122. https://www.nber.org/papers/w31122
4. Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, T. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337-371. https://doi.org/10.1017/pan.2023.2
5. Feng, S., Ding, W., Liu, A., Wang, Z., Shi, W., Wang, Y., et al. (2025). When one LLM drools, multi-LLM collaboration rules. arXiv preprint. https://doi.org/10.48550/arXiv.2502.04506
6. Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. Proceedings of UIST, 1-22. https://doi.org/10.48550/arXiv.2304.03442
7. Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., et al. (2023). S3: Social-network simulation system with large language model-empowered agents. arXiv:2307.14984. https://doi.org/10.48550/arXiv.2307.14984
8. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? Computational Linguistics, 50(1), 237-291. https://aclanthology.org/2024.cl-1.8/
9. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT, 610-623. https://doi.org/10.1145/3442188.3445922
10. Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental influences. Psychological Bulletin, 135(6), 885-908. https://doi.org/10.1037/a0017376
11. Bowlby, J. (1969). Attachment and loss: Vol. 1. Attachment. Basic Books.
12. Cassidy, J., & Shaver, P. R. (Eds.). (2018). Handbook of attachment: Theory, research, and clinical applications (3rd ed.). Guilford Press.
13. Wass, S. V., Noreika, V., Georgieva, S., Clackson, K., Ball, G., & Leong, V. (2024). Parental neural scaffolding: A mechanism for dyadic co-regulation in caregiver-infant interactions. Developmental Cognitive Neuroscience, 66, 101445. https://doi.org/10.1371/journal.pbio.2006328
14. Tronick, E., & Reck, C. (2009). Infants of depressed mothers. Harvard Review of Psychiatry, 17(2), 147-156. https://doi.org/10.1080/10673220902899714
15. DiCorcia, J. A., & Tronick, E. (2011). Quotidian resilience: Exploring mechanisms that drive resilience from a perspective of everyday stress and coping. Neuroscience & Biobehavioral Reviews, 35(7), 1593-1602. https://doi.org/10.1016/j.neubiorev.2011.04.008
16. Chris L. Porter, Chongming Yang, Nathan A. Jorgensen, Cortney Evans-Stout, Development of mother-infant co-regulation: The role of infant vagal tone and temperament at 6, 9, and 12 months of age, Infant Behavior and Development, Volume 67, 2022,101708, ISSN 0163-6383, https://doi.org/10.1016/j.infbeh.2022.101708.

17. Rothbart, M. K. (2004). Infant behavior questionnaire-revised. Bowdoin College. https://research.bowdoin.edu/rothbart-temperament-questionnaires/instrument-descriptions/the-infant-behavior-questionnaire/
18. Petters, D. (2017). A new direction for attachment modelling: Simulating Q-set descriptors. Proceedings of the 15th International Conference on Cognitive Modelling (ICCM). https://acs.ist.psu.edu/papers/ICCM2017Proceedings.pdf
19. Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. Proceedings of the National Academy of Sciences, 120(13), e2215907120. https://doi.org/10.1073/pnas.2215907120