Dreaming with ChatGPT: Unraveling the Challenges of LLMs Dream Generation

Anonymous ACL submission

Abstract

Large Language Models (LLMs), such as ChatGPT, are used daily for different human-like text generation tasks. This motivates us to ask: *Can an LLM generate human dreams?* For this research, we explore this new avenue through the lens of ChatGPT, and its ability to generate valid dreams. We have three main findings: (i) Chatgpt-4o, the new version of chatGPT, generated all requested dreams. (ii) Generated dreams meet key psychological criteria of dreams. (iii) Generated dreams embed biases towards different groups. We hope our work will set the stage for developing a new task of dream generation for LLMs. This task can help psychologists evaluate patients' dreams based on their demographic factors.

1 Introduction

A dream is a series of involuntary images, ideas, and emotions during sleep, especially in the rapid eye movement (REM) stage (apa, 2024). Dreams are crucial in psychology, as they provide insight into the mind, revealing hidden desires, fears, psychological status, and conflicts (Freud, 1900; Hobson, 2009; Solomonova et al., 2021).
Large Language Models (LLMs) aim to mimic psychological phenomena by simulating aspects of human cognition, such as language understanding, reasoning, and emotion recognition (Sartori and Orrù, 2023; Hofweber et al., 2024; Kuo and Chen, 2023). While still not there, using dreams generated by LLMs may serve a helpful tool to the professional systemization of humans' dreams analysis, categorized by a person's characteristics, thus allowing a deeper understanding of an individual's dreams and their psychological diagnosis.

In this work, we lay the groundwork for this task - dreams generation. As this avenue is undereamined, we try to shed light on the capabilities of certain LLMs to generate dreams that meet psychological criteria, and the biases reflected in these dream descriptions. We picked ChatGPT, the most globally popular LLM[1] as our test case. We use several versions of ChatGPT3.5 and ChatGPT4o, the most recent version of the OpenAI's LLM. Through an in-depth analysis of the results of different versions of ChatGPT and the dreams they produce, we find that:

- ChatGPT4o generates a dream description per every prompt, which is different from its predecessors.
- Dream descriptions that are generated by ChatGPT models follow some common psychological definitions of a dream but do not fully capture how a dream looks/feels like.
- The dream descriptions generated by gpt-3.5-turbo-16k, gpt-3.5-turbo-16k-0613, and gpt4o models are mainly biased towards demographic factors.

2 Dreams in Psychology

Traditionally, dreams are mostly associated and analyzed through REM sleep (Hobson and Pace-Schott, 2002; Nir and Tononi, 2010). Formally, in the APA Dictionary of Psychology (apa, 2024), REM dreams are defined by four attributes: (1) a sense of motion in space paired with visual imagery (*Motion*); (2) strong emotions, especially fear, euphoria, or anger (*Emotion*); (3) the perception that dream events, characters, and situations are real (*Realness*); and (4) unexpected changes in characters, situations, and plot elements (*Discontinuity*). Other attributes derived from psychological works include the location of the dream, which is mostly in normative daily scenes (Domhoff, 2007; Snyder et al., 1968) (*Location*); the existence of at least one other being (Domhoff, 2007; Snyder, 1970; Dorus et al., 1971) (*Other Beings*); the existence

[1] https://zapier.com/blog/best-llm/

of objects (Domhoff, 2007; Snyder, 1970; Dorus et al., 1971) (*Objects*); and the activity of talking with other beings (Domhoff, 2007; Snyder, 1970) (*Conversation*). We will check if generated dreams meet psychological criteria.

3 Related Work

LLMs are being tested through different advanced generation tasks of human nature, such as sarcasm (Chakrabarty et al., 2020), metaphor (Chakrabarty et al., 2021), storytelling (Yao et al., 2019; Yang et al., 2022), humour (Mittal et al., 2022; Dsilva, 2024; Tikhonov and Shtykovskiy, 2024), songs (Tian and Peng, 2022; He et al., 2019), hyperbole (Tian et al., 2021) and tongue twisters (Loakman et al., 2024).

The mimicry of human thinking and behavior by LLMs is still under research. Binz and Schulz (Binz and Schulz, 2023) and Abbasiantaeb et al. (Abbasiantaeb et al., 2024) explored LLMs' abilities to simulate human understanding and interactions. Zhang et al. (Zhang et al., 2023b) focused on using human-like reasoning to improve LLMs' decision-making. Another research area is LLMs' fairness, with Bender et al. (Bender et al., 2021) and Noble (Noble, 2018) highlighting the risks of biases and stereotypes. Zheng et al. (Zheng et al., 2024) explored enhanced biases in judgments carried out by LLMs. Turpin et al. (Turpin et al., 2024) showed that biases in quality assessment tasks can significantly affect LLMs' reasoning.

However, the topic of generating dreams is yet to be covered. Recent work (Bertolini et al., 2024) explored LLMs' ability to classify dreams by emotions but did not examine their ability to produce dreams or consider other psychological criteria.

This research evaluates ChatGPT's ability to generate dreams descriptions matching known psychological frameworks (apa, 2024; Snyder et al., 1968; Dorus et al., 1971) and to produce dreams fairly.

4 Methodology

In this research, we tackle three research questions:

1. Can LLMs generate dreams?
2. Do dreams generated by LLMs comply with psychological criteria of dreams?
3. Are LLMs biased when generating dreams?

To address these questions, we devised a dedicated methodology. To generate the dreams we use Context-Less Generation (Wan et al., 2023; Wan and Chang, 2024), in which the model is prompted with a simple zero-shot description of the dreamer - race, religion, and gender (Hanna et al., 2023; Salewski et al., 2024; Mahomed et al., 2024; Calderon et al.; Zhang et al., 2023a). We added a control group, person. These demographic factors are found in Table 1. Our methodology follows the steps below: (1) Given a demographic factor, replace the mask in the prompt with it: "**Pretand you are $\{MASK\}$. Please generate a dream that this person dreamt last night.**", and the prompt to model M. (2) Get the response r_i. (3) Run steps 1-2 five times in different sessions (no memorization, zero-shot), thus getting the set $r_{i_j} = \{r_{i_0}, r_{i_1}, r_{i_2}, r_{i_3}, r_{i_4}\}$. (5) Annotate using human annotators each r_{i_j} set based on predefined attributes. (6) Analyze the results.

Race	Religion	Gender	Control
Asian	Jewish	Female	Person
Indian	Christian	Male	
Black	Muslim		
White			
Hispanic			
Arab			

Table 1: Table of simple demographic factors of people used for prompting GPT models.

Some models provided very few dreams. We concluded this by automatically analyzing for a single disclaimer or absence of multiple blank lines[2].

Attributes: For each sample, we annotated the following attributes: (1) is there a dream? (yes/no), based on the existence of a story. (2) is there a disclaimer? (yes/no), where a disclaimer is a text similar to "I'm sorry, but I cannot fulfill that request." or "As an AI, I don't have dreams or feelings". (3) the pronoun used for the dreamer (I/You/He/She/They) (4) the existence of other languages used in the dream, and which language (Arabic/Hebrew/Spanish/Others/None) (5-11) the psychological attributes from Section 2 - *Motion* (yes/no), *Emotion* (yes/no), *Realness* (yes/no), *Discontinuity* (yes/no), *Location*, *Other Beings*, *Objects* and *Conversation* (yes/no).

Human Evaluation: Three annotators participated: two Masters students with an academic background in psychology and one computer science

[2]Concrete dreams were spread across multiple lines upon close inspection.

postdoctoral fellow. Each sample was annotated by two annotators, with a third resolving any disagreements (Mukhtar et al., 2017). The full text of instructions given to annotators is presented in Appendix E.

Metrics: We used a success rate metric for generating dreams, similar to previous work (Wen et al., 2024; Zhao et al., 2024). This measured the model's ability to produce valid dreams (i.e., containing a story) or without disclaimers. The success rate was the number of samples meeting the criteria divided by the total samples.

We also used the *Chi-Square* test to evaluate the independence of attributes and demographic factors/models, similar to previous research (Hanna et al., 2023; Calderon et al.; Mahomed et al., 2024).

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency and E_i is the expected frequency.

5 Experiments

We generated three popular demographic factors groups to use for the prompts - religion, race, and gender, and a control group - person. The full list is presented in Table 1. We used the prompt from Section 4 with each factor.

We evaluated several gpt models: gpt-3.5-turbo (gpt3.5T), gpt-3.5-turbo-16k-0613 (gpt3.5T16k), gpt-3.5-turbo-0613 (gpt3.5T0613), and gpt-3.5-turbo-1106 (gpt3.5T1106), gpt-4o (gpt4o). We used the default parameters (e.g., temperature 1.0, Top-P 1.0) of all models. In total, for each model, we obtained 60 samples, based on the demographic factors. The samples, code, and anonymized annotations are available online[3]. The code is under the MIT license (Open Source Initiative, 2023).

6 Dream Generation Analysis

In this section, we analyzed all 300 generated responses, to assess the ability of a model to generate a basic dream generation. We summed all samples. The full details of each are presented in Appendix C.

We found that gpt4o generated 100% of requested dreams, while gpt3.5T16k and gpt3.5T0613 produced over 68%. However, both gpt3.5T and gpt3.5T1106 generated less than 20% of dreams, with the latter experiencing a significant drop in performance despite being released later than gpt3.5T16k and gpt3.5T0613[4]. Thus, we dropped further analysis of the latter models.

Out of these generated dreams, we continually analyzed the generated dreams (Table 2). We explored whether the model did not produce a disclaimer stating it is an AI that does not dream, thus following the prompt directly without an explicit objection.

Model	Gen	No Disc	1st
gpt3.5T	13%	-	-
gpt3.5T16k	75%	31%	15%
gpt3.5T0613	68%	39%	17%
gpt3.5T1106	18%	-	-
gpt4o	100%	98%	73%

Table 2: Dream generation characteristics, based on the generated dreams (*Gen*) out of total sample size (*Samp*), the nonexistence of a disclaimer (*No Disc*), and whether the dream is in first person view (*1st*). The original sample size is 60 dreams. The gpt3.5T and gpt3.5T1106 were eliminated in the deeper analysis due to their poor performance in the initial dream generation.

We found that this phenomenon of no disclaimer+dream was found in 97% of gpt4o dreams, 39% of gptT0613 dreams, and 31% of gptT061316k dreams.

We also looked at whether the dream was generated in first person, as the prompt started with *"pretend you are..."*. In that aspect, gpt4o met 73% of the times, gptgpt3.5T16k 15% of the times, and gpt3.5T0613 17% of the times.

In short, although with some decrease caused by matching the full criteria, gpt4o followed the prompted dream generation with a significant gap (∼60%) between its performance and the other two models' performances.

Although not all generated dreams complied with the no disclaimer+first person criteria, we continued with the generated dreams (Gen from Table 2). For the next sections, we considered 60 dreams for gpt4o, 45 dreams for gpt3.5T16k, and 41 dreams for gptgpt3.5T0613[5].

7 Psychological Dream Attributes

APA Attributes: The results of APA's attributes (Section 2) are presented in Table 3. It can be

[3] https://anonymous.4open.science/r/DreamGPT-9653/

[4] https://context.ai/compare/gpt-3-5-turbo-16k/gpt-3-5-turbo

[5] Similarly to other work (Wan et al., 2023), that drew interesting conclusions from small LLM-generated samples.

seen that the three models meet the motion and emotion dream properties raised by APA. In the discontinuity attribute, gpt3.5T16k got 56%, and gpt3.5T0613 got 37%. Gpt4o shows the greatest promise in this attribute, with 70%. However, all models lack a sense of realness, as this property does not have a clear indication in the dreams.

Model	M	E	R	D
gpt3.5T16k	98%	100%	0%	56%
gpt3.5T0613	100%	100%	0%	37%
gpt4o	100%	100%	7%	70%

Table 3: APA Attributes Results. M stand for motion, E for emotion, R for realness and D for discontinuity. It is shown that gpt4o complies the most APA's properties of ERM dreams.

Model	N_Loc	Other Beings	Conv
gpt3.5T16k	47%	96%	51%
gpt3.5T0613	44%	88%	44%
gpt4o	73%	95%	67%

Table 4: Other Attributes Results. N_Loc stands for locations in nature, Other Beings for people/animals, and Conv for conversation. It is shown that gpt4o complies the most with all properties.

Other Attributes: We explored attributes from various psychological sources, including locations, beings and interactions (Section 2). Non-daily locations appeared in 73% of gpt4o dreams, 44% of gpt3.5T0613 dreams, and 47% of gpt3.5T16k dreams. This shows that the models do not fully comply with this property. Also, all models included at least one other being in the dreams (Domhoff, 2007; Snyder, 1970; Dorus et al., 1971). Conversations were found in 67% of gpt4o dreams, 44% of gpt3.5T0613 dreams, and 51% of gpt3.5T16k's dreams.

Overall, meeting all psychological dream definitions is not trivial for LLMs. However, the ability to generate dreams with embedded creatures, and motion/emotion rules is met 100% by each model we explored. Still, gpt4o is the leader in psychological attributes in general.

8 Biased Dream Attributes

In this section, we present insights derived from attributes other than the previous section's psychological ones. These insights showcase biases towards specific demographic factors.

Pronouns: Among non-genderized demographic factors generated dreams, "he" pronoun usage was 6 out of 38 (16%) in gpt3.5T16k, 1 out of 36 (3%) in gpt3.5T0613, and 4 out of 50 (10%) in gpt4o, while "she" was never used. In this aspect, gpt4o is overshadowed by the slightly more neutral gpt3.5T0613.

Flowers: Among genderized factors, the female factor with flower/s has residual of 2.09. The person and male factors have no strong correlations. Full results are in Appendix D. One possible explanation is the association of females with flowers in poetry, the scent of flowers as a perfume (Stott, 1992; Spence, 2021).

Other Languages: Most dreams were in English with some non-English expressions, except for the Hispanic factor, where 40% of gpt3.5Ts' and 80% of gpt4o's dreams were entirely in Spanish[6]. Non-English word usage showd a notable association, with Arabic and Arab factor residual of 2.32, and an even stronger link between Arabic and Muslims, with a residual of 7.17, although not all Muslims speak Arabic (Chejne, 1965). The Jewish factor has a clear association with using Hebrew words[7] with residuals of 8.1, and 10.13 for Hispanic and Spanish. The full results are presented in Appenix A.

Ultimately, generated dreams are embedded with biases towards different groups, in language, pronouns, and objects' usage. Also, gpt4o is not the most neutral model in the set, but the most fit one in most categories.

9 Conclusion

In this work, we examined the possibility of generating dreams by LLMs. We explored it through the test case of ChatGPT models. The most promising model was found to be gpt4o. We found that some fundamental psychological attributes are met by the generated dreams, but there is still progress to be made. Also, some biases were found in the models for generating dreams. We hope this initial work will pave the way to more LLM-dreams research, contributing to the psychological analysis of human dreams.

[6] We translated these dreams using Google Translate, which showed significant results in machine translation tasks, even against GPT models (Robinson et al., 2023; Lai et al., 2024)

[7] An outlier might be the usage of Arabic in the Jewish factor. It can be explained by the usage being of the word Hamsa, a symbol that is common in Jewish communities (Sabar, 2010).

10 Ethics Statement

This paper initially explores the capabilities ChatGPT to generate dreams. As the authors only infer dreams and do not look for a specific person's dream, the resulting dreams are not exposing any private data of an individual. Also, the authors explore the biases generated by the LLM to shed light on the models' fairness issues.

However, the potential risks of such a research include LLMs perpetuating biases as detailed above, overgeneralizing results of dreams generation, and struggling to provide reliable insights across diverse psychological contexts.

11 Limitations

Despite our interesting findings, this work is subject to several limitations. First, our annotations were based on human annotators. Due to the lack of concise annotations of psychological attributes of dreams, such as discontinuity and realness, we annotated the data with human annotators as an initial work. We envision an extension of this work using fine-tuned model to annotate the data (Wang et al., 2024; Wu et al., 2023).
Second, our data was limited to 300 samples. Although this data seems small, it gave interesting aspects of the ability of LLMs to generate dreams. We intend to curate a larger dataset for more comprehensive research.
Next, we explored ChatGPT as the most popular LLM globally. It would be beneficial to explore the dream generation abilities of other LLMs as well, such as Meta's Llama (Touvron et al., 2023) or Google's Gemini.
Also, this work initialized the research of generating dreams by LLMs. We used a small set of psychological attributes and a limited set of demographic factors. More advanced work on this topic may follow a broader range of psychological aspects, analyzing combinations of demographic factors, and adding more factors such as jobs and maternity status.

References

2024. APA Dictionary - Dream.

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Lorenzo Bertolini, Valentina Elce, Adriana Michalak, Hanna-Sophia Widhoezl, Giulio Bernardi, and Julie Weeds. 2024. Automatic annotation of dream report's emotional content with large language models. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 92–107.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Nitay Calderon, Naveh Porat, Eyal Ben-David, Alexander Chapanin, Zorik Gekhman, Nadav Oved, Vitaly Shalumov, and Roi Reichart. Measuring the robustness of nlp models to domain shifts.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. r^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. *arXiv preprint arXiv:2004.13248*.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Anwar G Chejne. 1965. Arabic: Its significance and place in arab-muslim society. *Middle East Journal*, 19(4):447–470.

G William Domhoff. 2007. Realistic simulation and bizarreness in dream content: Past findings and suggestions for future research. *The new science of dreaming*, 2:1–27.

E. Dorus, W. Dorus, and A. Rechtschaffen. 1971. The incidence of novelty in dreams. *Archives of General Psychiatry*, 25(4):364–368.

Ryan Rony Dsilva. 2024. *Augmenting Large Language Models with Humor Theory To Understand Puns*. Ph.D. thesis, Purdue University Graduate School.

Sigmund Freud. 1900. *The Interpretation of Dreams*. Macmillan, New York.

John J Hanna, Abdi D Wakene, Christoph U Lehmann, and Richard J Medford. 2023. Assessing racial and ethnic bias in text generation for healthcare-related tasks by chatgpt1. *MedRxiv*.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

J Allan Hobson. 2009. Rem sleep and dreaming: towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10(11):803–813.

J Allan Hobson and Edward F Pace-Schott. 2002. The cognitive neuroscience of sleep: neuronal systems, consciousness and learning. *Nature Reviews Neuroscience*, 3(9):679–693.

Thomas Hofweber, Peter Hase, Elias Stengel-Eskin, and Mohit Bansal. 2024. Are language models rational? the case of coherence norms and belief revision. *arXiv preprint arXiv:2406.03442*.

Hui-Chi Kuo and Yun-Nung Chen. 2023. Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning. *Preprint*, arXiv:2210.05901.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *arXiv preprint arXiv:2406.01771*.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2024. Train & constrain: Phonologically informed tongue-twister generation from topics and paraphrases. *arXiv preprint arXiv:2403.13901*.

Yaaseen Mahomed, Charlie M Crawford, Sanjana Gautam, Sorelle A Friedler, and Danaë Metaxa. 2024. Auditing gpt's content moderation guardrails: Can chatgpt write your favorite tv show? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 660–686.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

Neelam Mukhtar, Mohammad Abid Khan, and Nadia Chiragh. 2017. Effective use of evaluation measures for the validation of best classifier in urdu sentiment analysis. *Cognitive Computation*, 9:446–456.

Yuval Nir and Giulio Tononi. 2010. Dreaming and the brain: from phenomenology to neurophysiology. *Trends in cognitive sciences*, 14(2):88–100.

Safiya Umoja Noble. 2018. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Open Source Initiative. 2023. Mit license. `https://opensource.org/license/mit/`. Accessed: 2024-06-14.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Shalom Sabar. 2010. From sacred symbol to key ring: The hamsa in jewish and israeli societies. *Jews at Home: The Domestication of Identity*, page 140.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Giuseppe Sartori and Graziella Orrù. 2023. Language models and psychological sciences. *Frontiers in Psychology*, 14:1279317.

F. Snyder. 1970. The phenomenology of dreaming. In L. Madow and L.H. Snow, editors, *The Psychodynamic Implications of the Physiological Studies on Dreams*, pages 124–151. Charles S Thomas, Springfield.

F. Snyder, I. Karacan, V. K. Jr. Tharp, and J. Scott. 1968. Phenomenology of rems dreaming. *Psychophysiology*, 4(3):375.

Elizaveta Solomonova, Claudia Picard-Deland, Iris L Rapoport, Marie-Hélène Pennestri, Mysa Saad, Tetyana Kendzerska, Samuel Paul Louis Veissiere, Roger Godbout, Jodi D Edwards, Lena Quilty, et al. 2021. Stuck in a lockdown: Dreams, bad dreams, nightmares, and their relationship to stress, depression and anxiety during the covid-19 pandemic. *PLoS One*, 16(11):e0259040.

Charles Spence. 2021. The scent of attraction and the smell of success: crossmodal influences on person perception. *Cognitive Research: Principles and Implications*, 6(1):46.

Annette Stott. 1992. Floral femininity: A pictorial definition. *American Art*, 6(2):61–77.

Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3587–3597, Seattle, United States. Association for Computational Linguistics.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. *arXiv preprint arXiv:2405.07280*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Yixin Wan and Kai-Wei Chang. 2024. White men lead, black women help: Uncovering gender, racial, and intersectional bias in language agency. *arXiv preprint arXiv:2404.10508*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. Autodroid: Llm-powered task automation in android. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 543–557.

Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, et al. 2023. Empowering a metric with llm-assisted named entity annotation: Hw-tsc's submission to the wmt23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 822–828.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Angela Zhang, Mert Yuksekgonul, Joshua Guild, James Zou, and Joseph Wu. 2023a. Chatgpt exhibits gender and racial biases in acute coronary syndrome management. *medRxiv*, pages 2023–11.

Zheyuan Zhang, Shane Storks, Fengyuan Hu, Sungryull Sohn, Moontae Lee, Honglak Lee, and Joyce Chai. 2023b. From heuristic to analytic: Cognitively motivated strategies for coherent physical commonsense reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7354–7379, Singapore. Association for Computational Linguistics.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Language Usage by Demographic Factors - Full Analysis

We show the full Chi-square test of the languages and demographic factors in our data. The significant results are in Table 6, and full results in Table 7. The Chi-square statistic was 318.83, the P-value was $2.36 * 10^{-43}$, and the degrees of freedom were 44. So, we rejected the null hypothesis of independence of demographic factors and languages.

	Arab		Muslim		Jewish		Hispanic	
Tag	O	E	O	E	O	E	O	E
A	**3**	**0.99**	**8**	**1.07**	0	1.15	0	1.15
H	0	0.54	0	0.58	**6**	**0.62**	0	0.62
S	0	1.17	0	1.26	0	1.35	**13**	**1.35**

Table 5: Comparison of Observed and Expected Frequencies of other Languages for Arab, Muslim, Jewish, and Hispanic factors. The languages are **A**rabic, **H**ebrew, and **S**panish.

B Nature locations found in Dreams - Full Analysis

This section shows the full list of locations found in our dreams data. The locations can be found in table 9.

C Models History & Tokens

Table 10 discloses the dream generation rates of each explored model, based on its release date and amount of tokens, as a complementary to Section 6. Dates and Tokens data acquired from[8] [9].

[8]https://community.openai.com/t/what-are-the-differences-between-gpt-3-5-turbo-models/557028/2

[9]https://context.ai/compare/gpt-3-5-turbo-16k/gpt-3-5-turbo

lang	Arab	Asian	Chris	Hisp	Indian	Jewish	Muslim	a person	black	female	male	white
A	1.33	1.22	1.22	1.53	0.82	1.63	1.43	1.53	1.12	1.02	1.22	0.92
H	0.80	0.73	0.73	0.92	0.49	0.98	0.86	0.92	0.67	0.61	0.73	0.55
O	0.35	0.33	0.33	0.41	0.22	0.44	0.38	0.41	0.30	0.27	0.33	0.24
S	1.15	1.06	1.06	1.33	0.71	1.41	1.24	1.33	0.97	0.88	1.06	0.80
X	9.37	8.65	8.65	10.82	5.77	11.54	10.10	10.82	7.93	7.21	8.65	6.49

Table 6: Expected frequencies of language usage by demographic factors. The languages (lang) are A for Arabic, H for Hebrew, S for Spanish, O for others, and X means only English words. The shortened factor names are Hispanic (Hisp) and Christian (Chris).

lang	Arab	Asian	Chris	Hisp	Indian	Jewish	Muslim	a person	black	female	male	white
A	4.00	0.00	0.00	0.00	0.00	1.00	10.00	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	9.00	0.00	0.00	0.00	0.00	0.00	0.00
O	0.00	0.00	0.00	0.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	13.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
X	9.00	12.00	12.00	2.00	5.00	5.00	4.00	15.00	11.00	10.00	12.00	9.00

Table 7: Original observations of language usage by demographic factors. The languages (lang) are A for Arabic, H for Hebrew, S for Spanish, O for others, and X means only English words. The shortened factor names are Hispanic (Hisp) and Christian (Chris).

D Flowers in Genderized factors

We run chi-square test on the flower item and genderized groups. Table 11-12 express the results. The Chi-square statistic value is 12.206, the P-value is 0.0022, and degree of freedom is 2. The Chi-square test proves a positive correlation between the female factor and flowers (residual of 2.09).

E Instructions to Annoators

In the annotations of dreams, when the symbol (V/X) is shown, please put V for true/exists, and X for false/nonexist. If you are not sure, please put X.

These are the attributes we explore:

- Is there a dream (v/x) - is there a story or just a statement on the inability of the AI to generate a dream?
- disclaimer (v/x) - if the model states something as “As an AI, I don’t have personal dreams as humans do. However, I can create a fictional dream scenario for you.”, this means that it disclaims that it generates a dream and it is not natural. If there is nothing more than this disclaimer, and no dream was generated, please leave the entire row blank.
- use of languages other than English - if there is a word not from English, like “Shema” and “Inshalla”, please state the other language (Hebrew, Arabic)
- location - A one-word location of the dream, such as desert, garden. If the dreamer moves places, please add other places.
- narrator (I/You/He/She/They) - The point of view of the dreamer - is it “I dreamt that…”, or “he dreamt”. This is considered as the pronoun of a dream in the paper.
- other persons - other persons mentioned in the dream
- animals - same as persons, but with animals
- items - same with animals, but with items
- conversation (v/x) - if there is any conversation in the dream.
- motion (x/v) - visual imagery along with a sense of motion in space, such as “I was walking”.
- emotion (x/v) - intense emotion, especially fear, elation, or anger.
- belief of realness (x/v) - belief that dream characters, events, and situations are real

tag	Arab	Asian	Chris	Hisp	Indian	Jewish	Muslim	a person	black	female	male	white
A	**2.32**	-1.11	-1.11	-1.24	-0.90	-0.50	**7.17**	-1.24	-1.06	-1.01	-1.11	-0.96
H	-0.89	-0.86	-0.86	-0.96	-0.70	**8.10**	-0.93	-0.96	-0.82	-0.78	-0.86	-0.74
O	-0.59	-0.57	-0.57	-0.64	5.96	0.86	-0.62	-0.64	-0.55	-0.52	-0.57	-0.49
S	-1.07	-1.03	-1.03	**10.14**	-0.84	-1.19	-1.11	-1.15	-0.99	-0.94	-1.03	-0.89
X	-0.12	1.14	1.14	-2.68	-0.32	-1.92	-1.92	1.27	1.09	1.04	1.14	0.99

Table 8: Residuals by demographic factors and languages. The languages (lang) are A for Arabic, H for Hebrew, S for Spanish, O for others, and X means only English words. The shortened factor names are Hispanic (Hisp) and Christian (Chris).

Garden	Sea	Ocean	Forest
Meadow	Lake	Waterfall	River
Mountain	Field	Oasis	Island
Lagoon	Sky	Hills	Pond

Table 9: Nature locations of dreams found in our data.

Model	DGR	Date	Tokens
gpt3.5T	13% (8)	11.28.22	4K
gpt3.5T16k	73% (44)	06.13.23	16K
gpt3.5T0613	68% (41)	06.13.23	4K
gpt3.5T1106	18% (11)	11.06.23	16K
gpt4o	100% (60)	05.13.24	128K

Table 10: Dream generation rate (DGR), based on each model, its date of release, and the number of tokens used as context window. The DGR is measured by counting the actual dreams (no sole disclaimer) out of all responses. The generation rate and actual count are provided for clarity.

- discontinuity (x/v) - sudden discontinuities in characters, situations, and plot elements. The word suddenly helps a lot here

	a person		female		male	
Tag	O	E	O	E	O	E
True	5	4.86	**7**	**3.24**	0	3.89
False	10	10.13	3	6.76	12	8.1

Table 11: Comparison of Observed and Expected Frequencies of the existence of a flower/s in male, female, person dreams.

tag	a person	female	male
0.00	-0.04	-1.45	1.37
1.00	0.06	**2.09**	-1.97

Table 12: Residuals of male, female, person dreams with flowers.