




CatLLM: A Python package for Generating, Assigning, and Scoring Open-Ended Survey Data and Images

Chris Soria ^{1*}

¹ University of California, Berkeley, United States * These authors contributed equally.

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 08 June 2025

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

The rapid advancement of large language and vision models has created new opportunities for automated text and image analysis in social science research (Sachdeva & Nuenen, 2025; Schulze Buschoff et al., 2025; Yang et al., 2024). Researchers increasingly use these tools to code open-ended survey responses, categorize qualitative data, and analyze visual content at scale. However, challenges remain due to inconsistent output formats, diverse API interfaces, and the lack of standardized workflows for integrating model outputs into traditional statistical analysis pipelines (Rossi et al., 2024). CatLLM addresses these issues by providing a modular framework with specialized functions that enable researchers to work with consistent data structures across text and image analysis workflows while maintaining compatibility with standard statistical analysis tools.

Statement of need

Social scientists increasingly recognize the value of open-ended survey input for capturing rich, nuanced responses that closed-ended formats cannot provide. However, many researchers avoid incorporating open-ended input into their surveys due to the substantial analysis challenges they present. The processing of open-ended responses is notoriously time-intensive, requiring manual categorization and careful interpretation that can quickly become overwhelming with large datasets. Even when researchers do include open-ended questions, quantitative researchers often fail to fully utilize the resulting qualitative data due to limited time, resources, or expertise in analysis techniques. This analysis burden not only increases research costs but also creates practical barriers that prevent researchers from leveraging the deeper insights that open-ended responses can provide.

Current solutions present several limitations for academic researchers analyzing open-ended survey data. General-purpose natural language processing libraries such as NLTK require significant programming knowledge and often involve complex workflows for custom model training, while tools like spaCy, though more user-friendly, still require domain expertise for specialized applications. Commercial platforms like Dedoose or Atlas.ti focus primarily on manual coding workflows and lack integration with modern language models. While some researchers have begun using large language models (LLMs) directly through web interfaces, this approach lacks standardization, reproducibility, and systematic output formatting necessary for quantitative analysis.

CatLLM addresses these gaps by providing a standardized, free-to-use, interface for applying state-of-the-art language and vision models to common research tasks without requiring machine learning expertise. The package enables researchers to transform qualitative data into quantitative datasets suitable for statistical analysis, bridging the gap between traditional qualitative methods and computational approaches. Recent research demonstrates that LLMs

from OpenAI and Anthropic, particularly GPT-4, can effectively replicate human analysis performance in content analysis tasks, with some studies showing LLMs achieving higher inter-rater reliability than human annotators in sentiment analysis and political leaning assessments (Bojić et al., 2025). Unlike existing tools, CatLLM improves reproducible, structured outputs while supporting multiple AI providers and maintaining cost efficiency through built-in optimization features.

Table with 5 columns: Survey Response, Financial, Family, Housing Features, New Job. It shows how three different survey responses are categorized into these five areas with binary scores (0 or 1).

Figure 1: Example of CatLLM Assigning Categories to Move Reason Survey Responses

The software has demonstrated practical impact across diverse research domains. It has been successfully applied in studies examining demographic differences in LLM performance using the UC Berkeley Social Networks Study (Soria, 2025), categorizing occupational data according to Standard Occupational Classification codes, and implementing automated scoring for cognitive assessments in the Caribbean-American Dementia and Aging Study (Llibre-Guerra et al., 2021). These applications demonstrate the package’s versatility in addressing real-world research challenges that require systematic analysis of unstructured data at scale.

The package can be easily installed and implemented:

Code snippets for installing the package using pip and importing it in Python.

For comprehensive documentation and detailed installation instructions, see https://github.com/chrissoria/cat-llm.

Features

The CatLLM package processes user-provided text (open-ended survey responses) or image data and returns structured data objects. The package enables users to customize function behavior by incorporating their specific research questions and background theoretical frameworks, allowing the language models to generate more contextually relevant and theoretically grounded outputs tailored to their analytical objectives.

The package extends this framework through specialized capabilities:

- Binary Image Classification: Applies classification frameworks to vision models, determining the presence or absence of specific categories within images for systematic visual content analysis.
Flexible Image Feature Extraction: Extracts diverse data types from images, returning numeric, string, or categorical outputs rather than limiting analysis to binary classifications, enabling more nuanced visual data collection.
Drawing Quality Assessment: Compares user-generated drawings against reference images, producing quality scores based on similarity metrics for objective evaluation of visual reproduction tasks.

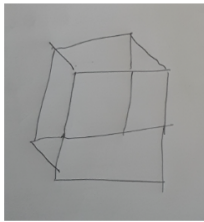
- **Standardized Cognitive Assessment Scoring:** Implements established CERAD protocols (Fillenbaum et al., 2008) for scoring geometric shape drawings, calculating standardized scores based on the presence of required visual elements for neuropsychological evaluation.
 - **Corpus-Level Theme Discovery:** Identifies and ranks themes across large text collections by systematically analyzing random corpus segments, extracting recurring topics, and prioritizing themes based on their frequency and consistency across different sections.
- This modular approach provides researchers with consistent data structures across text and image analysis workflows while maintaining compatibility with standard statistical analysis tools.

Picture Column

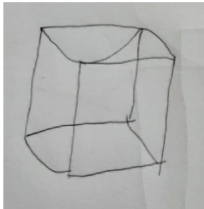
Score Column



0



2



3

Figure 2: Scoring Drawings of Cubes According to CERAD Rules Using CatLLM

Acknowledgements

This work was supported by the UC Berkeley Mentored Research Award. The author thanks Henry Tyler Dow for assistance in testing the functions on real data. The author also acknowledges the University of California, Berkeley for providing the institutional support that enabled this research.

Partial support was provided by the Center on the Economics and Demography of Aging, P30AG012839, and the Greater Good Science Center's Libby Fee Fellowship.

References

Bojić, L., Zagovora, O., Zelenkauskaitė, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., & Jovančević, A. (2025). Comparing large Language models and human annotators

- 95 in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm.
96 *Scientific Reports*, 15(1), 11477. <https://doi.org/10.1038/s41598-025-96508-3>
- 97 Fillenbaum, G. G., Belle, G. van, Morris, J. C., Mohs, R. C., Mirra, S. S., Davis, P. C.,
98 Tariot, P. N., Silverman, J. M., Clark, C. M., Welsh-Bohmer, K. A., & Heyman, A. (2008).
99 CERAD (Consortium to Establish a Registry for Alzheimer's Disease) The first 20 years.
100 *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, 4(2), 96–109.
101 <https://doi.org/10.1016/j.jalz.2007.08.005>
- 102 Llibre-Guerra, J. J., Li, J., Harrati, A., Jiménez-Velazquez, I., Acosta, D. M., Llibre-Rodriguez,
103 J. J., Liu, M.-M., & Dow, W. H. (2021). The Caribbean-American Dementia and Aging
104 Study (CADAS): A multinational initiative to address dementia in Caribbean populations.
105 *Alzheimer's & Dementia*, 17(S7), e053789. <https://doi.org/10.1002/alz.053789>
- 106 Rossi, L., Harrison, K., & Shklovski, I. (2024). The Problems of LLM-generated Data in Social
107 Science Research. *Sociologica*, 18(2), 145–168. [https://doi.org/10.6092/issn.1971-8853/](https://doi.org/10.6092/issn.1971-8853/19576)
108 [19576](https://doi.org/10.6092/issn.1971-8853/19576)
- 109 Sachdeva, P. S., & Nuenen, T. van. (2025). *Normative Evaluation of Large Language Models*
110 *with Everyday Moral Dilemmas*. arXiv. <https://doi.org/10.48550/arXiv.2501.18081>
- 111 Schulze Buschoff, L. M., Akata, E., Bethge, M., & Schulz, E. (2025). Visual cognition in
112 multimodal large language models. *Nature Machine Intelligence*, 7(1), 96–106. <https://doi.org/10.1038/s42256-024-00963-y>
113 <https://doi.org/10.1038/s42256-024-00963-y>
- 114 Soria, C. (2025). *An Empirical Investigation into the Utility of Large Language Models in*
115 *Open-Ended Survey Data Categorization*. OSF. https://doi.org/10.31235/osf.io/wv6tk_v2
- 116 Yang, Z., Du, X., Li, J., Zheng, J., Poria, S., & Cambria, E. (2024). *Large Language*
117 *Models for Automated Open-domain Scientific Hypotheses Discovery*. arXiv. <https://doi.org/10.48550/arXiv.2309.02726>
118 <https://doi.org/10.48550/arXiv.2309.02726>