Automating psychological hypothesis generation with AI: when large language models meet causal graph (Appendix)

Song Tong [1,2,3,4,*], Kai Mao [5,*], Zhen Huang [2], Yukun Zhao[2,†], Kaiping Peng[1,2,3,4,†]

1. Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China
2. Positive Psychology Research Center, School of Social Sciences, Tsinghua University, Beijing, China
3. AI for Wellbeing Lab, Tsinghua University, Beijing, China
4. Institute for Global Industry, Tsinghua University, Beijing, China
5. Kindom KK, Tokyo, Japan

∗ Equal technical contribution

† Correspondence: {zhaoyukun, pengkp}@tsinghua.edu.cn

Appendix A

Method

Article Selection & Cost Analysis.

Before delving into the specifics of the extraction process, it's crucial to emphasize the importance of cost analysis, especially when dealing with large-scale data processing. In the realm of data-intensive research, the sheer volume of articles and the intricate nature of text extraction entail significant computational and financial resources. Understanding the associated costs is paramount not only for optimal resource allocation but also for ensuring the scalability and feasibility of the project. With these considerations in mind, extracting causal knowledge from texts requires the utilization of language models, such as GPT-4, to process each paper. Given the costs associated with API usage, we projected the total expenses for different corpus sizes. Key determinants of the costs include: Token Count and API Constraints. Total number of tokens (word segments) across all texts. GPT-4 charges per thousand tokens for both inputs and outputs. In addition, GPT-4 capped at 60 requests per minute and 150k tokens per minute at the time of this research. Extraction procedures must comply with these thresholds. For our curated subset of 140k articles, with each abstract at 500 words and main content at 5,000 words, the estimates are 40 million tokens for 40k articles. The GPT-4 pricing is approximate to 40,000 USD. Based on this analysis, our choice was to extract 43,312 articles, representing around 40 million tokens. This strikes a balance between comprehensive coverage and cost-efficiency.

Differentiating Causality and Correlation by GPT-4

A experiment is conducted to investigate the capability of GPT-4 for differentiating 'causality' and 'correlation'. The experiment involved four graduate students (1 male, mean age 31 $\pm$ 10.23), each well acquainted with a collection of six psychological articles. This familiarity was intended to mitigate the inherent difficulty in assessing the presence of causality or correlation within academic texts. From an initial set of 21 articles, after excluding amenable to effective processing by PyPDF and excluding review articles, GPT-4 extracted 238 causal and 51 correlational concept pairs. Detailed methodologies and GPT prompt are provided in Section 'Causal knowledge extraction method' and Table 2. This resulted in an average of 11.33 pairs of causal concepts per article (SD = 7.20).

The students as evaluators were then surveyed to assess these pairs of identified concepts within the articles they provided, assigned with categorizing each as 'existent or not' and as 'causal or correlational', based on the descriptions of the articles. This unique perspective aimed to leverage their detailed understanding of the content to validate extracted concept pairs. The evaluation results are in Table A1. The preliminary statistical analysis of the 289 relationships highlighted that 87.54% (253/289) were recognized as existing, with 74.31% (188/289) classified as causality by evaluators. Notably, when GPT-4 identified a concept pair as causality, 86.98% (207/238) of the relationships were acknowledged, with 65.55% (156/238) agreed upon as causality, and only 13.02% (31/238) potentially not mentioned in the papers.

Table A1

Expert evaluation of relationships extracted by GPT-4 from their provided papers.

	Causality #	**Correlation #**	**Total**
GPT-4			
Relationship Extracted	238	51	289
Expert Evaluated			
Relationship Exist	207	46	253
Causality	*(156)*	*(32)*	*(188)*
Correlation	*(51)*	*(14)*	*(65)*
Nonrelationship	31	5	36

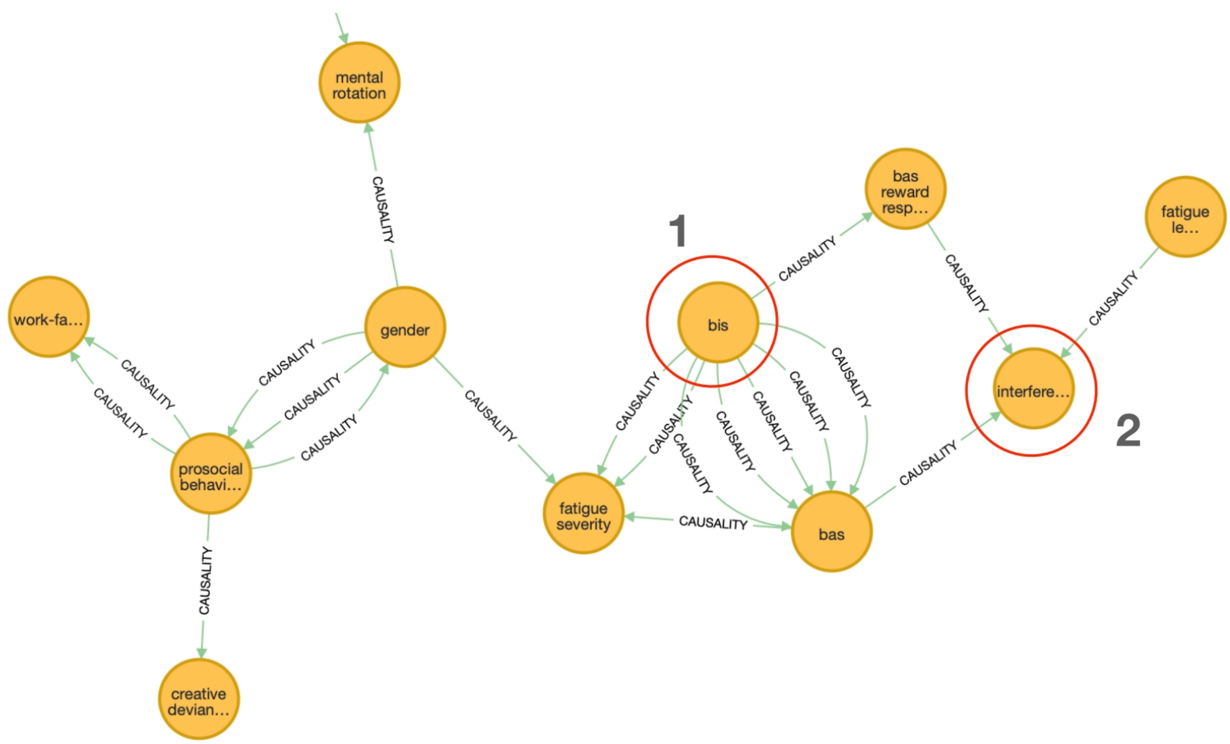


Figure A1. The illustration of hypothesis generation by link prediction.

Appendix B

Results

Details for Topic Analysis

Word cloud comparison. Figure B1 showcases visual representations of term and theme frequencies for different models and groups. A preliminary analysis suggests:

Control-Human (a): This word cloud seems to emphasize terms related to individual well-being and psychological health. Notable terms include 'relationship', 'happiness', 'self', and 'experience'. The presence of 'robot' and 'AI' suggests topics about technology's relationship with human psychology was atrracted by Ph.D students.

Control-Claude (b): The themes here seem to be oriented around positivity and growth. Key terms such as 'will', 'increase', 'greater', and 'positive' stand out.

Random-selected LLMCG (c): The terms in this word cloud underscore social connections, individual autonomy, and competence. Words such as 'social', 'individual', 'autonomy', and 'competence' are dominant. Themes of satisfaction, resilience, and cultural aspects can also be deciphered.

Expert-selected LLMCG (d): Here, the emphasis seems to be on community, personal feelings, and shared experiences. 'Support', 'sense', 'one', 'we', and 'social' are recurrent terms, highlighting collective experiences and social interconnectedness.

Connection graph analysis. The connection graphs in Figure B2 depict relationships between various themes and concepts for different groups.

Control-Human (a): The graph for this model suggests a notable interplay between artificial intelligence themes, such as 'Robot Companionship' and 'AI generating music/classic music', and human well-being factors like 'Heart rate variability (HRV) and electrodermal activity measures' and 'Life quality based ESM data'. This suggests research or perspectives focusing on how AI, robotics or algorithms can impact and measure human well-being.

Control-Claude (b): This graph emphasizes different facets of well-being, from

'Emotional Well-being' to 'Workplace Well-being'. It also considers both positive elements, such as 'Growth mindset' and 'Shared novel experiences', and potential challenges, like 'Reducing anxiety symptoms'. This indicates a holistic view of well-being.

Random-selected LLMCG (c): There's a strong focus on societal and structural determinants of well-being, such as 'Economic condition and financial hardship', 'Autonomy/Competence', and 'Management of health-related issues'. It seems to highlight the broader environmental and cultural factors affecting individual well-being.

Expert-selected LLMCG (d): This graph reflects more nuanced interconnections between personal experiences and environmental factors. Notable themes include the 'Living in walkable, mixed-use neighborhoods', 'Exposure to nature', and the 'Integration of all influences into empowerment'. It suggests a focus on how diverse life experiences, settings, and exposures can interplay to shape an individual's well-being.

Deep Semantic Analysis on Hypothesis Examples

Figure B4 illustrates the grouping of various cases shown in Table B4, highlighting the proximity of hypotheses C4 and C8 within the deep semantic space of BERT (conner, bottom left). Both C4 and C8 involve interactions with technological entities that provide social support, suggesting a thematic overlap and highlighting the impact of social support mechanisms on well-being. Expanding this group to include C3, which also revolves around the theme of narrative-based therapeutic recovery, suggesting a broader category of technologically mediated psychological support. In contrast, the distance between the C7 and C4 and C8 group could indicate that their thematic links are less direct, which introduces an entirely different element, humor, combined with mindfulness, representing another separate avenue to improve well-being.

Table B1

Inquiry on hypothesis generation for PhD students.

Dear PhD students, I would greatly appreciate it if each of you could propose at least three original research hypotheses related to well-being. The specific requirements are as follows: Each hypothesis should involve well-being. That is, well-being (e.g. life satisfaction, positive emotions, psychological health, *etc.*) should be either the independent variable, or the dependent variable, or the mediator in the research hypotheses you propose. Please ensure the research hypotheses you submit are novel. Related personnel will check for duplication upon receiving the hypotheses. If the duplication check finds that identical research questions have already been investigated in prior publications, it will unfortunately have to be considered unsatisfactory. In such cases, I would need to kindly request you resubmit with different hypotheses. The subsidy for this task is 100 RMB per person.

Algorithm 1 Compute intra-group pairwise distances

Require:
- F_g^i: The 2D t-SNE feature representation of the i^{th} idea in group g, derived from the respective high-dimensional embedding.
- G: The set of all groups, with g representing an individual group.

Ensure:
- D_g: A dictionary containing lists of pairwise distances for each group g.

Initialize D_g as an empty dictionary
for each group $g \in G$ **do**
    Let N_g be the number of data points in group g
    Let $F_g^1, F_g^2, ..., F_g^{N_g}$ be the feature representations of the ideas in group g
    Initialize list $L_g = []$
    for $i = 1$ to N_g **do**
        for $j = 1$ to N_g, $j \neq i$ **do**
            Let $d_g^{i,j} = \sqrt{(x_g^i - x_g^j)^2 + (y_g^i - y_g^j)^2}$ ▷ Compute the Euclidean distance between F_g^i and F_g^j
            Append $d_g^{i,j}$ to L_g
        end for
    end for
    Set $D_g = L_g$ ▷ Store pairwise distances for group g in D_g
end for
return D_g

Table B2

Detailed prompt for hypothesis generation in Claude-2 model.

Please generate original research hypotheses.The specific requirements are as follows: 1. Each hypothesis should involve well-being. That is, the variables used to measure well-being (e.g. life satisfaction, positive emotions, psychological health, etc.) should be either the independent variable, or the dependent variable, or the mediator in the research hypotheses you propose. 2. In your research hypothesis, it is essential to include the independent variable and the dependent variable, as well as the directional relationship between them. If possible, include mediating variables. Other variables, such as moderating variables, are not mandatory. 3. Please ensure the research hypotheses you submit are novel. That means you should conduct this task in two steps. After generating hypothesis, please check for duplication before submitting them. If the duplication check finds that identical research questions have already been investigated in prior publications, replace them with different novel hypotheses. 4. Please output in format for csv file, which is a table with columns for the independent variable (IV), dependent variable (DV), the relationships between IV and DV (either positive or negative), mediator, moderator, and description of the hypothesis in natural language. The columns for mediators and moderators can be left blank if your hypothesis doesn't involve them.

Table B3

Evaluation criteria for hypotheses by expert reviewers.

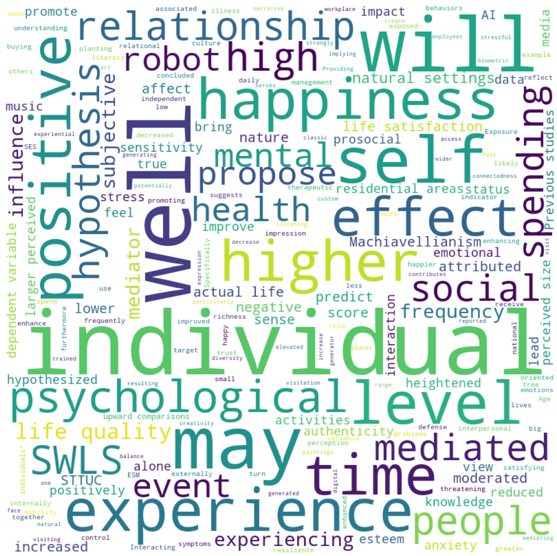
Dear Reviewers:

Thank you for participating in this evaluation process. In the table provided, you will find various psychological research ideas. We kindly request you to evaluate each hypothesis based on two key criteria: Usefulness and Novelty.

Novelty: This criterion focuses on the originality and uniqueness of the hypothesis. **How new or unprecedented is the concept?** A hypothesis that introduces fresh perspectives or challenges existing views would score high for novelty.

Usefulness: Please rate the practical relevance and applicability of each hypothesis. **Would it be beneficial for researchers, practitioners, or the general public?** A hypothesis that provides actionable insights or has potential to drive meaningful change would be rated high on this criterion.

Please provide **a score (1-10)** for each criterion for every hypothesis. Your expert opinions will significantly aid in understanding and prioritizing these hypotheses.



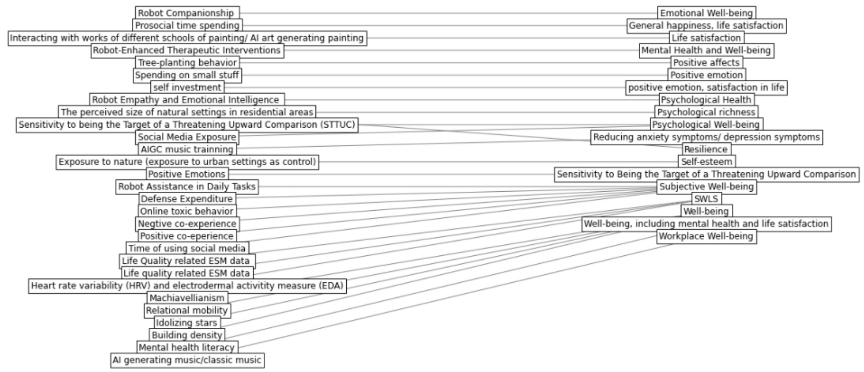
(a) Control-Human

(b) Control-Claude

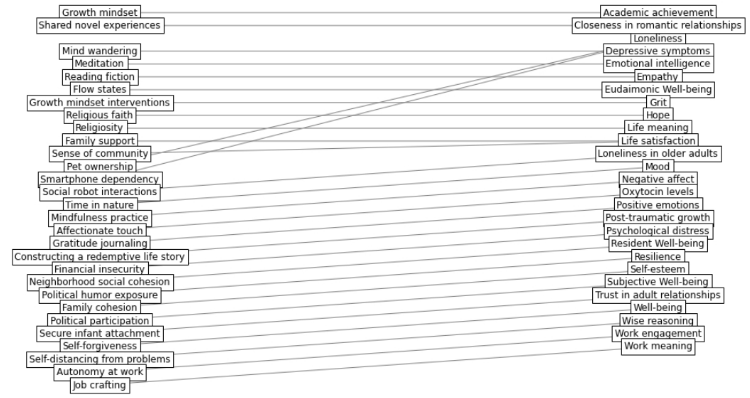
(c) Random-selected LLMCG

(d) Expert-selected LLMCG

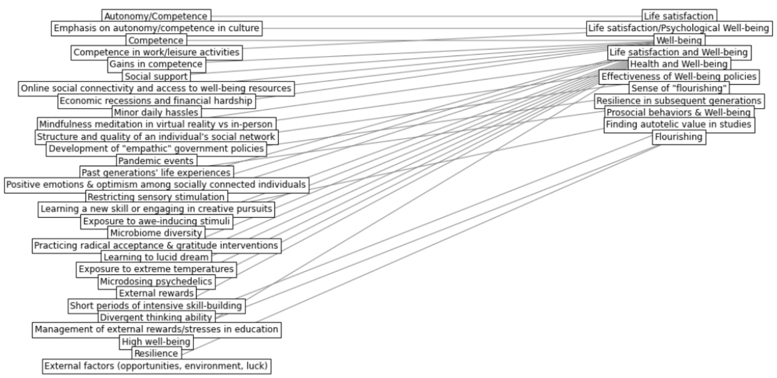
Figure B1. Word cloud comparisons of different models and groups.



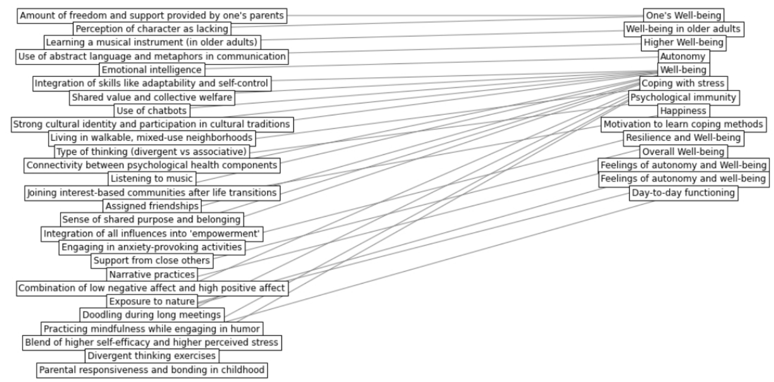
(a) Control-Human



(b) Control-Claude



(c) Random-selected LLMCG



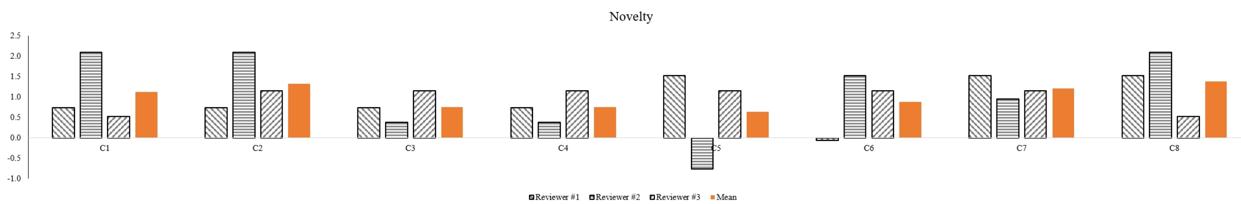
(d) Expert-selected LLMCG

Figure B2. Conceptual linkages and interactions across different models and groups.

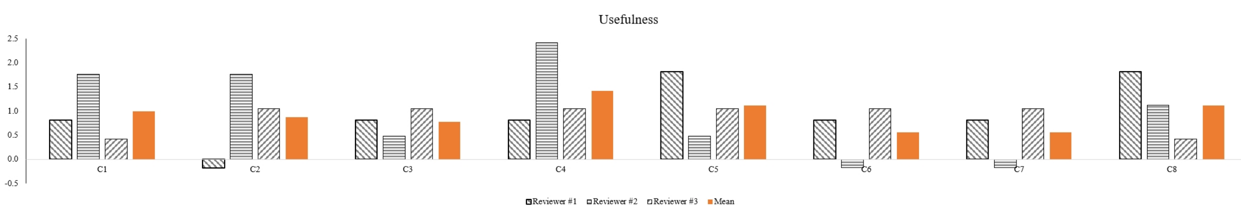
Table B4

The innovative hypotheses from different groups.

Case #	Research hypotheses	Novelty	Usefulness
Control-Human			
C1	Individuals interacting with robots that exhibit empathy and emotional intelligence will experience improved psychological health, and this relationship will be mediated by their enhanced emotional regulation capabilities.	1.119	0.997
C2	Listening to music generated by an AI trained on one's biometric data/narrative of self will lead to greater felt understanding and social connectedness, resulting in reduced anxiety symptoms.	1.330	0.875
Control-Claude			
C3	Creating a coherent narrative where negative experiences lead to positive outcomes will facilitate post-traumatic growth by enabling meaning-making.	0.758	0.777
C4	Interacting with social robots will help reduce loneliness in older adults by providing perceived social support.	0.758	1.422
Random-selected LLMCG			
C5	Virtual resilience: Online social connectivity and access to well-being resources can build 'virtual resilience' and enhance well-being during stressful events like pandemics.	0.641	1.109
C6	Exposure to awe-inducing stimuli increases prosocial behaviors and well-being through activating similar neural circuits.	0.876	0.561
Expert-selected LLMCG			
C7	Practicing mindfulness while engaging in humor (e.g., laughing yoga) will have an additive effect on well-being. Combining mindfulness techniques with humor and laughter may enhance the benefits of each practice on well-being. The positive effects of mindfulness and humor could build on each other.	1.212	0.561
C8	Use of chatbots can fulfill the need for communication and enhance connectedness, thereby promoting well-being in socially isolated individuals.	1.383	1.114



(a) Novelty



(b) Usefulness

Figure B3. Comparisons of (a) novelty and (b) usefulness scores among different reviewers of the hypotheses shown in Table B4.

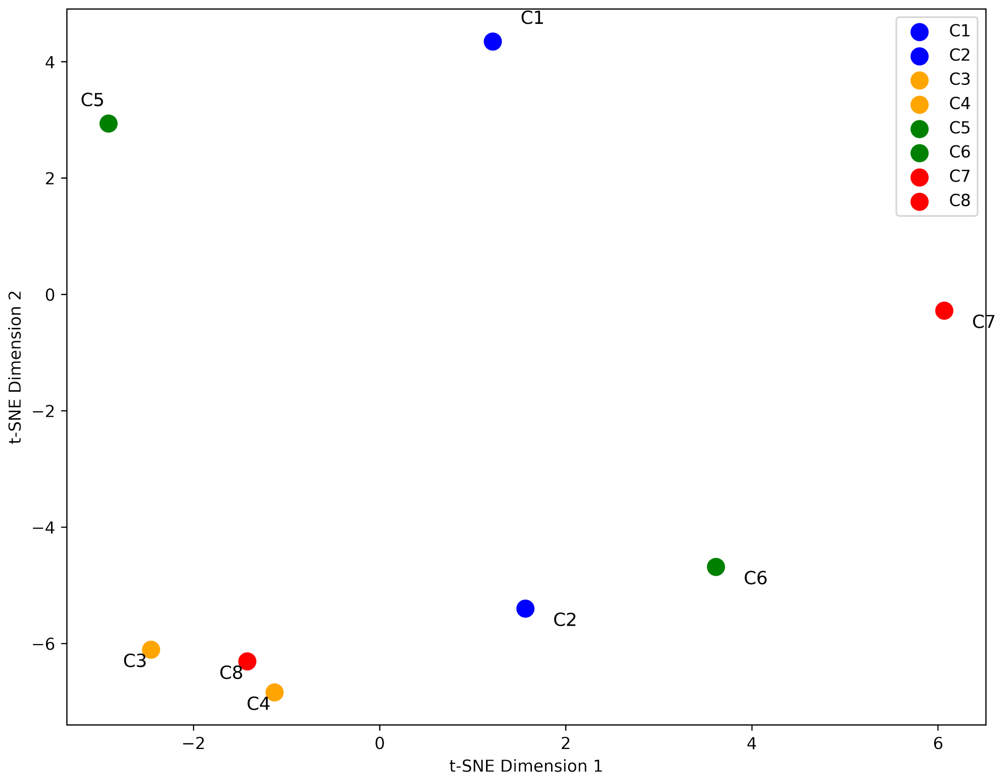
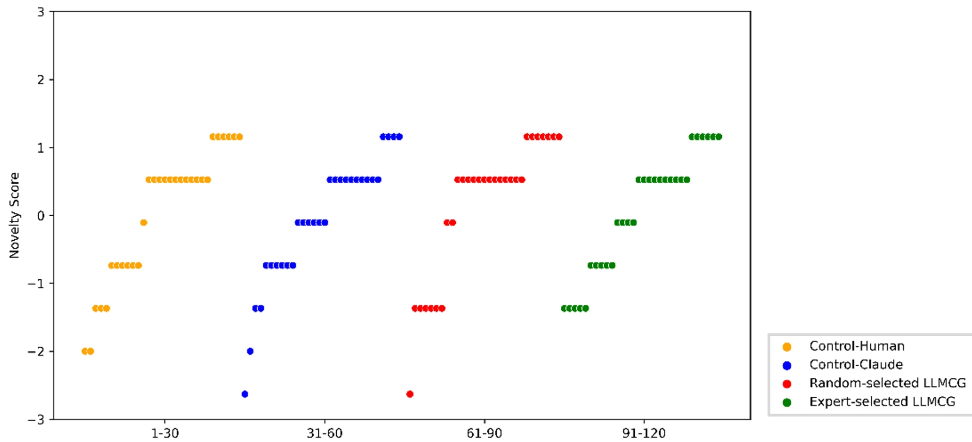


Figure B4. The t-SNE visualization of semantic representations for the hypotheses shown in Table B4.

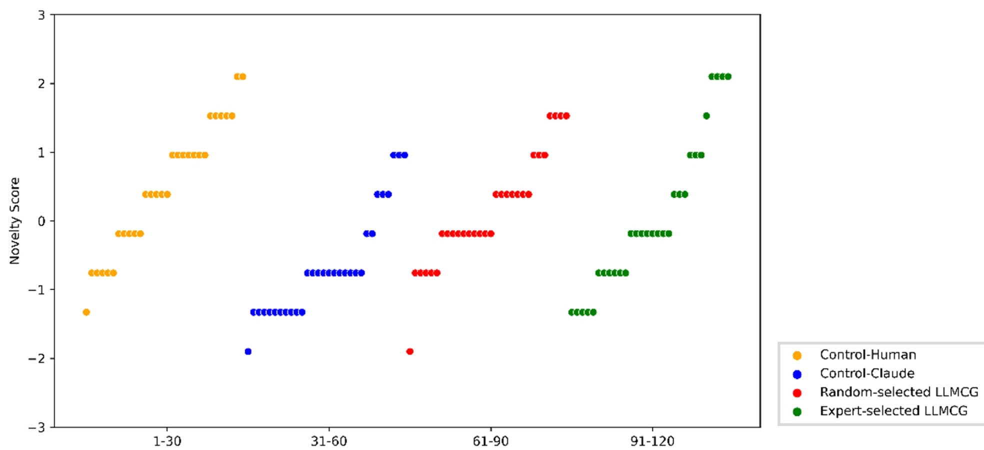
Table B5

Inter-rater reliability analysis for novelty and usefulness.

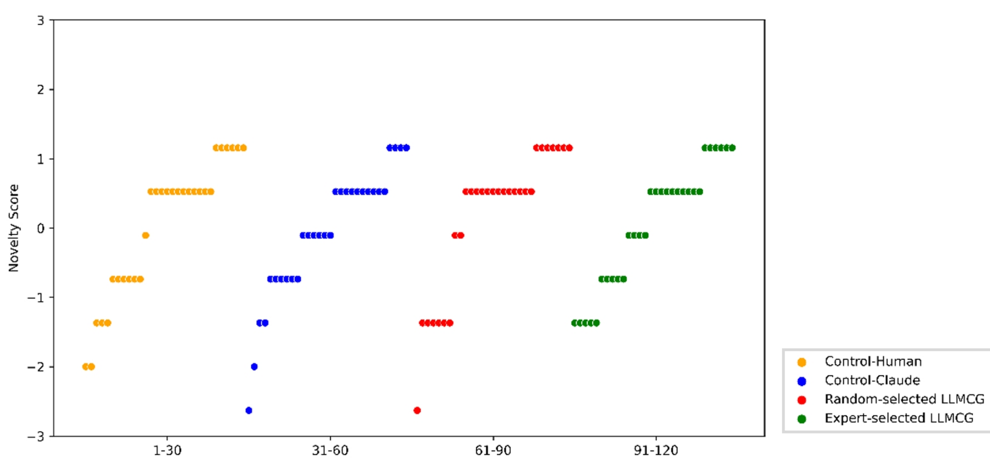
Comparison	**Spearman** r	p**-value**
Novelty		
Reviewer #1 vs. Reviewer #2	0.387	< 0.001
Reviewer #1 vs. Reviewer #3	0.172	0.059
Reviewer #2 vs. Reviewer #3	0.069	0.453
Usefulness		
Reviewer #1 vs. Reviewer #2	0.177	0.053
Reviewer #1 vs. Reviewer #3	0.221	0.015
Reviewer #2 vs. Reviewer #3	0.376	< 0.001



(a) Reviewer #1



(b) Reviewer #2



(c) Reviewer #3

Figure B5. Comparisons of novelty scores among different reviewers.

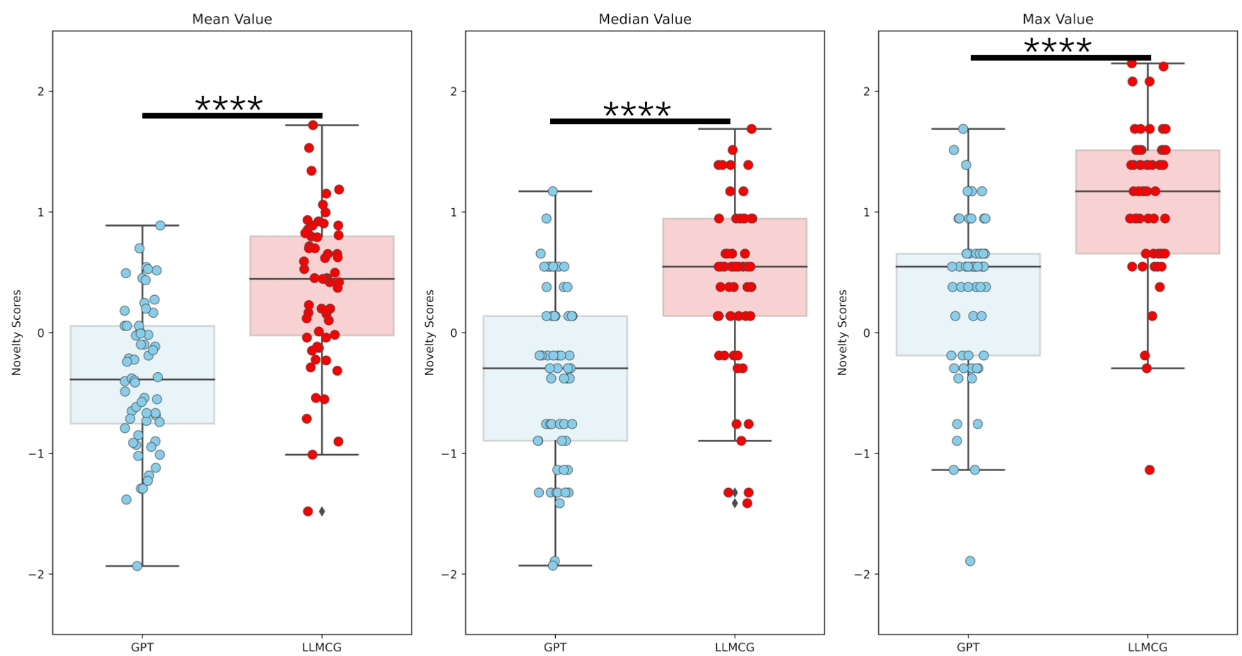


Figure B6. Comparative analysis between the GPT-4 and LLMCG groups. **Note:** $****$ denotes $p < 0.0001$.