Do LLMs Know Spring is Green? A Synesthesia Study of LLMs Response

Anonymous ACL submission

Abstract

Large Language Models (LLMs) exhibit emergent capabilities beyond core language tasks, and demonstrate certain cognitive alignment between humans. However, the potential of LLMs to simulate human-like cross-modal cognitive alignments, such as synesthesia, remains unexplored. Synesthesia involves consistent associations between concepts and sensory experiences (e.g., linking numbers to colors), a phenomenon also reflected in cross-modal correspondences observed even in non-synesthetes. In this work, we conduct the study of whether modern LLMs replicate such synesthetic alignment by evaluating their responses on color association tasks across diverse conceptual domains: digits, letters, temporal concepts (e.g., days, months), spatial directions, and abstract entities. Using standardized prompts, we analyze responses from multiple LLMs and compare them to human data collected from 260 participants. Colors are mapped to a perceptually uniform space (CIELAB), with alignment quantified via the CIEDE2000 metric. Our results reveal that LLMs show significant alignment with human consensual patterns, particularly for temporal concepts like seasons and months, achieving color differences comparable to human variability. However, abstract concepts (e.g., directions) exhibit greater divergence. Cultural influences (e.g., Western vs. Chinese contexts) impact alignment, while gender differences in humans do not translate to LLMs. Model size and architecture also affect performance, with larger models demonstrating stronger alignment. These findings highlight LLMs' ability to capture certain cross-modal associations, offering insights into their implicit grounding of abstract concepts and implications for multimodal applications requiring sensory-conceptual integration.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable generalization capabilities beyond core language tasks (Floridi and Chiriatti, 2020; Touvron et al., 2023; Liu et al., 2024), hinting at emergent competencies in domains like color perception (Pi et al., 2024; Abdou et al., 2021), spatial reasoning (Chen et al., 2024; Fu et al., 2024), and orientation understanding (Yang et al., 2025; Stogiannidis et al., 2025). Despite being trained only on textual data, recent studies suggest that LLM representations implicitly capture aspects of grounded concepts (Pavlick, 2023; Harnad, 2024).

However, human cognition contains more complex abilities, such as the phenomenon of synesthesia (Ward, 2013), which is a perceptual crossing of senses or concepts (Hubbard, 2007; Spector and Maurer, 2013). In synesthesia, stimulation of one cognitive pathway triggers involuntary experiences in another (Grossenbacher and Lovelace, 2001; Rich and Mattingley, 2002). For instance, some individuals consistently perceive specific colors when thinking of particular numbers or letters (known as grapheme–color synesthesia) (Cytowic and Eagleman, 2011; Parise, 2016). More broadly, even non-synesthetes exhibit cross-modal associations (Spence, 2011) such as the famous Bouba–Kiki effect (Ramachandran and Hubbard, 2001), where people intuitively match nonsense words like "bouba" with round shapes and "kiki" with spiky shapes (Maurer et al., 2006). These cross-modal correspondences appear to reflect fundamental alignments in human cognitive system (Boroditsky et al., 2009; Spence, 2011), which are widespread and often consistent across different people (Marks, 1987; Lacey et al., 2021), yet can also show individual or cultural variations (Spence and Deroy, 2012).

Given the above findings, an intriguing research question arises: can modern LLMs simulate synesthesia-like cognitive alignments across modalities? Addressing this question is crucial for practical applications of LLMs. For example, when tasked with designing a promotional poster for spring travel, an LLM should ideally select appropriate colors (e.g., green). However, if the model lacks synesthetic cognitive capabilities, it may fail to produce aesthetically appealing results. Prior research has extensively probed LLMs' abilities on basic perceptual (Yuksekgonul et al., 2022) and cognitive tasks (Misra et al., 2021). However, it remains underexplored whether LLMs possess any synesthesia alignment abilities. This leaves a significant research gap: we lack a clear understanding of whether advanced language models exhibit any form of synesthetic alignment akin to humans, and what implications this might have for their integration into multimodal tasks.

Our work aims to fill this gap by conducting the first systematic study of synesthesia alignment in LLMs. We design a suite of synesthesic color association tasks inspired by human synesthesia, covering a diverse range of conceptual domains, including digits (0–9), letters (A–Z), temporal concepts (e.g. days of week, months), cardinal directions (e.g. north, south, etc.), and other spatial or abstract concepts, and ask LLMs to link each concept with a color. Using a standardized questionnaire-style prompting, we evaluate multiple mainstream LLMs on these tasks, capturing their preferred "color" for each concept. To benchmark performance, we also collected a dataset of 260 human participants' responses on the same tasks, allowing direct comparison between model-generated associations and human intuition. For rigorous analysis, we represent colors in a perceptually uniform color space and quantify differences using the CIEDE2000 color difference metric (Sharma et al., 2005), ensuring an objective measure of how closely an LLM's color-choice aligns with human consensual patterns. This experimental design enables us to probe whether LLMs can align abstract concepts with concrete sensory dimensions in a manner comparable to human cross-modal cognition. The results indicate that while there are notable differences between human participants and LLMs in synesthetic responses to abstract concepts, alignment is observed in temporal concepts, such as seasons and months. Additionally, individual identity differences, such as gender and culture, exhibit varying impacts on synesthetic performance. Gender differences do not affect synesthesia of LLMs, whereas cultural influences are observed. Differences caused by model type and sizes are also observed.

In summary, our contributions are as follows:

- **First exploration of synesthetic alignment in LLMs**: We present the first systematic investigation into whether LLMs exhibit synesthesia-like cross-modal cognitive alignments, a previously unexamined aspect of their capabilities.

- **Synesthesia task and human comparison**: We introduce a synesthesia evaluation framework consisting of color-association tasks across multiple concept domains (numbers, letters, time, directions, space), and we gather responses from both several trending LLMs and 260 human subjects for comprehensive comparison.

- **Objective alignment analysis**: We propose an evaluation method using the CIEDE2000 color difference standard to quantify the alignment between model-generated color associations and human synesthetic patterns, ensuring results are measured with perceptual accuracy.

2 Related Work

2.1 Perceptual and Cognitive Tasks of LLMs

Despite being trained only on textual data, recent studies suggest that LLM representations implicitly capture aspects of grounded concepts (Pavlick, 2023; Harnad, 2024). For example, LLMs are able to encode certain spatial relationships without explicit spatial inputs (Ji and Gao, 2023; Hu et al., 2024). Such cognitive abilities are of vital importance for downstream applications in human–computer interaction and embodied intelligence (Anderson, 2003; Bisk et al., 2020a). An LLM capable of reasoning about space, color, or direction can more naturally interface with the physical world (Driess et al., 2023; Pan et al., 2024), enhancing interactive systems ranging from visual dialog agents (Schumann et al., 2024) to robotic assistants (Ahn et al., 2022).

2.2 Cognitive Alignment of LLMs

Prior research has extensively probed LLMs' alignment on basic perceptual (Yuksekgonul et al., 2022) and cognitive tasks (Misra et al., 2021), from understanding physical commonsense (Bisk et al., 2020b) and visual attributes (Park et al., 2023), to theory-of-mind (Kosinski, 2023; Strachan et al., 2024) and social reasoning (Shapira et al., 2023; Gandhi et al., 2023). These studies show that LLMs can sometimes achieve human-level performance on certain cognitive tests (Binz and Schulz, 2023; Webb et al., 2023; Hubert et al., 2024) (for instance, GPT-4 matches or exceeds humans on many theory-of-mind tasks (Bubeck et al., 2023; Kosinski, 2023)).

Until now, a few isolated efforts have begun to test cross-modal preferences in LLMs (Loakman et al., 2024) (e.g. probing vision–language models for the Bouba–Kiki sound-shape mapping), but with inconclusive results (Verhoef et al., 2024).

2.3 Human Cross-Modal Perceptual Capability

Emerging evidence indicates that cross-modal perceptual mapping mechanisms analogous to synesthetic experiences may develop in the general population through environmental learning. Synesthesia, a distinct neurocognitive phenomenon, is characterized by automatic elicitation of supplementary sensory experiences triggered by specific stimulus attributes (Ward, 2013). Typical manifestations include chromatic representations of temporal units (e.g., color associations for weekdays (Cytowic and Eagleman, 2011)) and spatial localization of color perception (Arend et al., 2016). Recent interdisciplinary investigations have revealed systematic perceptual mapping phenomena within non-synesthetic populations. Current scholarship underscores the metaphorical significance of space-color/time-color associations in human cognition (Bremner et al., 2013). However, critical knowledge gaps persist regarding the neurocognitive foundations of these associations and their interaction with environmental learning mechanisms.

3 Method

Understanding the alignment between LLMs and human synesthetic patterns holds significant implications for cognitive modeling. However, the variable space of synesthetic is vast and complex. Human synesthetic responses are influenced by a multitude of factors, including cultural background, linguistic framing, conceptual familiarity, and subjective perception. Meanwhile, LLM outputs reflect biases rooted in their training data and architectural constraints.

To address this complexity, our experimental design aims to balance between variable control and interpretive richness. Rather than attempting to account for every possible factor, we adopt a principled approach to simplification—selecting conceptual domains where prior literature suggests relatively stable cross-modal associations. This allows us to reduce extraneous noise while preserving theoretical relevance.

Accordingly, we designed a comparative experiment to evaluate the responses of humans and LLMs to identical synesthetic stimuli. Our methodology consists of three core components: (i) controlled elicitation of color associations through standardized questionnaires, (ii) rigorous computational analysis of the resulting response patterns, and (iii) cross-modal comparison metrics to quantify human–LLM alignment in perceptual space.

3.1 Stimulus Design

Our synesthetic stimuli were adapted from the standardized *Synesthesia Battery* framework, with modifications to enable cross-modal human-LLM comparison. The questionnaire comprised **85 items across 8 categories**, as is demonstrated in Table 1, designed to cover both universal and culture-general associations.

The category selection was rigorously validated by referencing established protocols from the *Synesthesia Battery* studies, ensuring methodological consistency with prior research. Our stimulus set was designed to incorporate both concrete (e.g., numbers) and abstract (e.g., temporal) concepts across different cognitive levels, while deliberately excluding culturally specific items such as zodiac signs to maintain cross-cultural applicability. This balanced approach provides comprehensive coverage of synesthetic inducers while minimizing cultural biases in the assessment.

3.2 Language Localization

To control for linguistic bias:

- **Human participants** receive questionnaires in their native languages (e.g., Simplified Chinese for Chinese speakers).

- **LLMs** are queried in their dominant training languages or official language of the assigned cultural identity. We consider two kinds of models: (i) Chinese models (Doubao, DeepSeek) with native Chinese prompts; and (ii) English models (GPT series) with native English prompts containing identical semantic content. Building on this, LLMs use their mainstream training language when no identity is assigned; otherwise,

Categories	Questionnaire Contents	# Items
Numbers	0, 1, 2, 3, 4, 5, 6, 7, 8, 9	10
Weekdays	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday	7
Letters	A, B, C, D, E, F, G, H, I ,J, K, L, M, N, O, P, Q..., X, Y, Z	26
Months	January, February, ..., December	12
Seasons	Spring, Summer, Fall, Winter	4
Spatial Orientations	*Basic*: Up, Down, Left, Right, Front, Back, Center *Cardinal*: East, West, South, North *Intermediate*: Southeast, Northeast, Southwest, Northwest	15
Abstract Concepts	*Dimensional*: Far/Near, High/Low *Density*: Dense/Sparse, Deep/Shallow	8
Temporal Concepts	Past, Present, Future	3

Table 1: A detailed listing of questionnaire categories along with their specific contents. Each category encompasses a variety of items designed to comprehensively cover multiple cognitive dimensions such as numbers, temporal concepts, spatial orientations, and abstract ideas. This table serves as a foundational framework for subsequent comparative analyses by clearly outlining the distribution of items across different conceptual domains.

they use the official language of the assigned cultural identity.

3.3 Response Format Design

To address perceptual variability in color naming among individuals (e.g., divergent RGB interpretations of "dark red"), we implemented distinct response protocols to bypass linguistic ambiguity:

- **Human participants**: we implemented a constrained 15-color selection interface grounded in the Berlin-Kay basic color theory. All participants were required to select one color for each concept item. If none of the 15 colors was considered suitable, they could write down the color they thought most appropriate and provide a rationale. The procedure took approximately 5-10 minutes per participant to complete.[1]

- **LLM models**: we enforced a strict RGB output format specification (e.g., `"[1]A:(255,0,0)"`) to enable automated parsing.

3.4 Metrics

We conduct a fine-grained quantitative anslysis to evaluate the alignment between LLMs and human synesthetic color associations. For each concept, we first identify the *dominant color*—defined as the most frequently selected color by human participants and the highest-probability prediction generated by each LLM. The perceptual difference between the human and model-assigned colors was then calculated using the CIEDE2000 formula (ΔE_{00}), a widely used metric for quantifying perceptual color dissimilarity. For detailed introduction and formulation, please refer to Appendix A.

4 Experiment

4.1 Data Collection

- **Human participants** We distributed an online questionnaire via Tencent Docs (a Chinese cloud-based survey platform comparable to Google Forms) to collect human synesthetic associations. The survey required approximately 10–15 minutes for completion and yielded 289 valid responses after excluding incomplete or inconsistent submissions. Participants were presented with identical stimulus words used for LLM evaluation to ensure direct comparability.

- **LLMs** We evaluated the following models:
  - **Doubao**: doubao-1-5-pro-32k, doubao-1-5-lite-32k, doubao-1-5-vision-pro-32k (AI, 2024) (3 items)
  - **Deepseek**:deepseek-r1-distill-qwen-32b (Bai et al., 2023), deepseek-r1 (Guo et al., 2025), deepseek-v3 (Liu et al., 2024)(3 items)
  - **ChatGPT**: gpt-4 (Achiam et al., 2023), gpt-4-32k, gpt-4o (Hurst et al., 2024), gpt-4o-

[1]Prior to the study, all participants provided informed consent, which explicitly outlined the purpose of data collection and its use for research purposes. The procedures were approved by the Ethics Committee of the XXX (masked due to double blind review policy). All data were analyzed using IBM SPSS Statistics version 20.0.

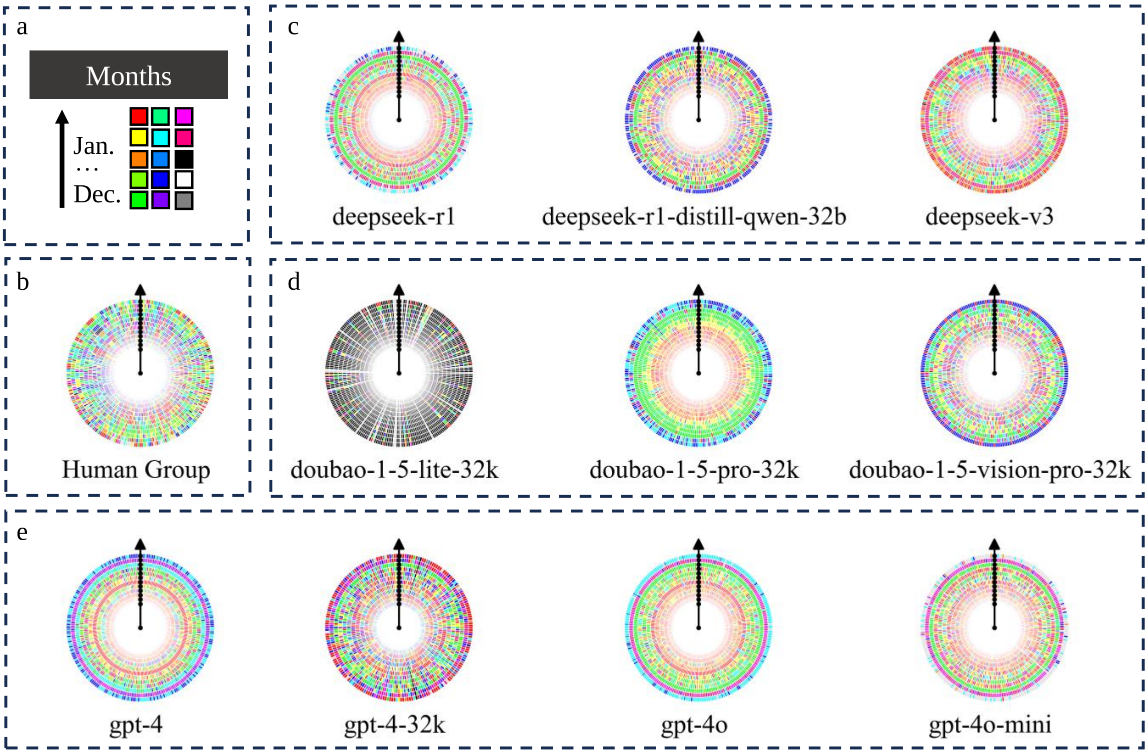

Figure 1: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the month-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents months from December (center) to January (outer edge), and the angular direction indicates cleaned response data.

mini (4 items)

We implemented two types of experimental settings: one with default identity (no assigned cultural or gender identity) and another with explicit cultural and gender assignments. All API parameters remained at their platform-default settings throughout the experiment, with no manual adjustment of temperature, top_p, or other generation parameters. In the identity-assigned setting, we assigned six cultures (Chinese, American, Japanese, Korean, British, and Russian) and two genders to the models.

4.2 Data Processing

Human responses: After applying response completeness and attention checks to the initial 289 submissions, we retained 260 valid human questionnaires from online, yielding an 89.7% valid response rate. Among these, 143 were male and 114 female participants, primarily drawn from the Chinese cultural context with ages concentrated between 20 and 30 years old.

LLM responses: For the responses generated by large language models (LLMs), we first excluded all outputs that did not conform to the predefined format (e.g., textual descriptions instead of numerical RGB values). Under the default (no assigned identity) condition, each model produced 300 samples. When cultural and gender identities were assigned, each model generated 50 samples for each culture-gender combination. For both the default and identity-assigned settings, we constructed balanced datasets for the same series of LLMs respectively. To ensure balanced representation across models, we employed a stratified sampling strategy, randomly selecting an equal number of qualified responses from each model to form balanced datasets containing 300 samples each. This sampling process strictly adhered to two principles: (i) maintaining equal contribution weights across all models, and (ii) ensuring uniform coverage of all stimulus words.

4.3 Data Analysis

To examine broader patterns, we grouped the questions into several *conceptual categories* (e.g., spatial directions, temporal concepts, sequences, letters) and computed the average ΔE_{00} within each category. This group-level aggregation enabled us to assess how consistently different semantic domains are mapped to color by LLMs compared to humans.

To provide a comprehensive view of model-human alignment, we computed the residuals for each individual model, question, and concept triplet. These residuals, defined as the CIEDE2000 color distance (ΔE_{00}) between the model-assigned and human-assigned dominant colors, serve as a fine-grained metric of perceptual deviation. All computed residuals are presented in the appendix to support transparency and reproducibility of our analysis.

5 Results

In order to compare the concept-color associations between the human and LLMs (deepseek and doubao). A three-way log-linear analysis: concept (85) × color (15) × group (2), on the choosing frequency of colors showed that there was a significant difference in concept-color associations between the two groups (e,g., weekdays, χ^2 = 1791.978, *df* = 168, $p < .0001$), except for the letters-color associations.

5.1 Overall

Statistical analysis revealed non-random color associations for all concepts including human and LLMs (for examples, Numbers, human: $\chi^2 = 1149.626$; GPT: $\chi^2 = 3153.075$; deepseek: $\chi^2 = 7496.427$, doubao: $\chi^2 = 5730.481$; all $df = 126$ and $p < .0001$), and some colors were chosen more frequently than others for specific concept. Each concept exhibited different patterns (e.g., number 0-color associations, human: $\chi^2 = 345.054$, $df = 13$, $p < .0001$). For number 0-color association in human group, white was chosen more frequently than other colors (with an adjusted residual $z = 23.09$). Other significant associations were observed. For examples, 1 with red ($z = 5.40$) and 9 with black ($z = 4.03$). Overall, our results demonstrate that *temporal concepts*—particularly months and seasons—yield the smallest average color difference between humans and LLMs.

5.2 Differences in Color Preferences

Human participants and LLMs exhibit pronounced differences in synesthetic responses. As shown in Appendix B, the color selections of LLMs form a distinct and concentrated circular pattern, whereas human responses appear much sparser and more dispersed. This divergence manifests not only in hue preference—with humans favoring blue tones and LLMs leaning toward reds. LLMs demonstrate a far higher degree of uniformity in color choices, indicating a more systematic and stable associative mechanism, while human synesthesia is shaped by individual differences and perceptual variability. These findings underscore the fundamental differences between human and AI cross-modal representation systems.

Notably, human responses revealed a primacy effect in sequential concepts (e.g., days of the week, months), with a significantly higher tendency to assign red hues to the first item in each sequence. This pattern may reflect attentional salience or learned associations in cognitive processing. In contrast, LLMs did not exhibit such a primacy effect; their color assignments appeared evenly distributed across the sequence, suggesting that their associative mappings are less influenced by ordinal positioning and more likely governed by statistical co-occurrence or embedding structure.

5.3 Characteristics of Conceptual Alignment

Between human participants and LLMs, color associations for seasonal and temperature-related concepts exhibit a high degree of consistency. As shown in the Figure 1, "July" is commonly associated with warm hues, while "Winter" is generally linked to cool tones. Figure 4 illustrates the color differences between humans and LLMs. These concepts show the smallest color differences between humans and LLMs, with seasons having the highest consistency, followed by months. This cross-modal consistency indicates that LLMs can effectively encode sensory mappings related to natural regularities. The relatively small color distances in these concepts further corroborate the ability of LLMs to capture synesthetic patterns associated with natural phenomena and shared cultural contexts.

5.4 Differences Due to Model Size and Type

LLMs of different size and architectures showed varying performance in the color synesthesia task. Generally, larger models tended to produce more distinct and differentiated color choices, while smaller models leaned towards more conservative outputs in grayscale or low-saturation colors. Significant differences were also observed among models within the same series, indicating that model scale is not the sole determinant of performance.

Models \ Questions		Numbers									
		0	1	2	3	4	5	6	7	8	9
Human	Male	35/143 23/143 21/143	25/143 15/143 15/143	27/143 17/143 17/143	26/143 18/143 16/143	25/143 25/143 15/143	18/143 17/143 17/143	25/143 19/143 17/143	23/143 17/143 16/143	20/143 14/143 13/143	19/143 14/143 13/143
	Female	53/117 11/117 9/117	26/117 15/117 15/117	23/117 19/117 14/117	21/117 15/117 14/117	19/117 15/117 13/117	17/117 14/117 14/117	17/117 16/117 12/117	21/117 18/117 13/117	15/117 15/117 13/117	14/117 12/117 10/117
GPT	Male	111/150 27/150 12/150	108/150 16/150 9/150	65/150 21/150 18/150	38/150 29/150 26/150	37/150 26/150 23/150	23/150 18/150 17/150	27/150 19/150 19/150	26/150 22/150 19/150	29/150 25/150 16/150	25/150 24/150 21/150
	Female	114/150 25/150 11/150	101/150 17/150 14/150	87/150 21/150 12/150	39/150 32/150 22/150	26/150 24/150 23/150	24/150 20/150 18/150	24/150 20/150 18/150	20/150 20/150 19/150	26/150 23/150 22/150	26/150 20/150 19/150
Deepseek	Male	82/150 62/150 3/150	61/150 49/150 33/150	64/150 39/150 16/150	75/150 26/150 12/150	59/150 22/150 18/150	56/150 27/150 24/150	50/150 24/150 22/150	39/150 38/150 30/150	33/150 28/150 20/150	50/150 18/150 15/150
	Female	82/150 61/150 7/150	57/150 50/150 40/150	71/150 31/150 20/150	85/150 24/150 11/150	58/150 26/150 22/150	63/150 26/150 23/150	60/150 25/150 25/150	43/150 40/150 32/150	41/150 26/150 17/150	34/150 30/150 18/150
Doubao	Male	150/150	127/150 23/150	51/150 49/150 28/150	49/150 47/150 28/150	44/150 43/150 28/150	45/150 28/150 19/150	49/150 32/150 28/150	35/150 29/150 28/150	54/150 29/150 17/150	60/150 34/150 19/150
	Female	146/150 4/150	133/150 17/150	51/150 48/150 22/150	48/150 42/150 22/150	39/150 36/150 34/150	39/150 33/150 18/150	49/150 28/150 23/150	37/150 32/150 24/150	37/150 36/150 26/150	51/150 30/150 29/150

Figure 2: Top-3 most frequent color choices for number-related concepts across different models. Each cell represents the top three RGB colors selected for a given number item by a specific model under a given gender identity, along with the number of responses selecting each color and the total number of valid responses. While no consistent gender-based differences were observed within models, there are clear discrepancies between human responses and those generated by large language models (LLMs), suggesting a fundamental divergence in synesthetic patterns.

5.5 Differences Due to Individual Identity Variations

Individual identity factors significantly influence synesthetic experiences, highlighting the diversity and complexity of synesthesia.

In terms of gender, the experimental results (shown in Figure 2) do not show a significant gender difference. Possible explanations include insufficient sample size, the specific item's limited sensitivity to gender effects. Future research may explore potential gender influences on color synesthesia by increasing sample size or employing more detailed categorization methods. In terms of culture, LLMs exhibit noticeable changes in synesthetic patterns under different cultural contexts (shown in Figure 3). Although all models were exposed to the same semantic content, their color associations varied when assigned identities from distinct cultural backgrounds. This suggests that LLMs may internalize and reflect culture-specific associations encoded during training. These variations highlight the influence of cultural framing on AI-generated cross-modal mappings and point to the importance of identity conditioning in studying AI perception.

6 Discussion

Despite meaningful progress in examining the alignment of synesthetic associations between LLMs and humans, this study has several limitations. First, human participants were predominantly from China, resulting in a relatively homogenous cultural background that may limit the generalizability of the findings. Second, the sample size of human participants was modest, potentially reducing statistical power and robustness of conclusions. Third, the range of LLMs evaluated remains limited in terms of architectures and training paradigms, particularly regarding multimodal and cross-lingual models. Expanding the diversity and number of models will provide a more comprehensive understanding of synesthetic behavior across AI systems. Additionally, synesthesia measurements were based primarily on questionnaires, which may not fully capture dynamic or implicit synesthetic experiences; integrating physiological or neuroscientific measures could yield deeper insights. Lastly, given the complexity of cultural and individual differences in synesthesia, LLMs trained solely on language data lack the biological and emotional mechanisms underpinning human synesthetic perception, restricting their ability to fully emulate human-like synesthesia.

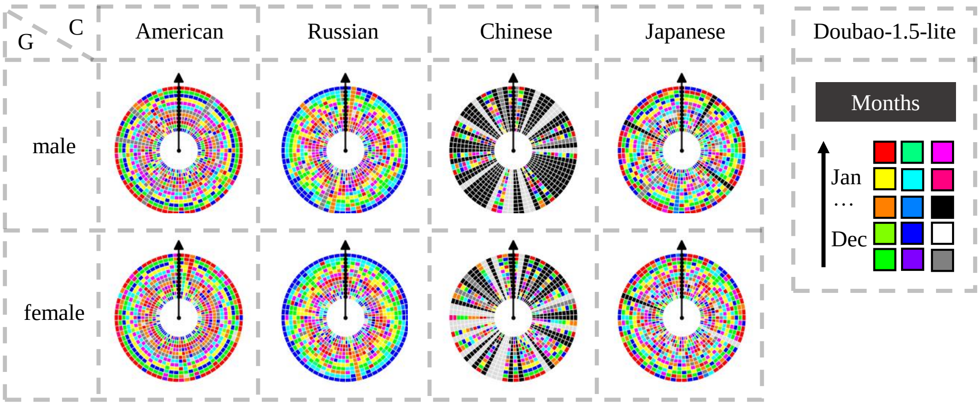

Figure 3: Color synesthetic patterns generated by the Doubao-1.5-lite model under different cultural and gender identity settings. The model shows markedly different responses depending on the assigned cultural context. Under the Chinese identity, the outputs tend to cluster around achromatic tones (e.g., black, white, gray), suggesting lower activation of synesthetic associations. In contrast, the American, Russian, and Japanese settings exhibit more vivid and consistent patterns, each reflecting distinct cultural color tendencies.

Figure 4: Color differences in synesthetic associations between large language models and human data across conceptual categories. For each concept, the most salient color was extracted, and the CIEDE2000 color difference was calculated between model and human responses. Average differences were then computed within each conceptual group. Notably, the smallest discrepancies occurred in time-related concepts such as months and seasons.

This study primarily elucidated the alignment between LLMs and human synesthetic associations across various conceptual domains. Future work can be advanced along two key directions. First, scaling up the scope of research by increasing both human and model samples. Expanding human participants' cultural and linguistic diversity, as well as sample sizes, will enhance the generalizability and statistical robustness of findings. Concurrently, incorporating a broader spectrum of LLM architectures, training paradigms, and multimodal integrations will enable a systematic evaluation of how model design influences synesthetic behavior, contributing to a comprehensive AI synesthesia cognitive map.

Second, deeper investigation into the internal structures and training mechanisms of LLMs that govern synesthetic expressions. While the current work focuses on descriptive results, future studies should examine how prompt engineering and language cues can modulate LLM synesthetic outputs, shedding light on the role of linguistic context in shaping cross-modal associations. Additionally, fine-tuning and targeted training approaches could be explored to guide models toward more human-like synesthetic representations. These efforts will not only deepen our understanding of LLM cognitive processes but also provide theoretical and practical foundations for developing AI with richer, human-aligned multimodal perception.

7 Conclusion

In this study, we compared synesthetic associations between large language models and humans, revealing both notable alignments and clear differences across conceptual domains. Findings highlight the significant impact of model scale and architecture on synesthetic behavior. Future efforts to scale up samples and probe model mechanisms will advance AI's multimodal cognitive capabilities. Overall, this work provides a valuable foundation for understanding and enhancing human-like perception in artificial intelligence.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

DeepSeek AI. 2024. Deepseek-moe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Michael L Anderson. 2003. Embodied cognition: A field guide. *Artificial intelligence*, 149(1):91–130.

Isabel Arend, Shiran Ofir, and Avishai Henik. 2016. What spatial coordinate defines color-space synesthesia? *Brain and Cognition*, 105:88–94.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, and 1 others. 2020a. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020b. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Lera Boroditsky and 1 others. 2009. How does our language shape the way we think. *What's next*, 6:116–129.

Andrew J Bremner, Serge Caparos, Jules Davidoff, Jan De Fockert, Karina J Linnell, and Charles Spence. 2013. "bouba" and "kiki" in namibia? a remote culture make similar shape–sound matches, but different shape–taste matches to westerners. *Cognition*, 126(2):165–172.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Richard E Cytowic and David M Eagleman. 2011. *Wednesday is indigo blue: Discovering the brain of synesthesia*. Mit Press.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, and 1 others. 2023. Palm-e: An embodied multimodal language model.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2024. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Peter G Grossenbacher and Christopher T Lovelace. 2001. Mechanisms of synesthesia: cognitive and physiological constraints. *Trends in cognitive sciences*, 5(1):36–41.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Stevan Harnad. 2024. Language writ large: Llms, chatgpt, grounding, meaning and understanding. *arXiv preprint arXiv:2402.02243*.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. 2024. Chain-of-symbol prompting for spatial reasoning in large language models. In *First Conference on Language Modeling*.

Edward M Hubbard. 2007. Neurophysiology of synesthesia. *Current psychiatry reports*, 9(3):193–199.

Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14(1):3440.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yuhan Ji and Song Gao. 2023. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations. *arXiv preprint arXiv:2307.03678*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Simon Lacey, Margaret Martinez, Nicole Steiner, Lynne C Nygaard, and K Sathian. 2021. Consistency and strength of grapheme-color associations are separable aspects of synesthetic experience. *Consciousness and cognition*, 91:103137.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tyler Loakman, Yucheng Li, and Chenghua Lin. 2024. With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. *arXiv preprint arXiv:2409.14917*.

Lawrence E Marks. 1987. On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of experimental psychology: Human perception and performance*, 13(3):384.

Daphne Maurer, Thanujeni Pathman, and Catherine J Mondloch. 2006. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental science*, 9(3):316–322.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2021. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

Bo Pan, Jiaying Lu, Ke Wang, Li Zheng, Zhen Wen, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. 2024. Agentcoord: Visually exploring coordination strategy for llm-based multi-agent collaboration. *arXiv preprint arXiv:2404.11943*.

Cesare V Parise. 2016. Crossmodal correspondences: Standing issues and experimental guidelines. *Multisensory research*, 29(1-3):7–28.

Jae Sung Park, Jack Hessel, Khyathi Chandu, Paul Pu Liang, Ximing Lu, Peter West, Youngjae Yu, Qiuyuan Huang, Jianfeng Gao, Ali Farhadi, and 1 others. 2023. Localized symbolic knowledge distillation for visual commonsense models. *Advances in Neural Information Processing Systems*, 36:11338–11352.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. 2024. Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133.

Vilayanur S Ramachandran and Edward M Hubbard. 2001. Synaesthesia–a window into perception, thought and language. *Journal of consciousness studies*, 8(12):3–34.

Anina N Rich and Jason B Mattingley. 2002. Anomalous perception in synaesthesia: a cognitive neuroscience perspective. *Nature Reviews Neuroscience*, 3(1):43–52.

Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. 2024. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18924–18933.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Gaurav Sharma, Wencheng Wu, and Edul N Dalal. 2005. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(1):21–30.

Ferrinne Spector and Daphne Maurer. 2013. Synesthesia: a new approach to understanding the development of perception.

Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73:971–995.

Charles Spence and Ophelia Deroy. 2012. Crossmodal correspondences: Innate or learned? *i-Perception*, 3(5):316–318.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. 2025. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tessa Verhoef, Kiana Shahrasbi, and Tom Kouwenhoven. 2024. What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. *arXiv preprint arXiv:2407.17974*.

Jamie Ward. 2013. Synesthesia. *Annual review of psychology*, 64(1):49–75.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, and 1 others. 2025. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9247–9255.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

A CIEDE2000

A.1 Formula

The CIEDE2000 formula takes into account not just the Euclidean distance between colors in the CIELAB space but also factors like lightness, chroma, and hue differences. The general form of the CIEDE2000 color difference formula is:

$$\Delta E_{00}^2 = \left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T \frac{\Delta C'}{k_C S_C}\frac{\Delta H'}{k_H S_H},$$

where $\Delta L'$ is Lightness difference, $\Delta C'$ denotes Chroma difference, $\Delta H'$ means Hue difference, S_L, S_C, S_H are Weighting functions for lightness, chroma, and hue, respectively. k_L, k_C, k_H are Parametric factors (usually set to 1 for standard conditions), and R_T = Rotation term accounting for interactions between chroma and hue differences. The detailed introduction to each parts are as follow:

- **Lightness Difference**: $\Delta L' = L_2^* - L_1^*$
- **Chroma Difference**: $C^* = \sqrt{(a^*)^2 + (b^*)^2}$ with $\Delta C' = C_2' - C_1'$.
- **Hue Difference**:

$$\Delta H' = 2\sqrt{C_1' C_2'} \sin\left(\frac{\Delta h'}{2}\right)$$

- **Weighting Functions**:
  - $S_L = 1 + \frac{0.015(L'-50)^2}{\sqrt{20+(L'-50)^2}}$
  - $S_C = 1 + 0.045C'$
  - $S_H = 1 + 0.015C'T$
- **Hue Rotation Term**: $R_T = -2R_C \sin(2\Delta\theta)$ with $R_C = \sqrt{\frac{C'^7}{C'^7+25^7}}$.

A.2 Usage in Academic Context

The CIEDE2000 formula is commonly used in image processing, textile engineering, printing, and quality control where precise color matching is critical. It is considered more accurate than earlier formulas (like CIELAB and CIE94) due to its nuanced handling of chroma and hue interactions, particularly in cases of significant hue angle differences.

B Supplementary Information for the Synesthesia Task

In this section, we provide the synesthesia prompt instructions used for LLMs under different cultural settings, along with additional detailed results. These include comparisons between LLMs and human participants across conceptual categories, per-item color residual analyses for each model series, and culture-specific outputs from different LLMs on the "months" category.

B.1 LLM prompt

Please, based on your default associations, sequentially provide the colors corresponding to the following nouns and represent them using RGB values (such as 255, 87, 51). Please answer in the order of the questions, in the format of [Question Number: Content of the Question]: (r, g, b), and do not omit any.
1:0
2:1
3:2
4:3
5:4
6:5
7:6
8:7
9:8
10:9
11:Monday
12:Tuesday
13:Wednesday
14:Thursday
15:Friday
16:Saturday
17:Sunday
18:January
19:February
20:March
21:April
22:May
23:June
24:July
25:August
26:September
27:October
28:November
29:December
30:Spring
31:Summer
32:Autumn
33:Winter
34:A
35:B
36:C
37:D
38:E
39:F
40:G
41:H
42:I
43:J
44:K
45:L
46:M
47:N
48:O
49:P
50:Q
51:R
52:S
53:T
54:U
55:V
56:W
57:X
58:Y
59:Z
60:Up
61:Down
62:Left
63:Center
64:Right
65:Forward
66:Backward
67:East
68:West
69:South
70:North
71:Southeast
72:Northeast
73:Southwest
74:Northwest
75:High
76:Low
77:Far
78:Near
79:Deep
80:Shallow
81:Sparse
82:Dense
83:Past
84:Present

85:Future

B.2 LLM and Human Responses by Conceptual Category

we provide the extra results of Table 1, as is illustrated in Figure 5 to Figure 15.

B.3 Per-Item Color Residuals by Model Series

We present the Per-Item Color Residuals by Model Series analysis, as is illustrated in Figure 16 to Figure 27.

Each cell displays the three colors with the largest residuals along with their corresponding residual values, making it easier for readers to assess the significance of the colors.If the residual is greater than 5, it is highlighted in bold. If the residual is less than 0, its opacity is set to 30%.

B.4 Culture-Specific Outputs of LLMs on the "Months" Category

We present the Culture-Specific Outputs of LLMs on the "Months" Category, as is illustrated in Figure 28 to Figure 35.

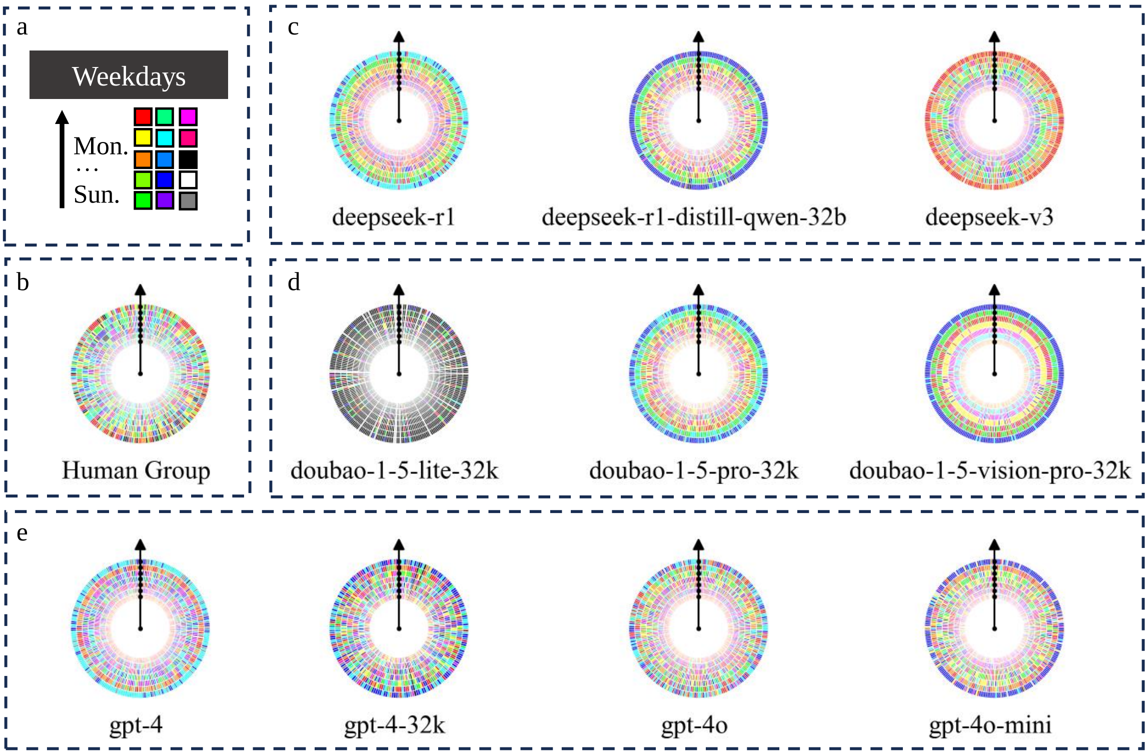

Figure 5: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the weekday-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers from Sunday (center) to Monday (outer edge), and the angular direction indicates cleaned response data.

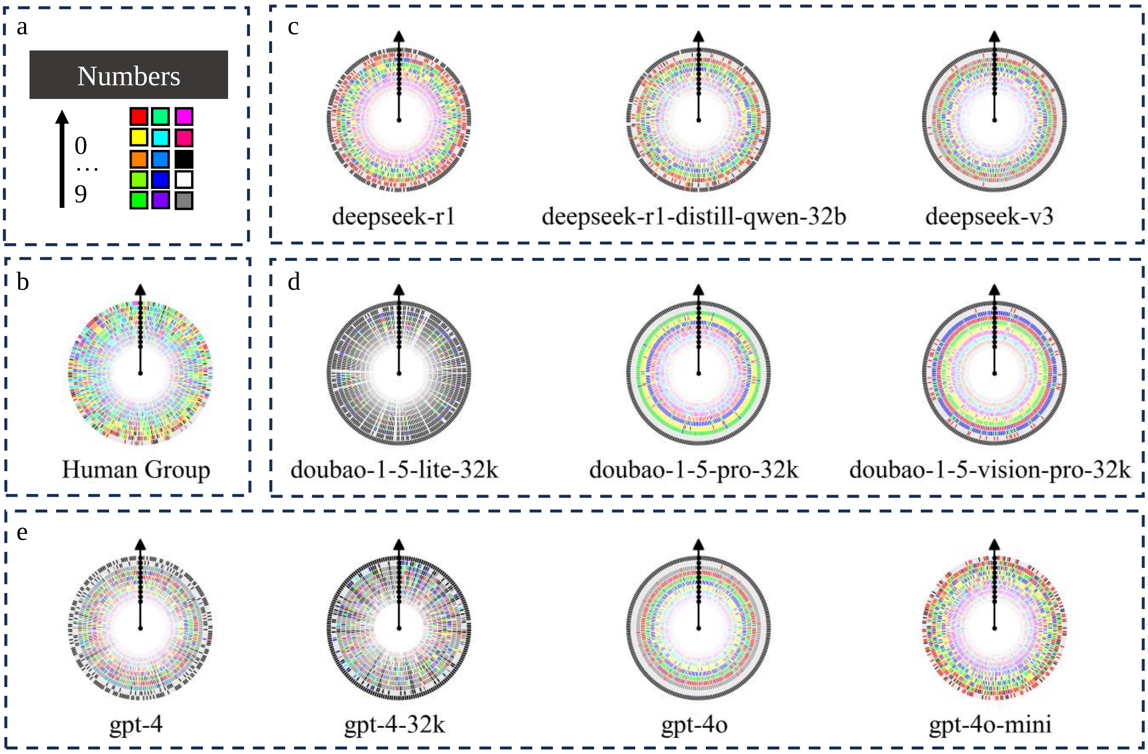

Figure 6: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the number-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers from 9 (center) to 0 (outer edge), and the angular direction indicates cleaned response data.

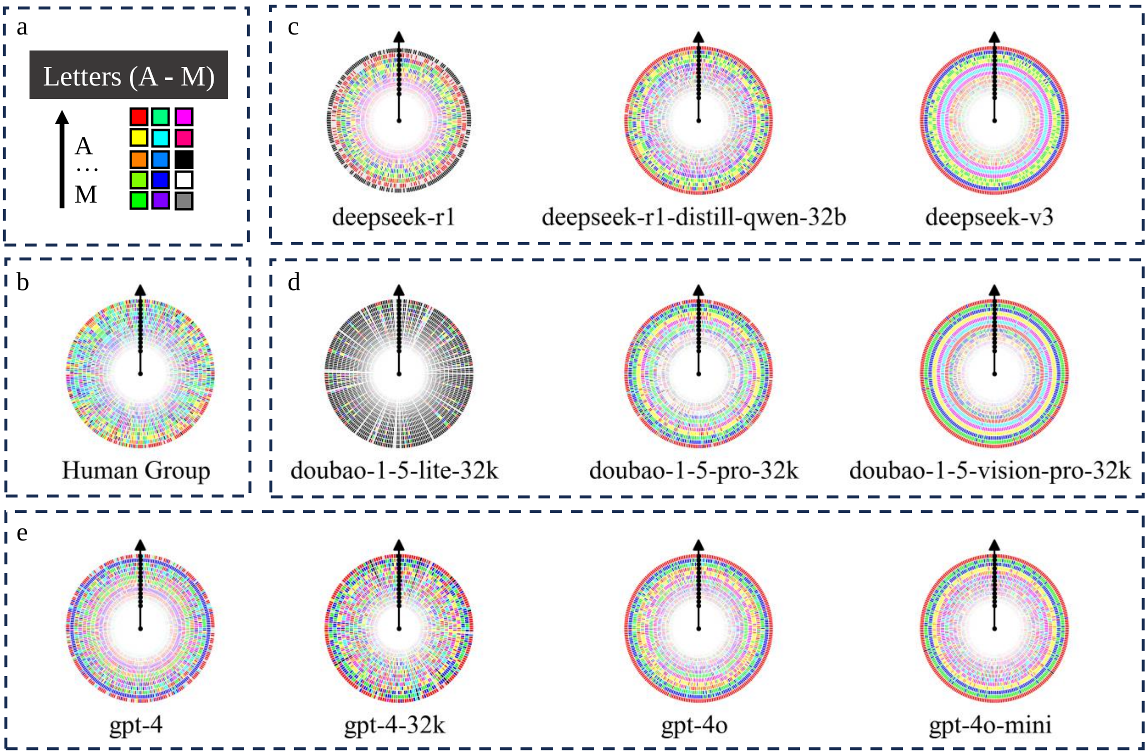

Figure 7: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the weekday-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers from A (outer edge) to M(center), and the angular direction indicates cleaned response data.

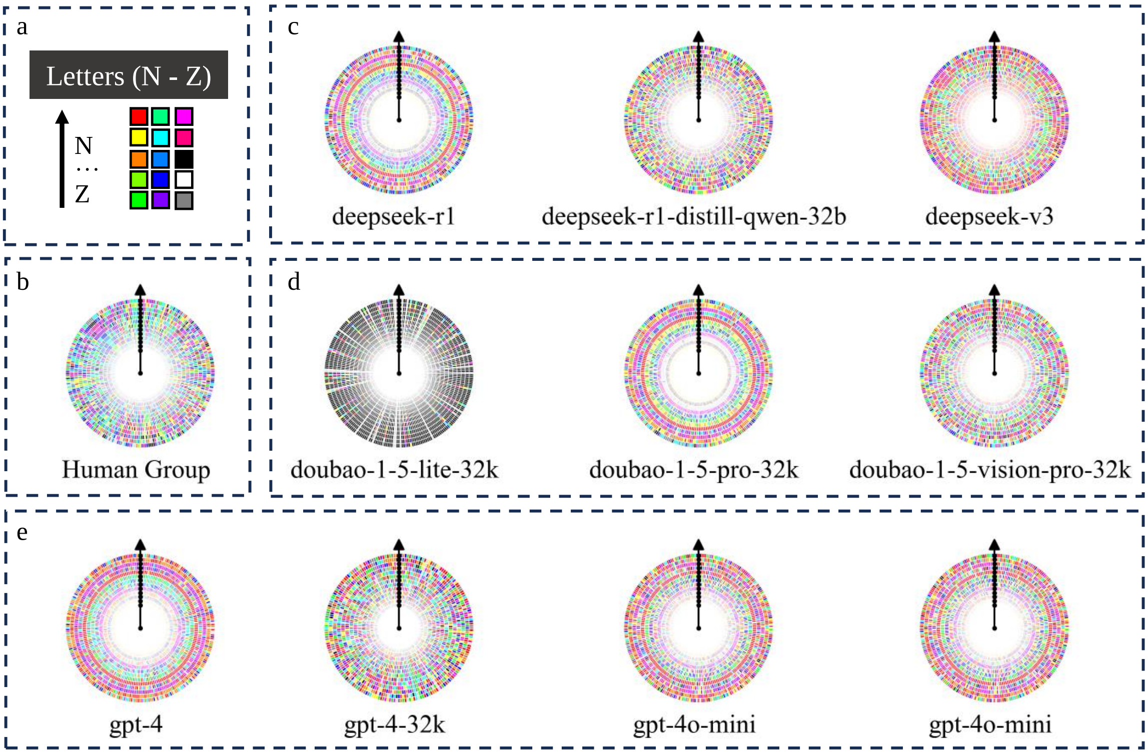

Figure 8: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the weekday-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers from N (outer edge) to Z(center), and the angular direction indicates cleaned response data.

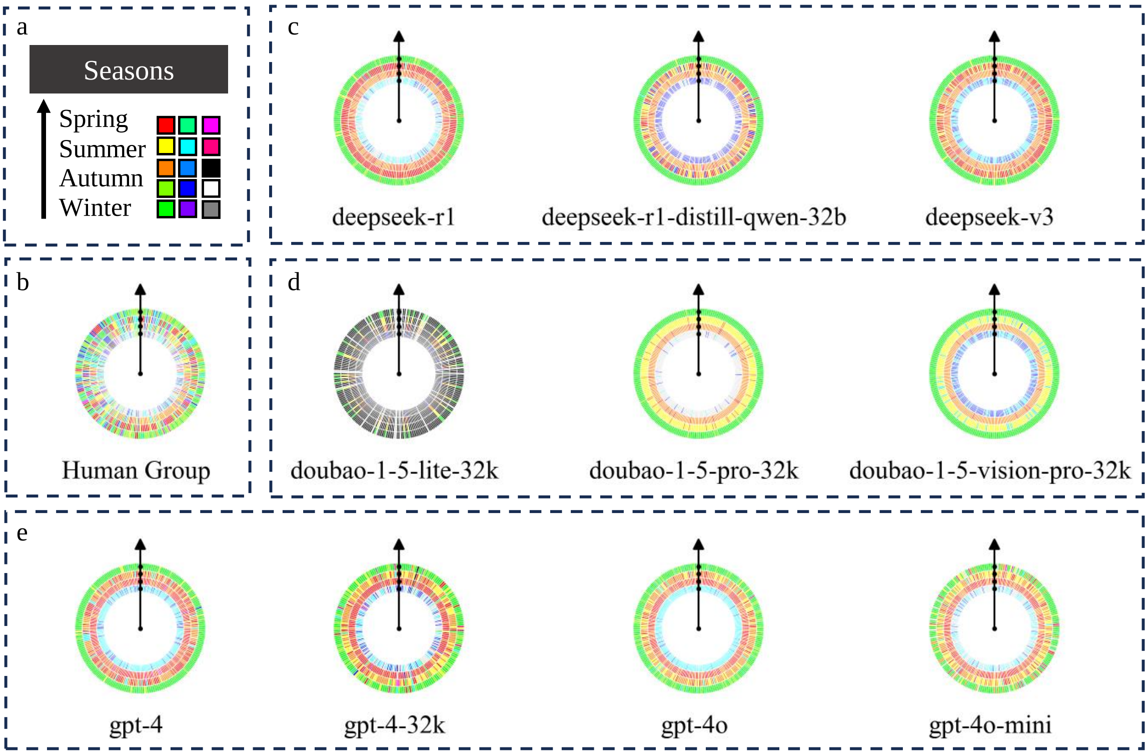

Figure 9: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the season-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers from spring (outer edge) to winter(center), and the angular direction indicates cleaned response data.

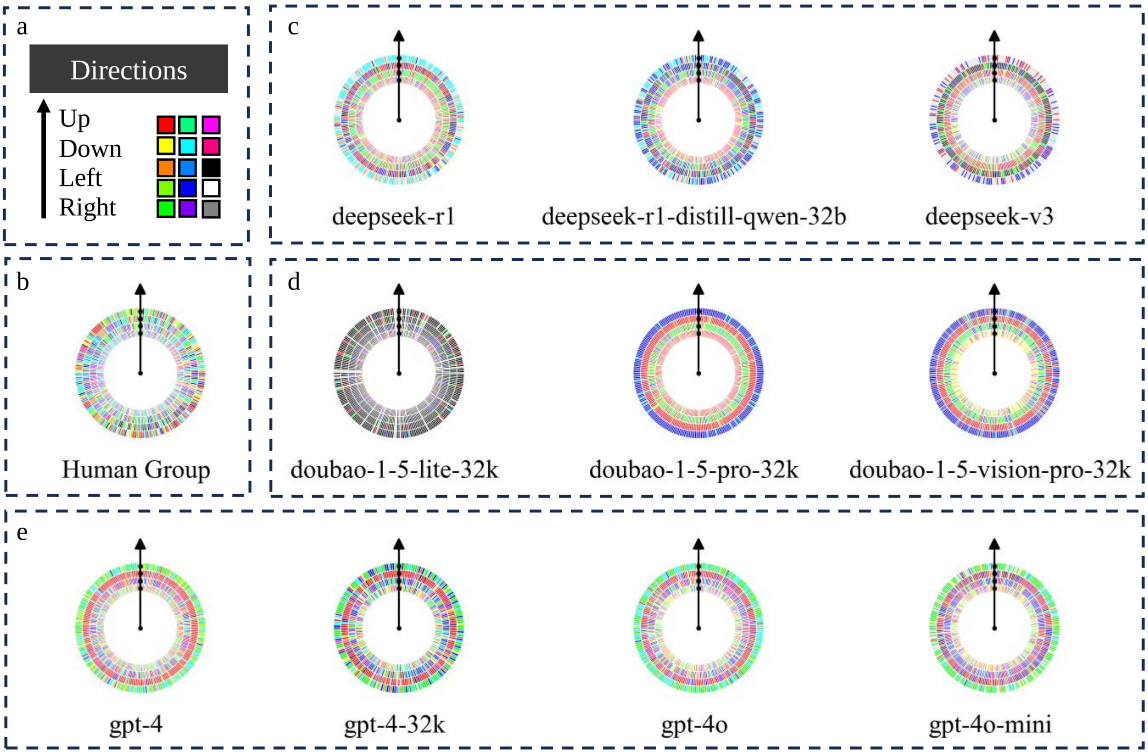

Figure 10: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the direction-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers of up, down, left and right, and the angular direction indicates cleaned response data.

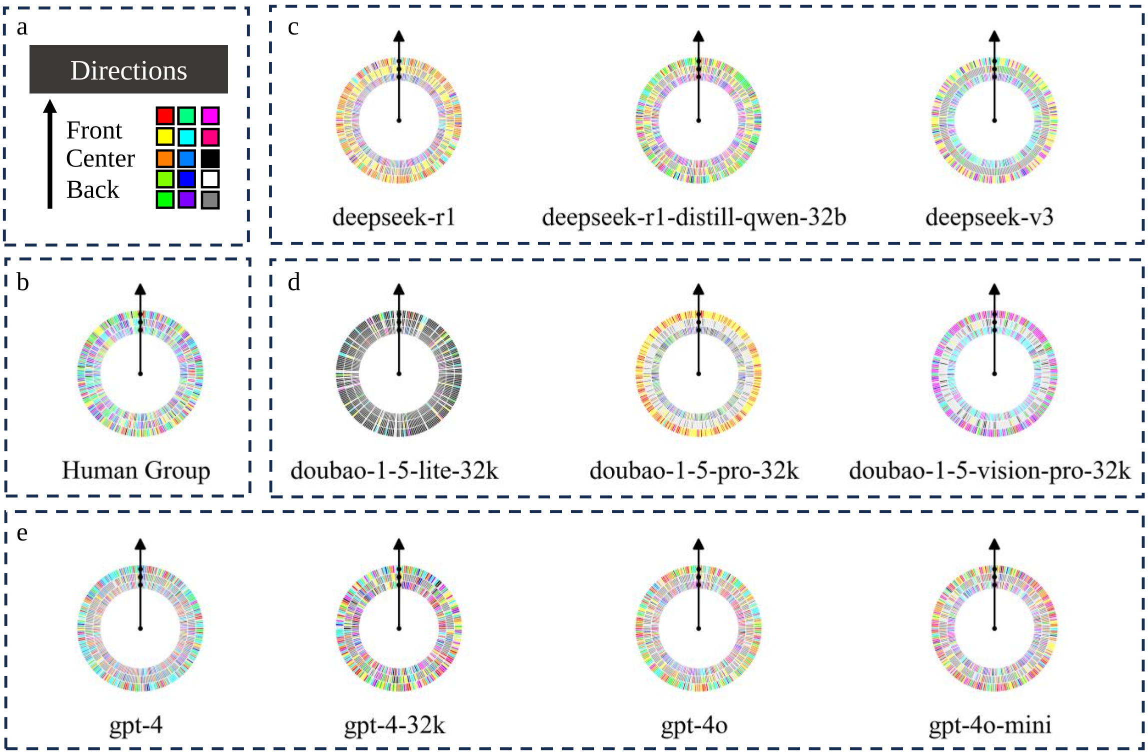

Figure 11: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the direction-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers of front, center and back, and the angular direction indicates cleaned response data.

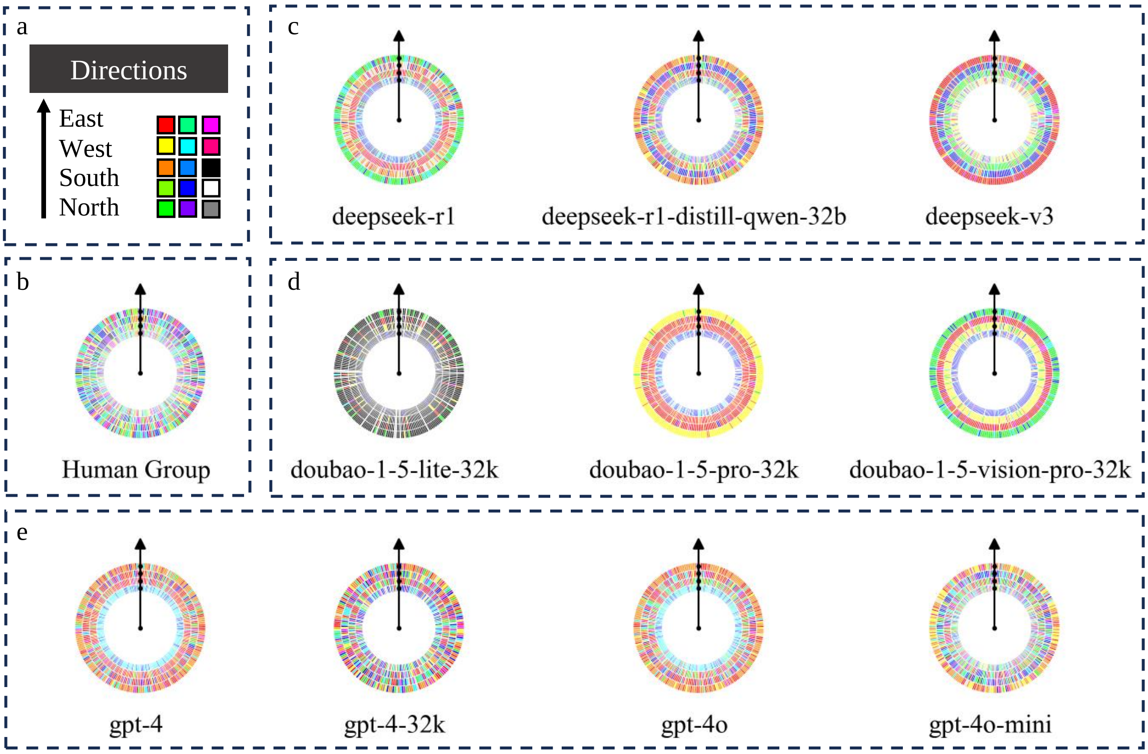

Figure 12: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the direction-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers of east, west, south and north, and the angular direction indicates cleaned response data.

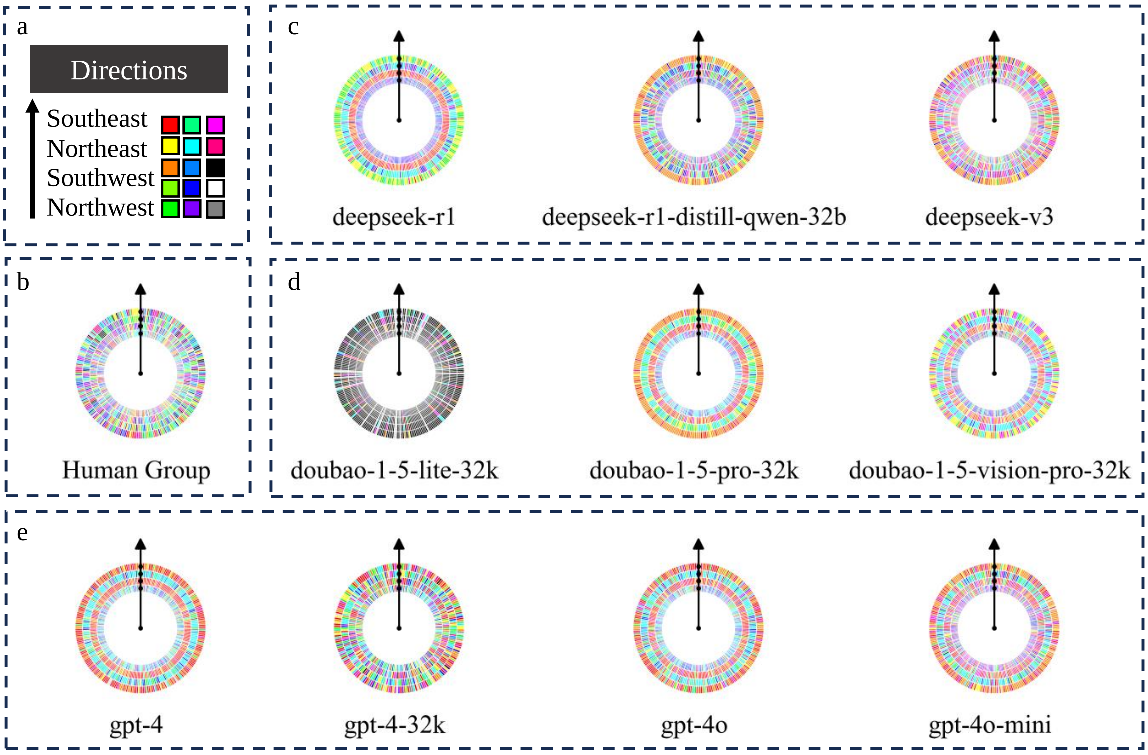

Figure 13: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the direction-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers of southeast, northeast, southwest and northwest, and the angular direction indicates cleaned response data.

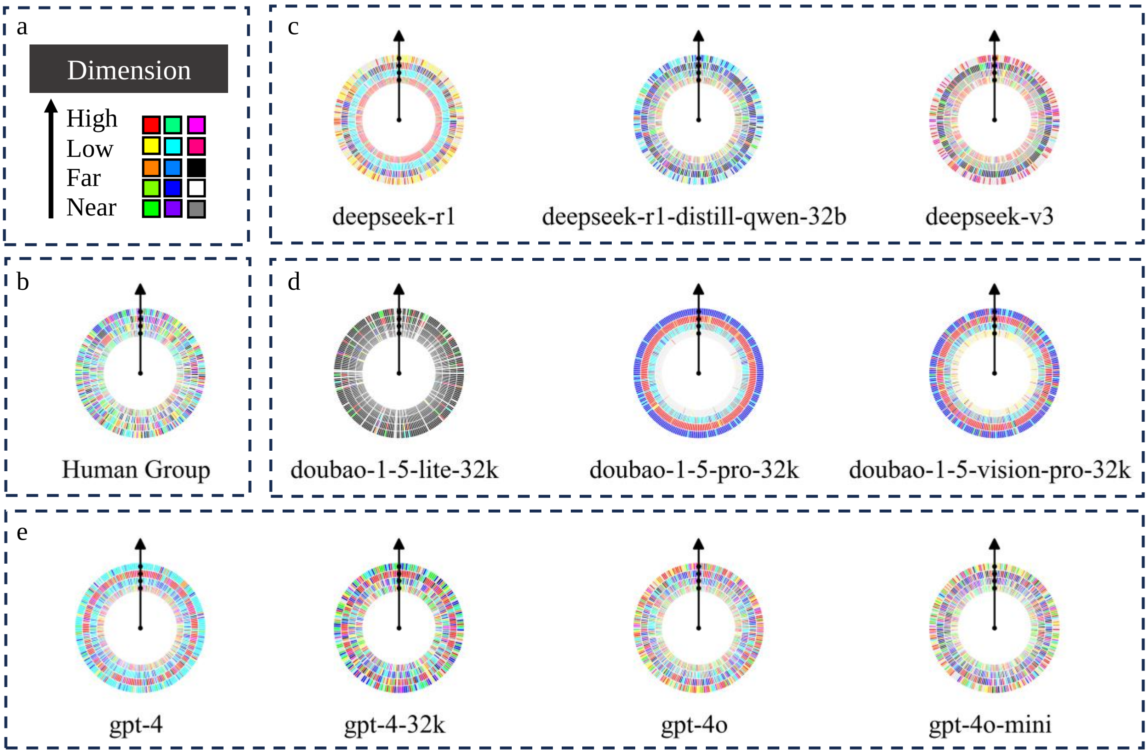

Figure 14: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the abstract-concept-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers of high, low, far and near, and the angular direction indicates cleaned response data.

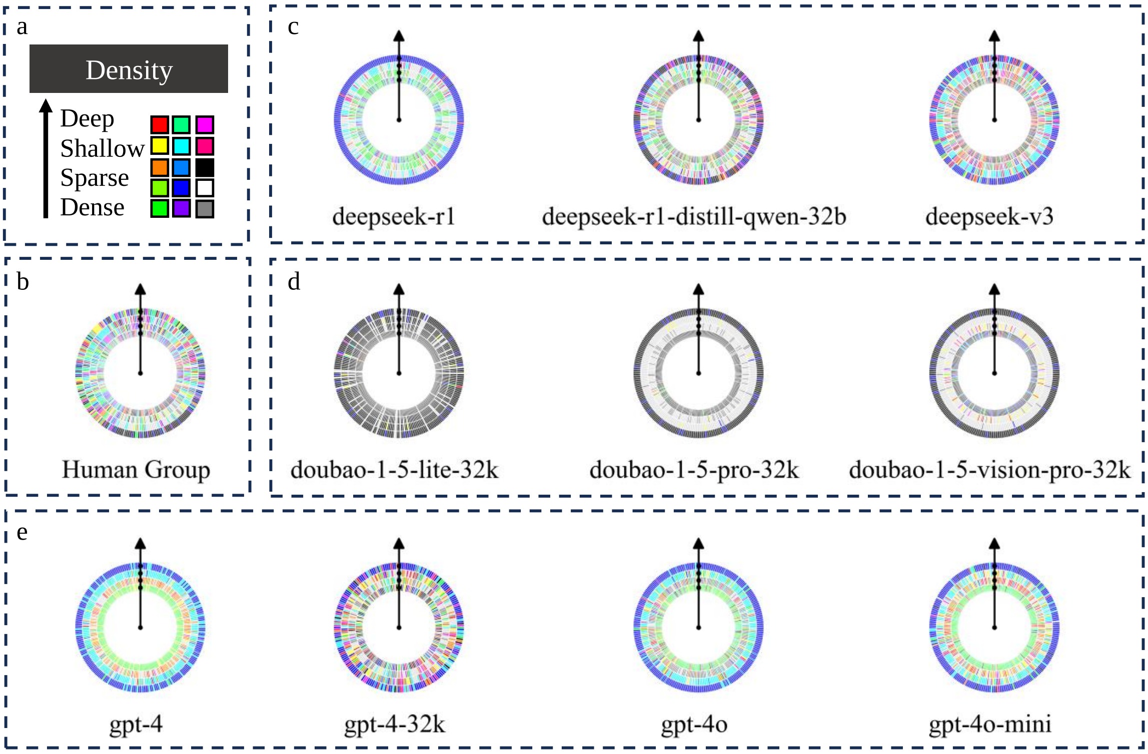

Figure 15: Figure (a) shows the legend, including the response colors and the definition of the radial coordinate. Figures (b), (c), (d), and (e) respectively present the density-concept-related synesthetic color responses from the human group, DeepSeek series, Doubao series, and ChatGPT series. The radial axis represents numbers of deep, shallow, sparse and dense, and the angular direction indicates cleaned response data.

M \ Q	0	1	2	3	4	5	6	7	8	9
Human	**23.09** 4.09 3.02	**5.40** **5.26** 2.64	**6.47** 4.28 2.23	**5.53** 4.22 1.90	4.82 2.98 2.07	2.85 2.70 1.78	3.18 2.70 2.61	4.38 3.84 2.63	**6.50** 2.25 1.95	4.03 3.13 2.68
GPT	**26.53** **8.34** -0.49	**26.75** 2.12 0.33	**17.10** **7.48** 0.41	**9.95** 4.74 3.69	**11.80** 4.29 1.78	**11.93** 3.19 2.07	**11.99** **5.72** 4.08	**13.85** 4.97 4.89	**10.96** 4.46 4.46	**6.30** 4.04 2.57
Deepseek	**50.54** -0.47 -0.67	**37.74** **10.09** -0.47	**16.69** **5.86** **5.74**	**23.13** 3.24 1.89	**18.45** **10.35** **5.32**	**18.42** **8.03** 2.44	**14.40** **12.61** 4.08	**17.33** **16.66** 4.78	**12.94** **9.63** **5.95**	**11.29** **6.27** **5.86**
Doubao	**26.99** -0.33 -2.86	**36.42** -0.33 -2.86	**16.71** **11.07** **5.71**	**17.16** **12.03** -0.33	**20.17** **8.91** -0.33	**17.50** **12.25** 1.14	**24.09** 4.40 -0.33	**30.17** 0.72 0.72	**14.23** **5.77** 3.10	**13.04** **6.09** 4.14

Figure 16: Color Residuals by Model Series for the Number Group

Q M	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Human	**10.41** **5.74** 3.16	**7.13** 2.97 2.03	3.99 1.62 1.58	3.10 2.34 2.18	3.86 2.84 2.19	1.93 1.84 1.43	**5.37** 3.31 1.95
GPT	**14.58** **10.64** **5.03**	**8.75** **5.89** 2.21	3.88 3.79 3.40	**5.15** 3.28 1.06	4.30 4.09 3.38	**9.05** **6.46** 2.98	**7.38** **6.86** 4.24
Deepseek	**10.83** **10.75** **5.25**	**15.98** 2.92 1.86	**7.11** **6.36** **6.08**	4.15 3.97 3.47	**7.23** 2.72 1.44	**10.41** **5.77** **5.64**	**11.50** **7.12** 2.66
Doubao	**30.22** **10.23** 0.14	**22.26** **8.04** 1.44	**15.19** 4.99 1.82	**22.78** 4.93 0.37	**15.17** **6.16** 1.65	**7.85** 2.36 1.86	**7.88** 2.38 1.88

Figure 17: Color Residuals by Model Series for the Weekday Group

Q M	A	B	C	D	E	F	G
Human	**19.62** 2.54 2.03	3.81 2.70 1.79	**7.93** 1.28 1.11	4.61 2.06 1.59	4.08 1.79 1.14	2.09 2.02 1.39	2.20 1.37 1.20
GPT	**34.30** 0.94 -0.56	**26.28** **13.63** -1.38	**14.45** **14.42** -0.46	**15.88** 4.42 0.73	**7.11** **6.97** **5.56**	**14.49** 2.73 1.82	**10.48** 2.34 1.58
Deepseek	**41.04** -0.20 -0.75	**35.23** 0.88 0.67	**27.80** **9.96** 0.31	**26.27** **15.15** -0.20	**23.68** **5.76** 0.88	**18.12** **12.26** 0.58	**14.51** 4.08 3.98
Doubao	**30.36** -0.78 -0.91	**25.25** **7.35** -0.37	**22.77** **9.63** -0.78	**34.71** -0.78 -1.05	**26.97** 1.74 0.57	**26.37** **5.31** -0.78	**13.40** **9.41** **5.94**

Figure 18: Color Residuals by Model Series for the Letters Group (from A to G)

Q M	H	I	J	K	L	M	N
Human	2.82 2.65 2.28	2.23 2.19 1.77	2.04 1.57 0.78	1.65 1.39 1.39	2.04 1.22 0.81	1.63 1.38 1.22	2.53 1.82 1.67
GPT	**7.27** 3.78 3.48	**20.38** 4.37 1.95	**5.24** 3.27 1.08	3.81 3.51 2.31	**8.28** **7.82** **7.57**	**5.69** 3.63 2.13	**5.99** 2.64 1.76
Deepseek	**20.82** 4.66 2.51	**23.65** **7.21** 2.76	**6.82** 3.27 2.90	**24.50** 3.52 0.41	**8.56** **5.22** 2.03	**6.29** 4.83 2.35	**6.76** **5.00** 4.03
Doubao	**15.28** **7.87** 2.15	**12.40** **7.87** **5.21**	**5.12** 4.94 3.63	**6.91** 4.68 0.83	**8.10** 2.70 2.15	**7.98** 4.94 2.37	**10.09** **5.85** 3.84

Figure 19: Color Residuals by Model Series for the Letters Group (from H to N)

M \ Q	O	P	Q	R	S	T
Human	**5.27** 2.34 0.61	2.04 1.82 1.16	1.72 1.25 0.86	1.00 0.88 0.82	2.30 1.44 1.41	2.46 2.04 0.73
GPT	**18.17** **7.59** -0.64	**14.49** **8.50** 4.66	2.92 2.63 2.44	**21.77** **7.27** 0.04	**14.22** 3.51 3.15	**8.15** **7.82** 3.02
Deepseek	**10.11** **7.75** 2.03	**8.90** 4.93 1.75	**6.02** **5.04** 3.42	**15.41** 2.76 1.62	**8.93** **5.79** 1.98	**8.09** **5.97** 1.24
Doubao	**20.92** **5.69** 2.06	**18.07** **11.41** 2.60	**18.07** **9.03** 4.72	**20.94** 2.60 0.46	**8.75** **6.15** 1.91	**6.10** **5.30** 4.35

Figure 20: Color Residuals by Model Series for the Letters Group (from O to T)

M \ Q	U	V	W	X	Y	Z
Human	3.60 1.59 0.67	2.40 1.20 0.92	1.34 1.13 0.88	2.53 1.73 1.36	1.97 1.48 1.45	3.98 1.49 1.41
GPT	**5.05** 3.67 3.22	**17.42** **11.41** -0.27	**32.44** 4.42 1.07	**10.41** **9.23** 3.61	**25.08** **7.09** 2.69	**12.08** **5.78** 1.59
Deepseek	**11.63** 2.38 1.78	**12.01** **8.71** 0.38	**15.15** 3.41 2.03	**18.76** 4.44 1.60	**13.94** 3.95 2.24	**16.03** 3.23 2.83
Doubao	**6.71** **5.17** 2.99	**19.40** 1.57 -0.12	**24.24** 2.24 1.91	**7.45** **7.32** 2.28	**27.11** 0.54 -0.39	**15.63** 2.85 2.32

Figure 21: Color Residuals by Model Series for the Letters Group (from U to Z)

M \ Q	Spring	Summer	Autumn	Winter
Human	**14.90** **6.53** 1.63	**13.42** 1.91 1.01	**14.69** **8.20** 0.53	**18.14** **5.24** **4.46**
GPT	**27.24** **10.09** 4.95	**17.12** **11.49** 0.00	**24.23** 2.08 1.73	**23.66** **14.93** **10.47**
Deepseek	**32.58** **5.50** -0.58	**18.53** **8.37** 3.72	**24.89** -0.58 -0.58	**18.08** **17.68** **15.50**
Doubao	**27.19** 2.45 2.36	**25.15** 2.45 0.07	**25.08** **7.12** 2.45	**17.01** **15.66** **11.03**

M \ Q	Past	Present	Future
Human	**5.73** **5.53** 2.34	2.26 2.08 1.92	3.13 1.39 1.39
GPT	**12.04** **6.85** 3.44	**9.15** **7.30** 4.32	**9.77** **8.39** 3.29
Deepseek	**14.52** **14.36** **13.42**	**17.55** **10.98** **7.91**	**19.89** **11.14** **6.40**
Doubao	**29.40** 3.17 2.00	**24.28** **8.28** 4.31	**24.79** **6.56** **5.33**

Figure 22: Color Residuals by Model Series for the Season and Time Group

M \ Q	Jan	Feb	Mar	Apr	May	Jun
Human	**6.43** 4.80 3.06	4.05 2.80 1.83	**8.98** 4.81 3.76	**6.81** 3.54 3.00	4.08 3.13 2.55	2.29 2.10 1.69
GPT	**14.67** **14.37** **5.77**	**23.53** **16.42** **13.78**	**26.21** 4.44 4.01	**14.03** **11.67** 3.34	**15.61** **5.09** 4.99	**10.36** **7.24** 2.28
Deepseek	**19.70** **11.52** **5.23**	**30.72** -0.28 -0.79	**34.30** 2.32 -1.09	**11.17** **8.76** 1.74	**5.62** 2.32 2.11	**8.99** **8.96** 4.92
Doubao	**25.60** 3.72 -0.07	**13.16** 3.44 2.00	**20.99** 1.95 0.42	**12.25** **8.07** 4.57	**8.67** **7.55** **7.23**	**11.55** 4.13 2.82

Figure 23: Color Residuals by Model Series for the Month Group (from Jan to Jun)

M \ Q	Jul	Aug	Sept	Oct	Nov	Dec
Human	**5.89** 4.98 1.32	**5.66** 3.79 1.66	**6.47** 1.97 1.93	2.54 1.66 1.18	**7.03** 3.13 3.01	**13.15** 4.19 4.01
GPT	**19.08** 0.30 -0.30	**12.98** 4.38 0.10	3.68 3.21 2.44	**14.00** **11.26** -0.30	**18.73** **7.48** 0.40	**15.51** **10.94** **10.76**
Deepseek	**12.62** 3.94 0.88	**9.52** 4.21 0.52	**8.10** 4.21 2.06	**10.13** **8.73** 4.72	**15.51** **10.83** **10.71**	**22.38** **13.25** **5.17**
Doubao	**12.83** 3.95 2.43	**10.05** 2.95 0.58	**6.20** **6.05** 2.43	**8.74** **5.87** 2.60	**9.37** **5.55** 2.43	**15.57** **6.34** 3.05

Figure 24: Color Residuals by Model Series for the Month Group (from Jul to Dec)

M \ Q	Up	Down	Front	Back	Left	Right	Center
Human	**12.03** 4.22 1.54	**6.21** 1.92 1.62	3.92 3.16 1.12	4.44 3.38 1.10	3.64 2.61 1.23	4.73 3.78 1.74	**7.20** 2.49 1.77
GPT	**19.49** **10.21** **5.46**	**19.67** **5.08** -0.33	3.50 3.44 2.72	**9.97** **6.71** 2.40	**15.81** 0.31 0.18	**5.86** 4.76 1.59	**37.52** 4.15 3.24
Deepseek	**26.53** **11.88** 2.38	**34.29** 3.09 2.82	**9.42** **5.26** 2.99	**10.73** **9.40** 1.90	**13.91** **7.39** 0.52	**7.72** 2.05 1.95	**32.39** **11.84** -0.60
Doubao	**24.36** 2.45 -0.34	**17.83** 2.20 0.22	**10.90** **5.66** **5.43**	**12.68** **8.36** 2.47	**24.59** 4.74 1.94	**8.52** **8.17** 2.95	**25.95** **12.68** 0.51

Figure 25: Color Residuals by Model Series for the Direction Group

M \ Q	E	W	S	N	SE	NE	SW	NW
Human	3.14 1.26 1.22	3.34 2.73 1.81	1.60 0.93 0.89	**5.01** 1.54 1.48	2.96 2.68 0.90	2.49 1.54 0.93	4.39 1.54 1.40	3.34 1.85 1.25
GPT	**17.92** 2.46 -0.33	3.97 3.51 2.66	**5.20** 4.64 4.36	**14.46** **6.64** 1.89	**12.36** 4.12 4.03	**14.80** **5.61** 3.44	**12.69** 3.49 1.41	**6.04** **5.39** 4.56
Deepseek	**8.55** **7.93** 3.64	**14.02** 4.91 1.84	**14.01** **10.67** 3.01	**7.10** **5.03** 4.63	**17.82** **8.37** **7.44**	**13.45** **7.19** **6.52**	**5.90** 4.31 4.11	**12.75** **9.73** **7.27**
Doubao	**14.20** **13.39** -0.34	**19.66** 1.76 0.43	**13.18** **6.36** 0.43	**23.19** 2.85 1.21	**18.68** **5.90** 4.53	**19.66** **7.58** 4.25	**11.18** **7.53** 3.14	**9.33** **8.17** **7.21**

Figure 26: Color Residuals by Model Series for the Direction Group

M \ Q	High	Low	Far	Near	Deep	Shallow	Sparse	Dense
Human	3.74 2.81 1.96	4.53 2.76 1.25	2.20 2.13 1.58	2.96 2.54 2.10	**11.23** 2.92 1.81	**9.25** 4.27 1.80	2.88 1.98 1.73	4.01 3.63 2.42
GPT	4.71 3.68 2.79	4.46 3.12 2.79	3.29 2.81 1.85	**14.87** 3.46 3.19	**28.34** 1.04 0.93	**14.37** 4.28 3.27	**11.42** **9.68** **7.98**	**22.17** **6.80** 4.75
Deepseek	**14.49** **9.54** 3.58	**20.20** 4.27 2.20	**13.70** **5.91** 1.67	**14.65** 3.43 1.29	**23.95** **6.42** 4.67	**11.66** **10.89** **5.39**	**9.57** **7.05** 4.51	**9.13** **8.52** 2.19
Doubao	**21.24** **20.22** 3.78	**13.31** **12.30** 2.24	**27.05** 3.77 1.98	**31.91** **8.20** 4.30	**17.68** 0.93 -0.53	**19.48** -0.36 -0.53	**13.45** 4.01 3.62	**16.54** -0.53 -0.69

Figure 27: Color Residuals by Model Series for the Abstract Concept Group

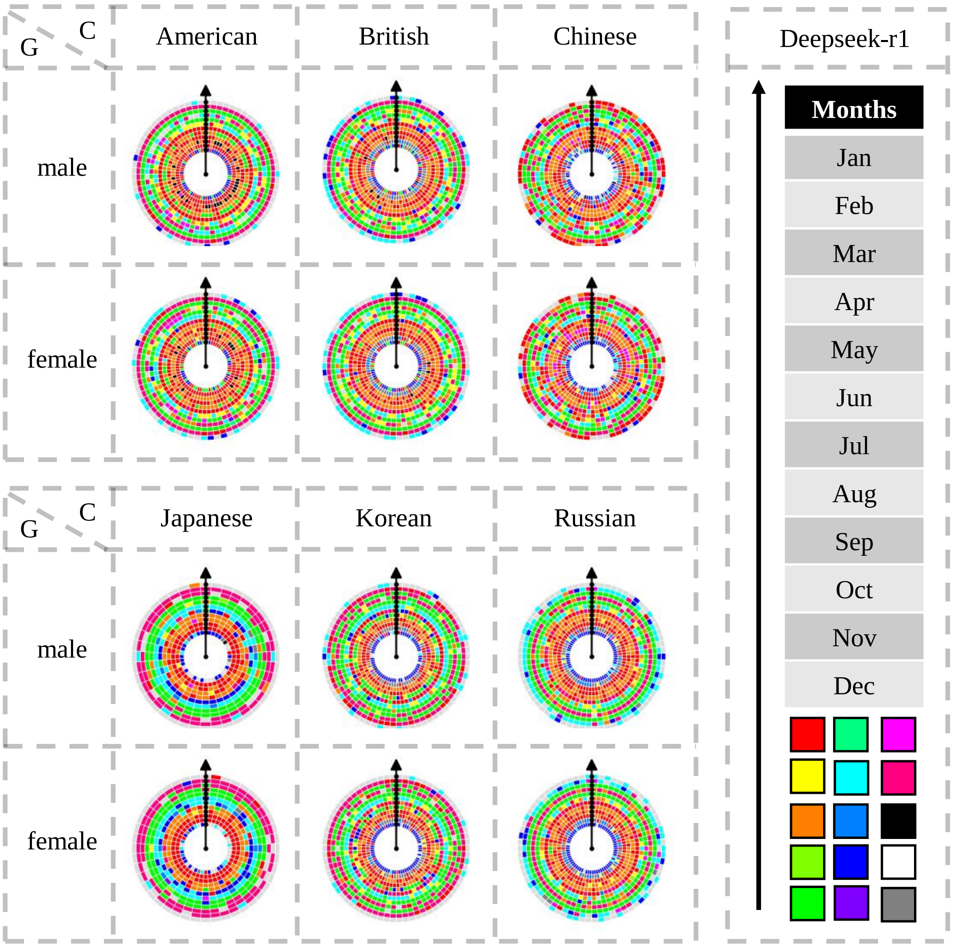

Figure 28: Culture-Specific Outputs of Deepseek-r1 on the "Months" Category

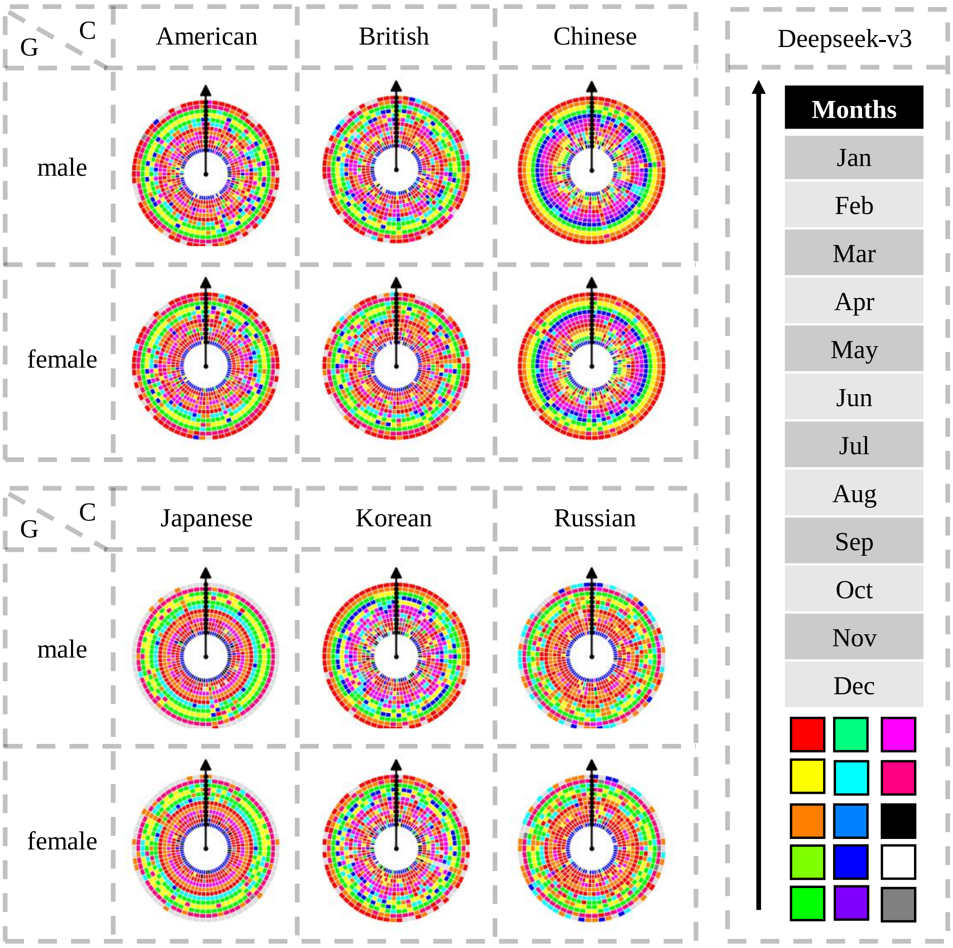

Figure 29: Culture-Specific Outputs of Deepseek-v3 on the "Months" Category

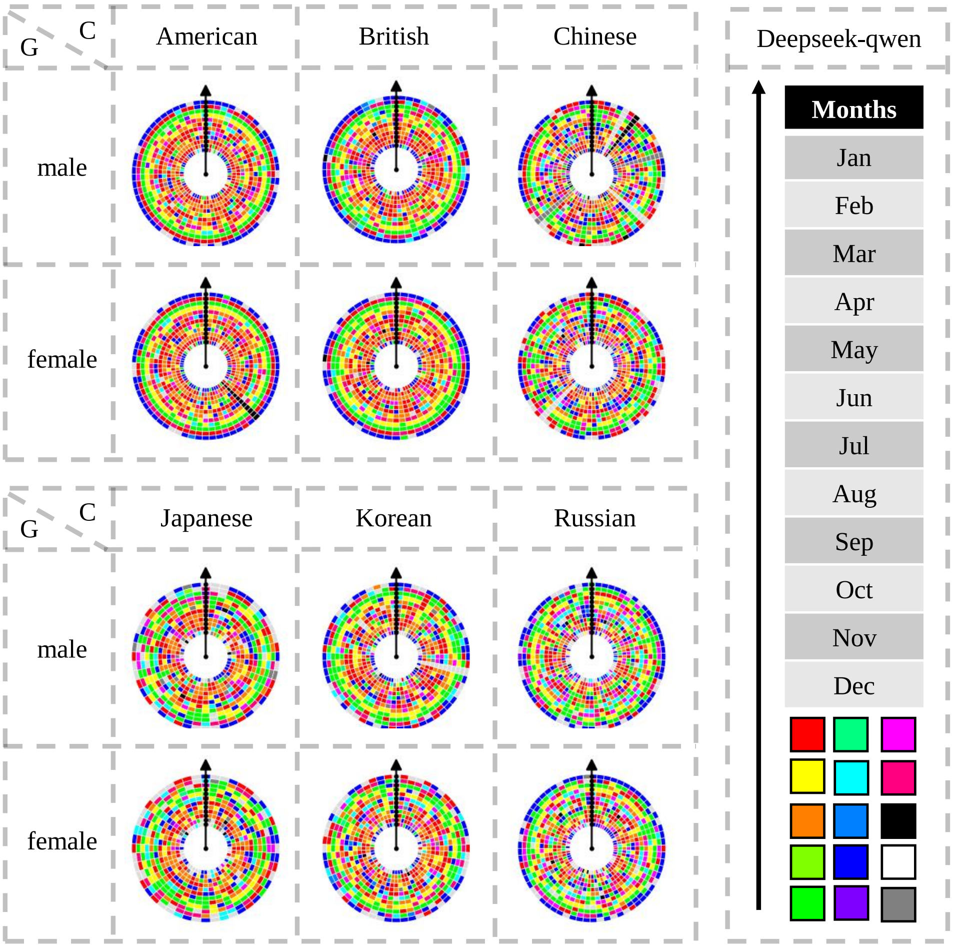

Figure 30: Culture-Specific Outputs of Deepseek-qwen on the "Months" Category

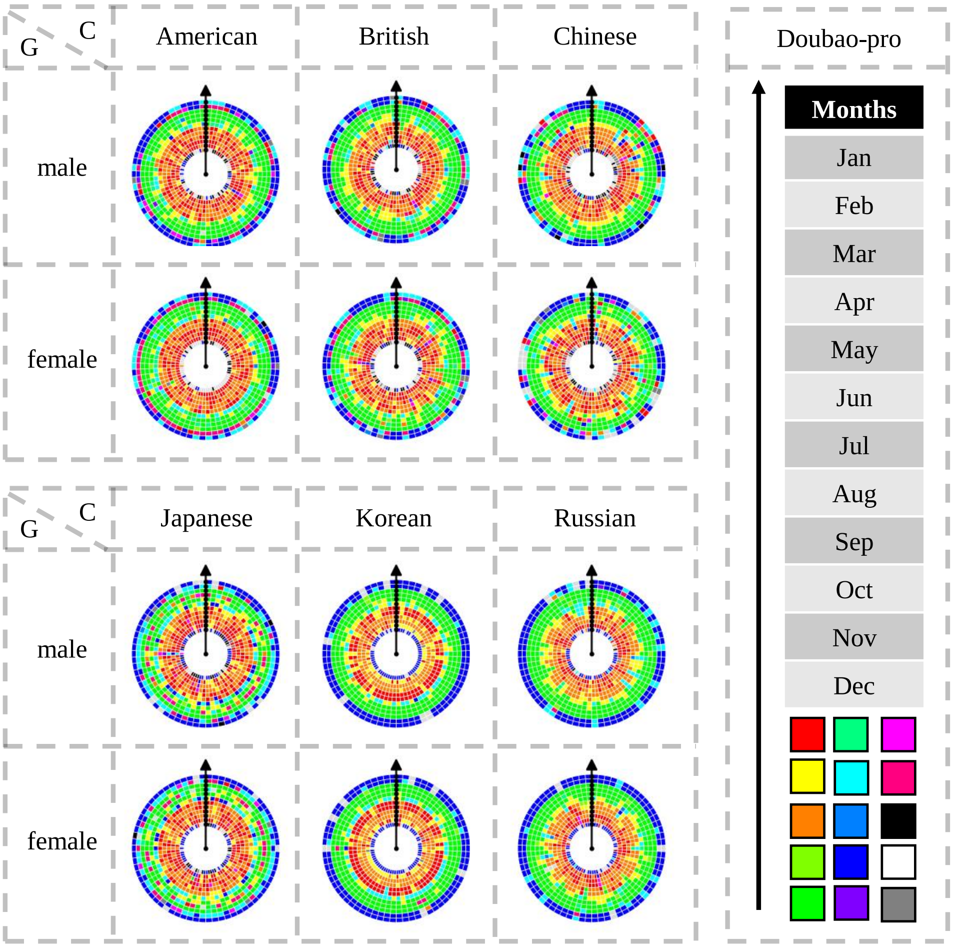

Figure 31: Culture-Specific Outputs of Doubao-pro on the "Months" Category

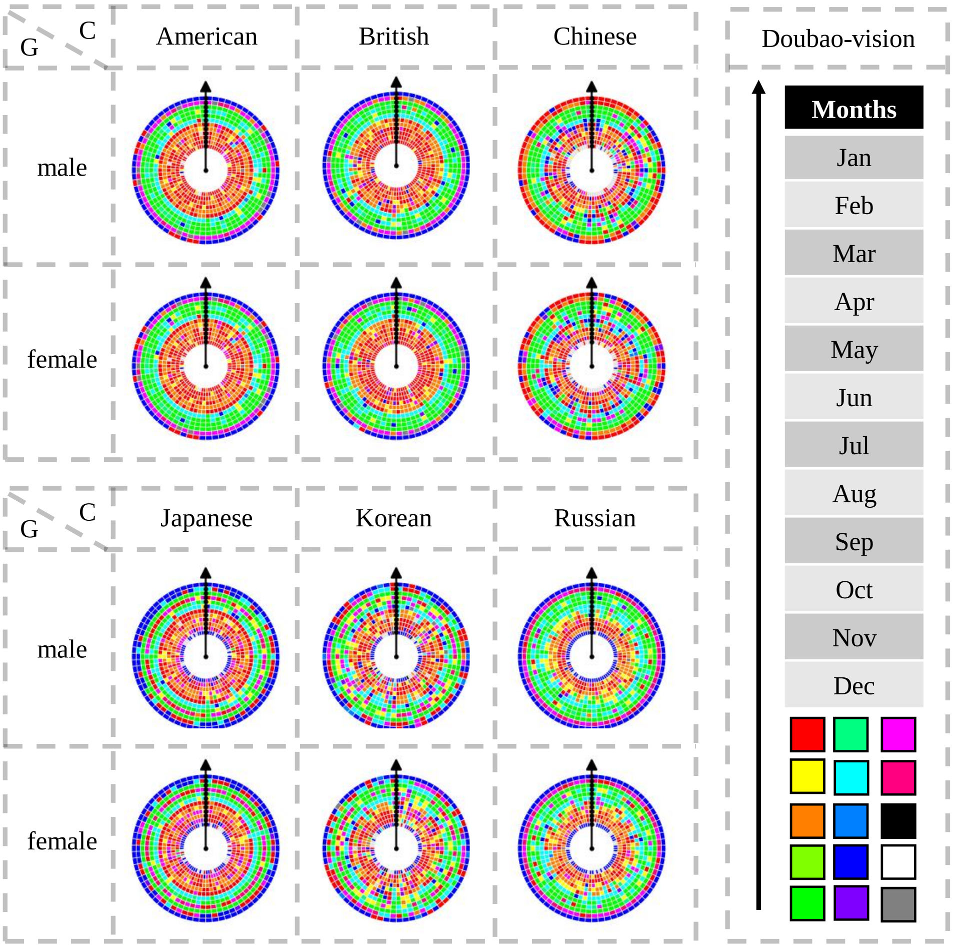

Figure 32: Culture-Specific Outputs of Doubao-vision on the "Months" Category

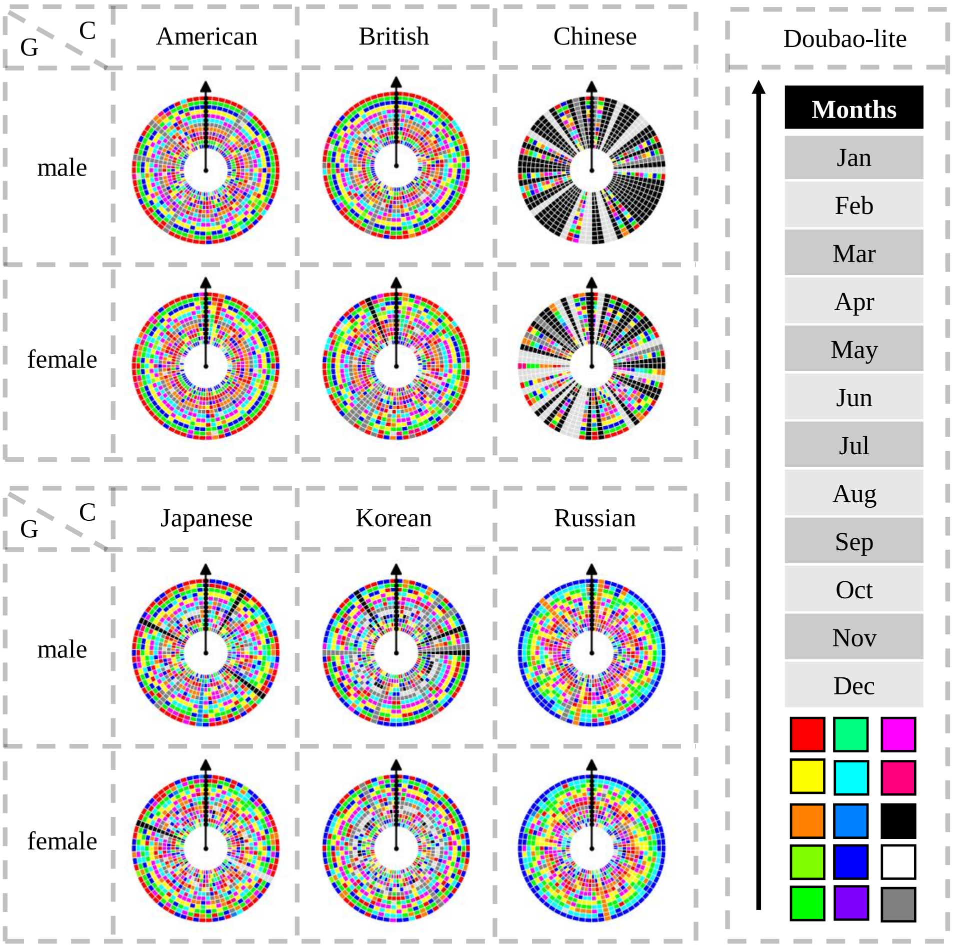

Figure 33: Culture-Specific Outputs of Doubao-lite on the "Months" Category

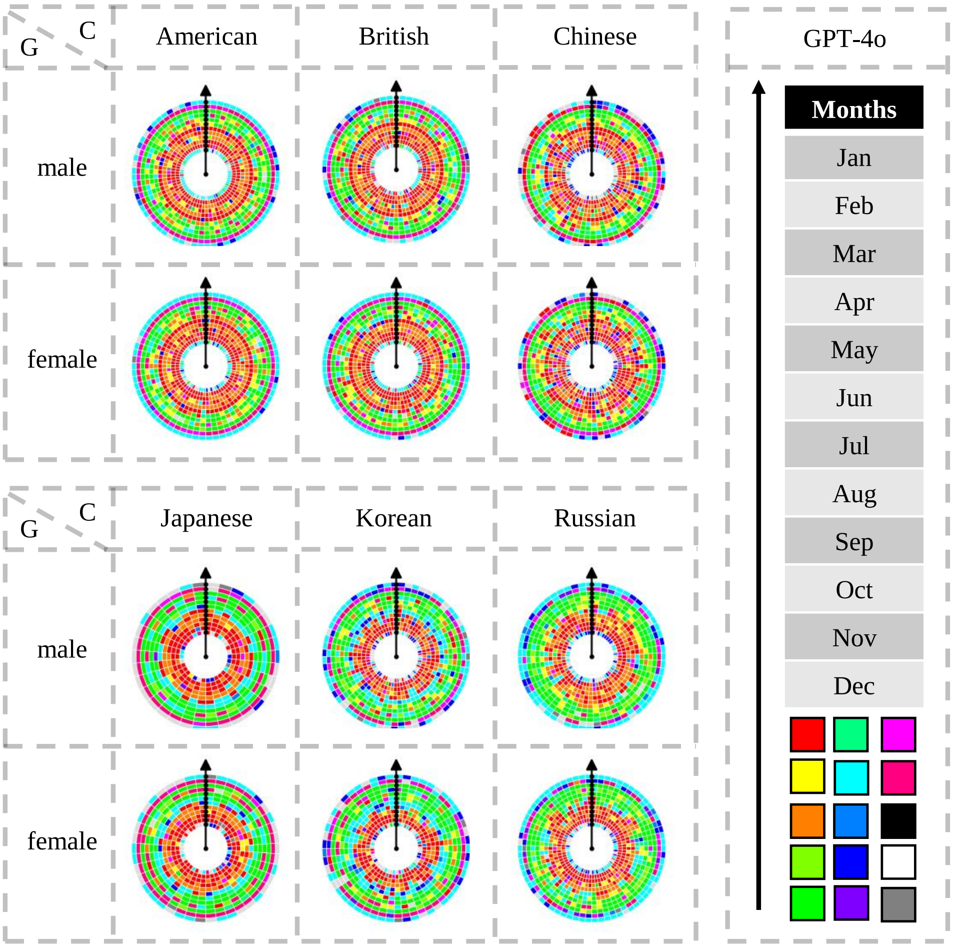

Figure 34: Culture-Specific Outputs of GPT-4o on the "Months" Category

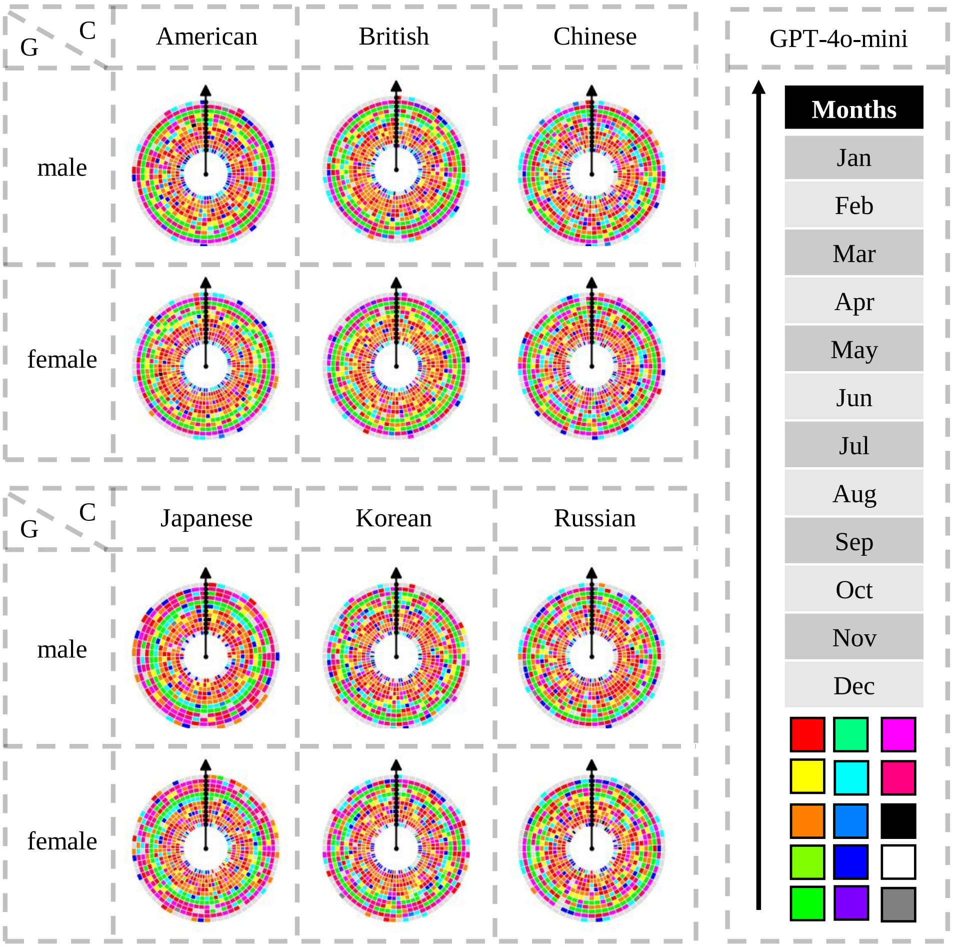

Figure 35: Culture-Specific Outputs of GPT-4o-mini on the "Months" Category