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

29 May 2025

Summary

The rapid advancement of large language and vision models has created new opportunities for automated text and image analysis in social science research (Schulze Buschoff et al. 2025; Yang et al. 2024; Sachdeva and Nuenen 2025). 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, Harrison, and Shklovski 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

Researchers across social sciences, digital humanities, and related fields increasingly need to analyze large volumes of unstructured text and image data. Traditional qualitative coding methods are time-intensive and often impractical for datasets with thousands of responses, while existing automated text analysis tools like topic modeling or sentiment analysis libraries require extensive preprocessing and domain expertise that many researchers lack.

Current solutions present several limitations for academic researchers. General-purpose natural language processing libraries such as spaCy or NLTK require significant programming knowledge and custom model training. 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 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. Unlike existing tools, `CatLLM` improves reproducible, structured outputs while supporting multiple AI providers and maintaining cost efficiency through built-in optimization features.

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

Survey Response	Financial	Family	Housing Features	New Job
Because I wanted a bigger house	0	0	1	0
I needed more money, so I got a new job	1	0	0	1
We started a family and wanted a bigger house	0	1	1	0

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

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:

```
pip install cat-llm

import catllm
```

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.

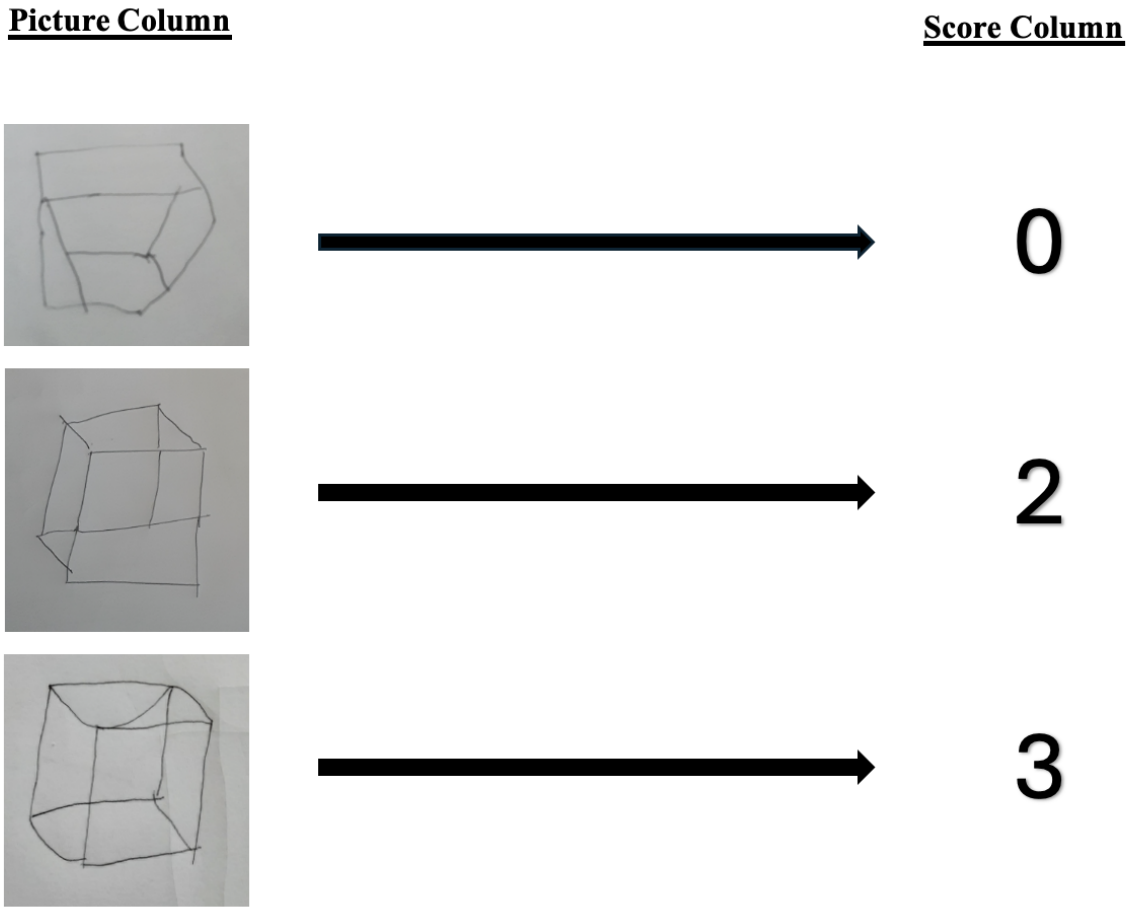

Figure 2: Scoring Drawings of Cubes According to CERAD Rules Using CatLLM

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

Fillenbaum, Gerda G., Gerald van Belle, John C. Morris, Richard C. Mohs, Suzanne S. Mirra, Patricia C. Davis, Pierre N. Tariot, et al. 2008. "CERAD (Consortium to Establish a Registry for Alzheimer's Disease) The First 20 Years." *Alzheimer's & Dementia : The Journal of the Alzheimer's Association* 4 (2): 96–109. https://doi.org/10.1016/j.jalz.2007.08.005.

Llibre-Guerra, Jorge J, Jing Li, Amal Harrati, Ivonne Jiménez-Velazquez, Daisy M Acosta, Juan J Llibre-Rodriguez, Mao-Mei Liu, and William H Dow. 2021. "The Caribbean-American Dementia and Aging Study (CADAS): A Multinational Initiative to Address Dementia in Caribbean Populations." *Alzheimer's & Dementia* 17 (S7): e053789. https://doi.org/10.1002/alz.053789.

Rossi, Luca, Katherine Harrison, and Irina Shklovski. 2024. "The Problems of LLM-Generated Data in Social Science Research." *Sociologica* 18 (2): 145–68. https://doi.org/10.6092/issn.1971-8853/19576.

Sachdeva, Pratik S., and Tom van Nuenen. 2025. "Normative Evaluation of Large Language Models with Everyday Moral Dilemmas." arXiv. https://doi.org/10.48550/arXiv.2501.18081.

Schulze Buschoff, Luca M., Elif Akata, Matthias Bethge, and Eric Schulz. 2025. "Visual Cognition in Multimodal Large Language Models." *Nature Machine Intelligence* 7 (1): 96–106. https://doi.org/10.1038/s42256-024-00963-y.

Soria, Chris. 2025. "An Empirical Investigation into the Utility of Large Language Models in Open-Ended Survey Data Categorization." OSF. https://doi.org/10.31235/osf.io/wv6tk_v2.

Yang, Zonglin, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. "Large Language Models for Automated Open-Domain Scientific Hypotheses Discovery." arXiv. https://doi.org/10.48550/arXiv.2309.02726.