LLMs Can Feel But Not See? AI vs Human Categorical Continuous Learning

Mario Garrido (mhg9251@nyu.edu)

Yuan Huang (yh2741@nyu.edu)

Introduction

Understanding how humans classify and update their hypotheses about categories is a fundamental question in cognitive science. The problem of categorization—how an intelligent agent should group stimuli into discrete concepts—is a significant focus of psychological research. Categorization extends themes from Western classical thought (Aristotle, 1984) and has clear interpretations at multiple levels of analysis (Anderson, 1990). It is foundational to understanding human cognition and advancing artificial intelligence (Cohen & Lefebvre, 2005). The study of human categorization represents a significant achievement in cognitive psychology, particularly through statistical modeling approaches (Nosofsky, 1998)(Griffiths, Steyvers, & Tenenbaum, 2007). The field's central debate focuses on whether category learning relies on abstract summary representations (prototype models) or stored examples of category members (exemplar models). Prototype models propose that humans form and store abstract representations capturing central category tendencies, while exemplar models suggest that categorization decisions stem from comparing new stimuli to stored instances in memory. Laboratory research has generally favored exemplar models due to their superior ability to represent complex category boundaries and irregular patterns (Reed, 1972). This empirical support for exemplar models suggests that human category learning may depend more on storing specific examples than maintaining abstract summaries. However, evidence indicates that both mechanisms may operate in different contexts or learning stages (Smith & Minda, 1998)(Ashby & Maddox, 2005). Furthermore, most of this work has relied on simplified stimuli, which limits its applicability to real-world categorization tasks (Nosofsky, 1988). Human categorization abilities, by contrast, emerge from interactions with the natural world. Extending existing theory to account for behavior in ecologically valid domains remains a key challenge. For instance, (Battleday, Peterson, & Griffiths, 2019) emphasized the importance of studying human categorization using complex, real-world stimuli. Their research demonstrated that combining cognitive models with convolutional neural network (CNN)-derived representations effectively predicts human classification behavior.

Prototype models leveraging CNN representations align closely with human categorization of natural images, suggesting their utility in understanding shared group characteristics. Furthermore, (Malaviya, Sucholutsky, Oktar, & Griffiths, 2022) introduced the concept of less-than-one-shot learning, where participants form prototypes without direct exposure to examples, underscoring the flexibility of prototype-based approaches in data-scarce environments.

Our project focuses on how humans update their hypotheses about categories, aligning with the supervised learning framework in cognitive science. Feedback is crucial for guiding categorization:

- **Deterministic Feedback:** Providing explicit feedback about the correct category facilitates rapid learning and reduces uncertainty.
- **Eliciting Reasoning:** Asking participants to articulate their reasoning provides insights into their strategies for hypothesis updating and mental model refinement.

(Ashby & Valentin, 2018) highlighted the importance of timing and the nature of feedback, noting that immediate feedback enhances performance. Incorporating tasks that prompt participants to explain their decisions and adjust their beliefs after receiving feedback can reveal the incremental processes underlying hypothesis refinement.

Methodological Considerations Several methodological considerations are critical for studying categorization:

- **Stimulus Selection:** Ensure stimuli include diverse subjects (e.g., cats, dogs, humans) and emphasize perceptually salient shared characteristics.
- **Feedback Design:** Use deterministic feedback to evaluate participants' learning strategies and hypothesis updating.
- **Measuring Complexity:** Apply the Minimum Description Length (MDL) principle to quantify the complexity of shared characteristics within categories (Fass & Feldman, 2002).

Our project's comparison between human and Large Language Model (LLM) categorization necessitates identical experimental conditions. Using the same stimuli,

instructions, feedback mechanisms, and evaluation metrics allows for a direct assessment of classification accuracy, strategies, and learning patterns. (Battleday et al., 2019) highlighted the potential for integrating cognitive and computational models to bridge human and machine categorization, offering valuable insights into data efficiency and conceptual learning. By studying human categorization in ecologically valid contexts and comparing it to LLM performance, our project advances our understanding of categorization processes. Integrating insights from cognitive science and machine learning provides a robust foundation for exploring how humans and machines classify complex visual information, contributing to broader inquiries into the nature of intelligence.

Methods and Models

Experimental Design

We employ a novel experimental design to recruit 10 subjects who will complete two sets of image categorization tasks using real-world stimuli. These tasks are designed to examine both perceptual and conceptual categorization processes:

- **Set 1 - Perceptual Categorization Task:** Subjects identify shared patterns based on concrete observable facts and details present in the images. For instance, objects, structures, etc...
- **Set 2 - Conceptual Categorization Task:** Subjects identify shared patterns that require interpreting the overall meaning and emotional impression conveyed by the images. For instance, shared emotions, implied context, etc...

Image Classification Task Each subject will perform two trials in succession. One for set 1, and then one for set 2.

For each set, subjects are presented with 9 images. 3 belonging to the A group, 3 belonging to the B group, and 3 belonging to neither (as a control/counterexamples). Images in a given group have a shared characteristic that is not present outside that group.

Set	Group A	Group B
Example	Dogs in image	Sad image
1	Text in the image	Wearing white shoes
2	Victory/Satisfaction after hard work/suffering	Abandoned/Lonely

Table 1: Ground Truth Categories for the example and sets.

Before the first set, subjects are given an example 3×3 grid with revealed ground-truth categories in order to explain the task and demonstrate both the existence of perceptual and conceptual patterns. [1]

At the beginning of each set, subjects are given time to examine and hypothesize the exclusive shared conditions of the set's groups A and B. Then, they are presented with 6 novel images in succession (in a random order). Two of these belong to group A, two to group B, and two to neither group. Subjects are tasked to answer for each:

1. What group they believe the current image belongs to.
2. Their reasoning.

Immediately after, subjects are given feedback. They are informed of what the correct group for the image is. Finally, they are given more time to examine all images and are asked to explain how their understanding of the groups' characteristics changed. An example for this process can be seen in table 2.

Round	Participant	Group A Understanding	Group B Understanding
Beginning	GPT-4o	Famous/iconic elements	Dynamic motion/activity
Round 1	GPT-4o	Printed/displayed text	Athletic/sports activity
Round 2	GPT-4o	Printed/displayed text	Professions requiring uniforms
Round 3	GPT-4o	Printed/displayed text	People in motion
Round 4	GPT-4o	Printed/displayed text	Active lifestyle/sports
Round 5	GPT-4o	Printed/displayed text	Active lifestyle/sports
Round 6	GPT-4o	Artistic/visual appeal	Active lifestyle/sports
Beginning	Human	Happy	Stuff after used, no alive people
Round 1	Human	Happy under static mode	Stuff after used, no alive people
Round 2	Human	Happy under static mode	Stuff after used, no alive people
Round 3	Human	Happy under static mode	Stuff after used, no alive people
Round 4	Human	Win after done something	Stuff after used, no alive people
Round 5	Human	Win after done something	Abandon it after use
Round 6	Human	Win after done something	Abandon it after use

Table 2: Comparison of Understanding Between GPT-4o and human subjects across rounds, showing an example of how their interpretations of Groups A and B changed over time.

(a) Set 1 (b) Set 2

Figure 1: Categorization tasks' example images.

Research Focus

Our experimental design aims to measure:

- **Classification Accuracy**: How well subjects classify images into the correct groups as the trial goes on.
- **Hypothesis Updating**: How subjects refine their beliefs about group characteristics based on successive new images and episodic feedback.
- **Reasoning Analysis**: The qualitative reasoning behind classification decisions, including the explicit articulation of shared patterns.

[1] As shown in Figure 12 in the Annex.

Comparative Analysis with LLMs

To benchmark human performance, we developed two LLM-based systems that work on the exact same inputs as the human subjects. One is designed to keep the exact description of previous images in its memory (exemplar model) and the other generates a prototype given the experimental constraints, and aims to update it after receiving feedback from each image. This system does not have access to previous images when deciding on a new image. Rather, it creates a prototype for each group when presented with the 3×3 grid, and updates it when experiencing a new image. Only the latest prototype will be used as contextual information when deciding on a new image.

The main difference in implementation other than this system, is that the exemplar model is a naive, 1:1 recreation of the experiment on human subjects. On the other hand, the prototype model has prompting that makes the model first identify all valid patterns across the image grid as an initial prototype, and then only discard and rewrite based on new examples. This makes the prototype model an inherently subtractive system, as no information gets added apart from what already was in the first prototype.

Both kinds of systems were tested with different base LLMs. We used Claude 3 Haiku, Claude 3.5 Sonnet V2 (Anthropic, 2024), and GPT-4o (OpenAI, 2023). Every model and system (either exemplar or prototype) combination was tested on 10 trials per set.

Comparison methods The systems are evaluated and compared with human answers using the following methods:

- **Change in understanding over 6 rounds** - For each round, we measure how closely the subject's understanding matches the true category definition. This is done by:
  1. Converting both the participant's description and the ground truth category definition into semantic embeddings.
  2. Calculating the cosine distance between these embeddings.
  3. Tracking how this distance changes across the 6 rounds.

  A lower distance indicates that the participant's understanding is more aligned with the true category definition.

- **Classification Accuracy** - For each test image, we check if the participant/model assigned it to the correct category (A, B, or Neither). The accuracy score represents the percentage of correct classifications out of all test images. This provides a straightforward way to compare human and AI performance on the same categorization task independently of the groups' definition hypotheses.

- **Similarity Map Visualization** - We create a 2D map that shows how similar or different the stated group defining characteristics are between different subjects for the same group/set. Using t-SNE – a dimensionality reduction technique – (Van der Maaten & Hinton, 2008) on the description's embeddings, descriptions that are more similar appear closer together on the map, while those that are more different appear farther apart. This helps visualize how similar human and AI hypothesized categories are for a given ground truth category.

- **Understanding Evolution Score** - Measures how much participants' category descriptions change between rounds. By comparing the embeddings of consecutive descriptions, we can quantify:
  - How drastically participants update their understanding after receiving feedback.
  - Whether humans and AI models show different patterns in how they refine their understanding.
  - At what point their understanding stabilizes.

This design allows us to explore not only the correctness of classifications but also the cognitive processes underlying hypothesis refinement, providing a robust framework for comparing human and machine category continuous learning in complex, real-world stimuli.

Results

Human performance

Distance from ground-truth Figure 2 shows how human participants' understanding of the groups tends to approach the ground truth more as the rounds go on. Some groups have immediately more obvious categories (Text in the image) while others seem to be more difficult across the board (Victory/Satisfaction after hard work/suffering). The 95% confidence intervals around the trend lines show that there is a noticeably large spread of values among human participants. However, the trend is for distance between the internalized category and ground truth to diminish, so there is a clear learning process.

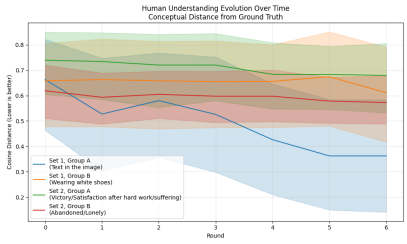


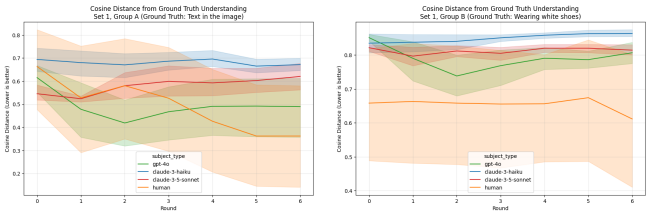
Figure 2: Human Understanding Evolution Over Time Conceptual Distance from Ground Truth

Exemplar Model-Based System Performance

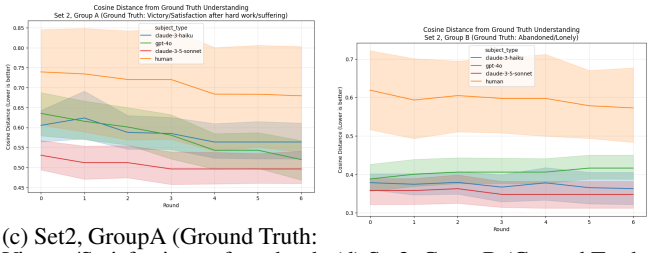
Semantic distance from understanding to ground-truth For the exemplar model system, Figure 3 shows the comparisons between different language models and human participants, measuring how close the hypothesis of the group characteristic gets to the ground truth. For **perceptual**

categorization tasks, humans have a noticeable advantage over language models. For this kind of task, language models seem to follow the same kind of trend indistinctly of capability or architecture. Even when using a temperature value of 0.7 across the board, humans also seem to have a far wider spread of values when compared to LLMs.

In the case of **conceptual categorization**, there is a reversal in this pattern. LLMs seem to strongly outperform humans in these kinds of tasks. In our experience, multimodal LLMs have a weakness when the task requires noticing details that take up a small portion of the image (for instance, white shoes). But they are far more likely to describe the mood and compositional choices in images, so they can leverage these patterns to reason correctly with a higher probability.



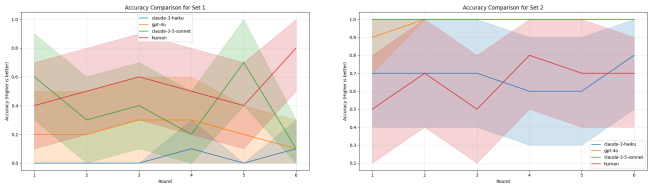
(a) Set1, GroupA (Ground Truth: Text in the image) (b) Set1, GroupB (Ground Truth: Wearing white shoes)



(c) Set2, GroupA (Ground Truth: Victory/Satisfaction after hard work/suffering) (d) Set2, GroupB (Ground Truth: Abandoned/Lonely)

Figure 3: Cosine Distance from Ground Truth Understanding between exemplar-based systems and human subjects

Accuracy Figure 4 shows the categorization category over the 6 rounds for the different types of subjects on sets 1 and 2. There is an agreement between the trend in these graphs and the trends shown in the previous section, as humans outperform LLMs consistently for the **perceptual categorization** tasks in set 1. For the case of the **conceptual categorization** in set 2, human performance seems to be slightly below LLM performance in general. It is noticeable that for the 10 trials run on each LLM system, both GPT-4o and Claude 3.5 Sonnet achieved a perfect accuracy on conceptual categorization after at most 1 round.

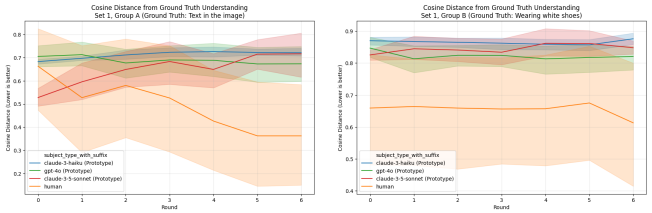


(a) Set 1 (b) Set 2

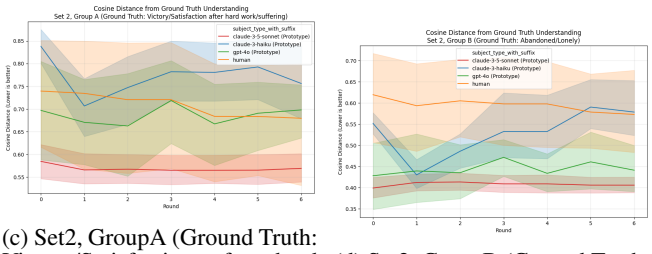
Figure 4: Accuracy Comparison between Human and Exemplar-agents

Prototype Model-based agent Performance

Semantic distance from understanding to ground-truth Figure 5 shows that modifying the system to use a prototype-based model causes the distance to the ground truth to increase consistently across all groups. This is consistent with empirical results discussed in the introduction, and it would appear that this modification makes the LLM system less aligned to human semantic understanding learning for this task.



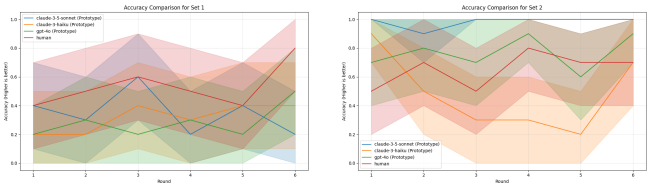
(a) Set1, GroupA (Ground Truth: Text in the image) (b) Set1, GroupB (Ground Truth: Wearing white shoes)



(c) Set2, GroupA (Ground Truth: Victory/Satisfaction after hard work/suffering) (d) Set2, GroupB (Ground Truth: Abandoned/Lonely)

Figure 5: Cosine Distance from Ground Truth Understanding between Prototype-Systems and Human subjects

Accuracy Figure 6 shows that in general, performance in set 2 dropped considerably across all models when using the prototype based model. GPT-4o seems to be the most affected. Surprisingly, accuracy in set 1 seem to have been unaffected. If anything, it went up. We will discuss this in detail in the next section.

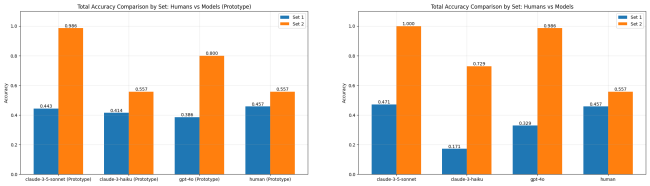


(a) Set 1 (b) Set 2

Figure 6: Accuracy Comparison between Human and Prototype-agents

Comparing human and LLM-system accuracy

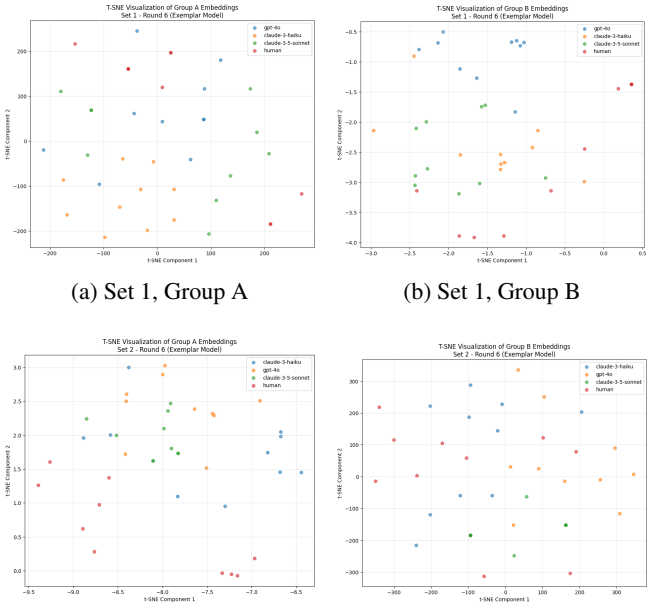
Figure 7 shows that although the prototype-based systems tend to perform worse on set 2 (conceptual categorization tasks), the accuracy in set 1 (perceptual categorization tasks) increases or stays basically the same across all models, and it matches human subjects' performance. This result is striking since from figures 3 and 5 it can be seen that the prototype models' understanding of groups tends to be more distant from the ground truth than in the exemplar case. This would imply that the prototype model was capable of generating alternative understandings of the groups to humans and exemplar models that brought their performance in perceptual categorization tasks up even though the alignment with ground truth dropped. Whether this is a result of decisions in prompt engineering changes between the two systems, or something inherent to the prototype model structure is a potential area of research for future work.



(a) Humans and Prototype-agents (b) Humans and Exemplar-agents

Figure 7: Total Accuracy Comparison by Set

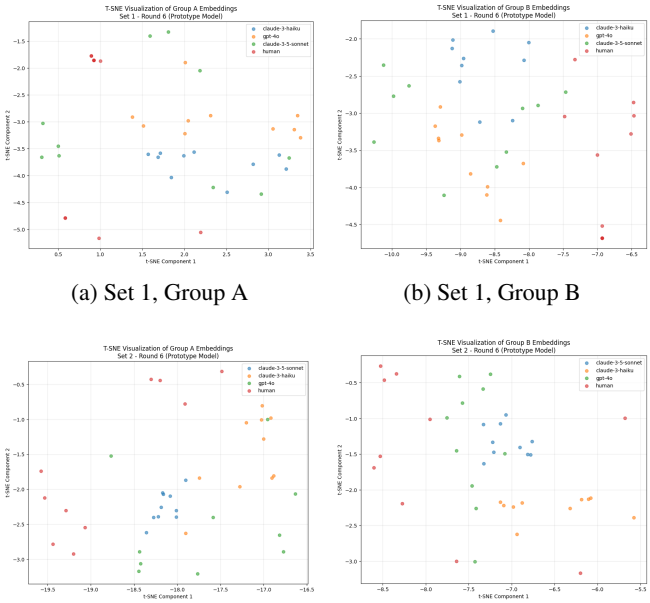
Embedding space similarity Figures 8 and 9 compare the t-SNE visualizations of the final rounds' group understanding embeddings for each group, divided by prototype or exemplar model. A pattern that arises when comparing both is that in the exemplar model, there seems to be a more uniform spatial distribution of the latent space representations when compared to the prototype model. This would imply that in general, the answers from the exemplar model are semantically more similar to humans' answers. Most of the maps for the prototype models show more spatial separation between human points and the LLM systems' points.



(a) Set 1, Group A (b) Set 1, Group B

(c) Set 2, Group A (d) Set 2, Group B

Figure 8: t-SNE Visualization of Embeddings of Humans and Exemplar-agents



(a) Set 1, Group A (b) Set 1, Group B

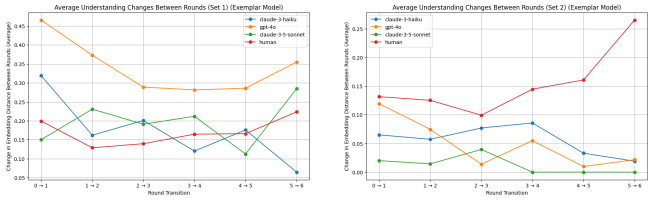
(c) Set 2, Group A (d) Set 2, Group B

Figure 9: t-SNE Visualization of Embeddings of Humans and Prototype-agents

Update information and reasoning over rounds between humans and agents

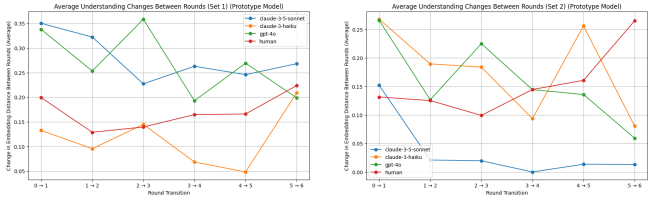
Figures 10 and 11 show the average change in distance between the embeddings of the group understanding. That is, how far on average is the change in semantic meaning from a given round to the next, after taking in a new example. It

is noticeable that conceptual change between rounds tends to be consistently higher in perceptual categorization tasks (set 1) than in conceptual categorization tasks (set 2). The biggest change that can be noticed is that the concept change dynamics of LLM based systems on set 2 changes radically when switching from an exemplar-based system to a prototype-based system. In exemplar systems, models seem to be less aligned with the rate of learning in humans, either too much change or too little. On the other hand, on prototype-based systems, LLMs seem to be closer in update rate to human subjects. The reason why LLMs seem to change their concepts more often on this system, is most likely because of the information bottleneck that is introduced by generating a prototype. This bottleneck will cause the current understanding to not fit the observed examples with more likelihood, which would make it tend to change more.



(a) Set 1 (b) Set 2

Figure 10: Average Understanding Changes Between Rounds for Humans and Exemplar-agents



(a) Set 1 (b) Set 2

Figure 11: Average Understanding Changes Between Rounds for Humans and Prototype-agents

Discussion

Our study of human and machine learning categorization processes reveals several important areas for methodological refinement and future research directions. The findings suggest both limitations in our current approach and promising avenues for advancement.

Methodological Considerations

Subject Recruitment and Sampling The current study's reliance on convenience sampling through friends and relatives introduces potential bias, despite achieving diversity in demographics. While our subjects represented varied backgrounds in terms of gender, race, culture, age, and occupation, the non-random selection limits the generalizability of our findings. Future iterations of this research should implement systematic randomization protocols aligned with established Human Subject Research guidelines.

Stimulus Selection and Presentation Our current image selection criteria rely on two categories: one based on perceptual facts or detailed patterns (e.g., wearing white shoes or text in the image), and the other requiring deeper conceptual understanding (e.g., emotions or abstract characteristics like "Victory/Satisfaction after hard work/suffering" or "Abandoned/Lonely"). However, our selection process was somewhat arbitrary, as we did not account for all factors and information contained in the images. For future improvements, we will adopt a more rigorous selection process to ensure that all stimuli are equally clear and representative of their intended category.

The randomized presentation of test images, while intended to prevent order effects, revealed unexpected consequences. Variations in information density across images led to performance disparities depending on the sequence. Early failures particularly impacted subject motivation and subsequent performance. Future designs should incorporate balanced information distribution across test sequences and consider adaptive ordering strategies.

Experimental Design Refinements

Sequential Testing Effects The current design's sequential presentation of Set 1 (perceptual) followed by Set 2 (conceptual) introduced potential carryover effects. Subjects' experiences with perceptual categorization in Set 1 appeared to influence their approach to conceptual categorization in Set 2, sometimes leading to overcorrection toward surface-level features. Future studies should consider between-subjects designs for different categorization types to minimize such interference.

Feedback Mechanisms In this study, we provided deterministic feedback (i.e., exact group information) after each response. We are also interested in exploring probabilistic feedback or feedback that only indicates right or wrong without specifying the correct group. During the experimental design phase, we found that subjects struggled significantly to discern patterns without correct group information. To evaluate probabilistic feedback effectively, future studies will require a larger pool of test images and extended learning periods for participants.

Future Research Directions

Individual Differences During testing, we observed variability in performance across subjects—some performed well on Set 1 but struggled with Set 2, and only two out of ten subjects performed equally well on both sets. We also collected data on subjects' MBTI personality types, though we did not analyze this in the current study. This information may be useful as a control variable for analyzing individual performance in future research.

Model Architecture Enhancement In this project, we used various LLMs to construct two agent architectures: the Exemplar model and the Prototype model. While these approaches were effective for our purposes, CNNs remain a crucial technique for image categorization tasks. In future work, we aim to integrate CNNs into our framework and compare their performance against both human participants and the LLM-based models. This comparison could provide deeper insights into the relative strengths and limitations of each approach.

Continuous Learning Assessment Understanding the differences in continuous learning between humans and machines is another one of our areas of interest. In future work, we could introduce a third set of images that combines both perceptual and conceptual categorization tasks. This would allow us to analyze how quickly humans and machines identify common patterns after being exposed to the first two sets.

Conclusion

Our study illuminates key aspects of human category learning through a novel experimental paradigm that leverages LLM performance as a comparative baseline. The findings reveal that humans excel in perceptual categorization tasks, demonstrating superior ability to identify and learn from fine-grained visual details. This suggests robust human capabilities in pattern recognition based on concrete, observable features—a fundamental cognitive skill that has likely been shaped by evolutionary pressures (Mattson, 2014).

Conversely, the relatively lower human performance in conceptual categorization tasks, especially when compared to LLM systems, highlights interesting constraints in human abstraction processes. This performance gap may reflect the cognitive cost of maintaining and updating complex, abstract category representations, particularly in time-constrained settings with limited examples.

The continuous learning patterns observed in our human subjects, characterized by gradual refinement in perceptual tasks but more variable progression in conceptual ones, provide insights into how humans naturally build and revise category representations. These findings contribute to our understanding of human category learning mechanisms and suggest potential approaches for designing systems that more accurately reflect human learning patterns in categorization tasks.

References

Anthropic. (2024). *Claude 3.5 sonnet.* `https://www.anthropic.com`. (Large language model)

Aristotle. (1984). *Categories* (J. Barnes, Ed. & J. L. Ackrill, Trans.). Oxford: Oxford University Press.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 5, pp. 1–41). John Wiley & Sons.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2019). Capturing human categorization of natural images at scale by combining deep networks and cognitive models. *arXiv preprint arXiv:1904.12690*.

Cohen, H., & Lefebvre, C. (Eds.). (2005). *Handbook of categorization in cognitive science*. Amsterdam: Elsevier.

Fass, D., & Feldman, J. (2002). Categorization under complexity: A unified mdl account of human learning of regular and irregular categories. In S. Becker & S. Thrun (Eds.), *Advances in neural information processing systems.*

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Malaviya, M., Sucholutsky, I., Oktar, K., & Griffiths, T. L. (2022). Can humans do less-than-one-shot learning? *arXiv preprint arXiv:2202.04670*.

Mattson, M. P. (2014). Superior pattern processing is the essence of the evolved human brain. *Frontiers in Neuroscience*, *8*, 265.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65.

Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to maddox and ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 322–339.

OpenAI. (2023). *Gpt-4o.* `https://www.openai.com/`. (Large language model)

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(86), 2579–2605.