The Journal of Open Source Software

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

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Ujjwal Karn

Reviewers:

- @Jiang22s
- @BinhL3

Submitted: 08 June 2025
Published: unpublished

Summary

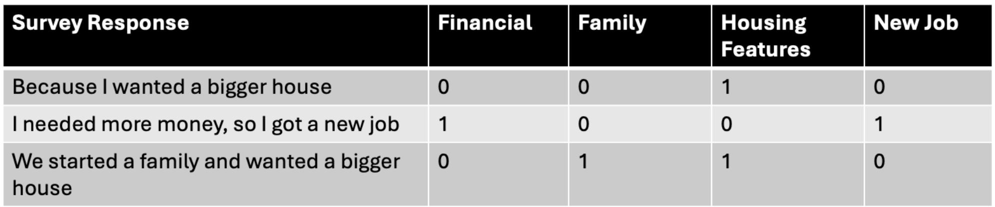
The rapid advancement of large language and vision models has created new opportunities for automated text and image analysis in social science research (Sachdeva & Nuenen, 2025; Schulze Buschoff et al., 2025; Yang et al., 2024). Researchers increasingly use these tools to code open-ended survey responses, categorize qualitative data, and analyze visual content at scale. Yet challenges persist due to inconsistent output formats, diverse API interfaces, and the lack of standardized workflows for integrating both model outputs and external data sources into traditional statistical analysis pipelines (Rossi et al., 2024). CatLLM addresses these issues by providing a modular framework with specialized functions that not only ensure consistent data structures across text and image analysis workflows, but also facilitate the automated retrieval of structured data from the web. The package handles different prompting strategies reactively through configurable parameters that allow users to switch between techniques such as Chain-of-Thought (CoT) (Wei et al., 2023), Chain-of-Verification (CoVe) (Dhuliawala et al., 2023), and step-back prompting (Zheng et al., 2024), enabling researchers to optimize model reasoning based on task complexity without requiring expertise in prompt engineering. This integration allows researchers to seamlessly combine large model outputs with real-world datasets, 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 diverse data sources—from open-ended survey responses and qualitative interviews to unstructured web content—into quantitative datasets suitable for statistical analysis, bridging the gap between traditional research 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). However, LLM outputs can be inconsistent across calls, posing challenges for reproducible qualitative analysis—CatLLM addresses this through frequency-based theme extraction that aggregates results across multiple independent calls rather than relying on single responses. Unlike existing tools, CatLLM provides reproducible, structured outputs while supporting multiple AI providers and maintaining cost efficiency through built-in optimization features.



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

The software has demonstrated practical impact across diverse research domains. It has been successfully applied by institutional researchers at UC Berkeley to track student experience and outcomes, 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:

```
pip install cat-llm

import catllm as cat
```

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

Features

The `CatLLM` package processes diverse data sources—including user-provided text (open-ended survey responses), image data, and unstructured content retrieved from the web—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:

- **Web Data Collection**: Available as part of the CatLLM ecosystem through the companion package `llm-web-research` (`pip install llm-web-research`). This package retrieves and structures unstructured content from web sources, transforming raw online data into standardized datasets suitable for analysis alongside survey and qualitative data. Unlike traditional web scraping approaches, `llm-web-research` prioritizes precision over quantity, using a multi-step verification pipeline to reduce false positives and flag ambiguous queries rather than returning potentially incorrect answers.
- **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.
- **Corpus-Level Theme Discovery**: Improves reproducibility in qualitative theme identification by making multiple independent calls across random corpus segments, then using a secondary model to standardize and consolidate the outputs. This frequency-based extraction method reduces the probabilistic variability inherent in single LLM calls—rather than relying on one potentially inconsistent response, the function surfaces only categories that recur across many independent iterations, producing a reliable, ranked list of themes most representative of the data.

This modular approach provides researchers with consistent data structures across text, image, and web data analysis workflows while maintaining compatibility with standard statistical analysis tools. In reliability testing across eight high-end language models processing 3,208 survey responses each (25,664 total classifications), the package produced valid structured output for 100% of successful API calls—the small number of failures (0–18 per model) were exclusively due to transient server errors rather than JSON parsing issues. Costs ranged from $0.38 (Mistral Medium) to $27.85 (GPT-5), with processing times from 23 minutes to over 7 hours depending on provider rate limits. Further research is needed to evaluate performance with smaller, less capable models.

The `image_multi_class` function has been applied to implement CERAD protocols (Fillenbaum et al., 2008) for scoring geometric shape drawings in the Caribbean-American Dementia and Aging Study (Llibre-Guerra et al., 2021), demonstrating how general-purpose image classification can be adapted to specialized research domains.

The Journal of Open Source Software

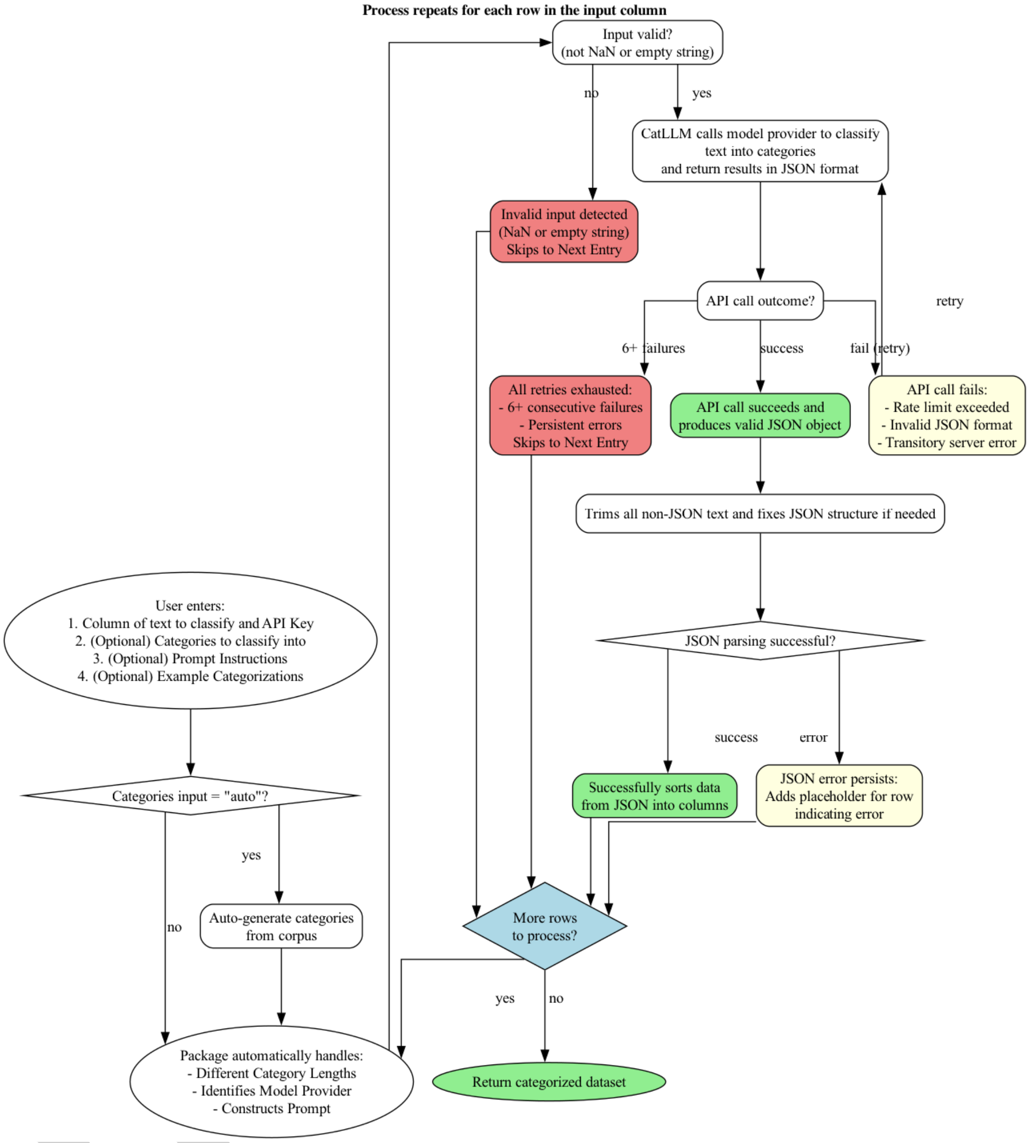

Process repeats for each row in the input column
Input valid?
(not NaN or empty string)
no
yes
CatLLM calls model provider to classify
text into categories
and return results in JSON format
Invalid input detected
(NaN or empty string)
Skips to Next Entry
API call outcome?
retry
6+ failures
success
fail (retry)
All retries exhausted:
- 6+ consecutive failures
- Persistent errors
Skips to Next Entry
API call succeeds and
produces valid JSON object
API call fails:
- Rate limit exceeded
- Invalid JSON format
- Transitory server error
Trims all non-JSON text and fixes JSON structure if needed
User enters:
1. Column of text to classify and API Key
2. (Optional) Categories to classify into
3. (Optional) Prompt Instructions
4. (Optional) Example Categorizations
JSON parsing successful?
success
error
Successfully sorts data
from JSON into columns
JSON error persists:
Adds placeholder for row
indicating error
Categories input = "auto"?
yes
no
Auto-generate categories
from corpus
More rows
to process?
yes
no
Package automatically handles:
- Different Category Lengths
- Identifies Model Provider
- Constructs Prompt
Return categorized dataset

Figure 2: CatLLM Classification Process Flow

Acknowledgements

This work was supported by the UC Berkeley Mentored Research Award. The author thanks Matthew Stenberg, Sara Quigley, Madeline Arnold, and 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 researPartial 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., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., & Jovančević, A. (2025). Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm.

Scientific Reports, *15*(1), 11477. https://doi.org/10.1038/s41598-025-96508-3

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). *Chain-of-Verification Reduces Hallucination in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2309.11495

Fillenbaum, G. G., Belle, G. van, Morris, J. C., Mohs, R. C., Mirra, S. S., Davis, P. C., Tariot, P. N., Silverman, J. M., Clark, C. M., Welsh-Bohmer, K. A., & Heyman, A. (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, J. J., Li, J., Harrati, A., Jiménez-Velazquez, I., Acosta, D. M., Llibre-Rodriguez, J. J., Liu, M.-M., & Dow, W. H. (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, L., Harrison, K., & Shklovski, I. (2024). The Problems of LLM-generated Data in Social Science Research. *Sociologica*, *18*(2), 145–168. https://doi.org/10.6092/issn.1971-8853/19576

Sachdeva, P. S., & Nuenen, T. van. (2025). *Normative Evaluation of Large Language Models with Everyday Moral Dilemmas*. arXiv. https://doi.org/10.48550/arXiv.2501.18081

Schulze Buschoff, L. M., Akata, E., Bethge, M., & Schulz, E. (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, C. (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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2201.11903

Yang, Z., Du, X., Li, J., Zheng, J., Poria, S., & Cambria, E. (2024). *Large Language Models for Automated Open-domain Scientific Hypotheses Discovery*. arXiv. https://doi.org/10.48550/arXiv.2309.02726

Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (2024). *Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2310.06117