

# Assessing Large Language Model Performance in Post-Certification-Examination Comment Categorization

Huaping Sun, Ph.D., MBA

Kristin O'Brien, M.Ed.

Colleen B. Kave, J.D.

David Shin, Ph.D.

Qiao Lin, Ph.D.

Jeffrey M. Lyness, M.D.

American Board of Psychiatry and Neurology

Correspondence: [hsun@abpn.org](mailto:hsun@abpn.org)

## Abstract

This study evaluates a large language model (LLM) for automating the analysis of post-certification-examination comments. A total of 1,406 Neurocritical Care certification examination comments were independently coded by two human raters for sentiment (*positive*, *negative*, and *mixed*) and nine predefined thematic categories. An LLM performed the same tasks under three conditions: baseline prompting, thematic definitions and coding rules, and thematic definitions plus human-coded training examples. Human raters demonstrated high agreement across sentiment and thematic categories. Baseline LLM-human agreement was generally moderate. Contrary to expectations, providing thematic definitions and coding rules decreased LLM-human agreement. Human-coded examples substantially improved LLM performance, especially in concrete, well-defined domains. These findings suggest that LLMs may support preliminary coding of examination feedback, but human review remains essential for complex and subjective comments.

## 1 Introduction

The American Board of Psychiatry and Neurology (ABPN, Deerfield, IL) provides candidates the opportunity to submit free-response comments immediately after completing certification examinations. These comments offer valuable feedback regarding examination content, administration, and overall testing experience. Over time, such feedback has informed content revisions and supported ongoing efforts to improve examination quality.

Despite their value, systematic analysis of post-examination comments presents several operational challenges. Traditionally, ABPN staff manually review and code comments for sentiment and thematic categories (e.g., question quality, content balance, testing platform

functionality). This process is time-consuming, labor-intensive, and subject to variability in human judgment. As the volume of candidate feedback continues to increase, there is a growing need for more scalable, efficient, and consistent analytical approaches to comment analysis.

Recent advances in artificial intelligence, particularly large language models (LLMs), offer a promising opportunity to automate aspects of qualitative feedback analysis ([Bennis & Mouwafaq, 2025](#); [Parfenova et al., 2024](#)). This study adopts a progressive evaluation framework to examine the feasibility and accuracy of using an LLM to classify post-examination comments by sentiment and thematic category. Specifically, the study investigates the extent to which LLM-generated classifications align with human judgments and whether additional guidance or training examples improve LLM coding performance.

## 2 Methods

### 2.1 Data Sample

The study sample included 1,406 free-text comments submitted through an optional post-examination survey following three Neurocritical Care (NCC) certification examination administrations: 587 comments from 2021, 350 from 2022, and 469 from 2024.

### 2.2 Coding Framework

Each comment was coded with respect to two dimensions: sentiment and thematic category. Sentiment analysis classified each comment as *positive*, *negative*, or *mixed*. Thematic categorization involved identifying the presence or absence of one or more of nine predefined categories. These categories were developed

through a preliminary review of NCC post-examination comments and were selected based on their relevance to future examination improvement (Table 1).

Two human raters independently reviewed an initial subset of comments to familiarize themselves with the coding framework and then met to discuss and align their interpretation of the coding criteria (Saldaña, 2025). Following this calibration process, each rater independently

coded all comments from the three examination years. The independent ratings were used to assess inter-rater reliability and to establish a benchmark for evaluating LLM performance.

After completing the initial coding, the same two raters re-examined the 2021 comments and resolved discrepancies through discussion. The resulting consensus ratings served as the training dataset used in the final stage of LLM evaluation.

| Category                          | Description                                                                                                                           |
|-----------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| Prefer Time by Section            | Desire for the exam to be timed by individual sections rather than pooled time across all sections.                                   |
| Not Enough Break Time             | Insufficient break periods; delays due to longer-than-expected check-in and check-out processes.                                      |
| Not Enough Total Time             | Feeling rushed or lacking adequate time to complete the examination.                                                                  |
| Poor Media Quality                | Issues with images or videos, such as poor resolution, darkness, lack of clarity, or confusing visuals.                               |
| Editorial Errors                  | Complaints about misspellings, grammatical errors, or duplicate answer choices (distractors).                                         |
| Poor Question Quality             | Questions perceived as unclear, overly vague or detailed, having multiple plausible answers, or lacking well-designed answer choices. |
| Lack of Content Balance           | Feedback about disproportionate coverage of certain topics, either too much or too little.                                            |
| Questions Not Reflecting Practice | Concerns that questions do not mirror real-world clinical practice.                                                                   |
| Exam Functions Need Improvement   | Suggestions for better usability features, such as easier ways to highlight, mark, or retrieve questions for later review.            |

Table 1. Categories for Neurocritical Care Post-Examination Comments

## 2.3 LLM Evaluation Conditions

To examine the extent to which increasing levels of guidance improve agreement between LLM-generated classifications and human ratings, the LLM was evaluated under three conditions:

- 1) Baseline condition. The LLM performed sentiment analysis and thematic categorization using a minimal prompt, with no additional information regarding category definitions or coding rules.
- 2) Thematic-definition-and-coding-rule condition. In addition to the base prompt, the LLM was provided with descriptions of each thematic category and coding rules linking sentiment classifications to thematic categories. Specifically, because all thematic

categories represented concerns, criticisms, or areas for improvement, category tags were applied only to negative comments or to the negative component of mixed comments.

- 3) Thematic-input-plus-training condition. In addition to thematic definitions and coding rules, the LLM was provided with the consensus ratings of the 2021 comments for training purposes. This condition was designed to evaluate whether exposure to domain-specific, human-rated examples would further improve LLM coding performance.

We hypothesized that coding accuracy would improve following the introduction of thematic guidance and would improve further after the inclusion of human-coded training examples.

## 2.4 LLM Implementation

Microsoft Copilot was used as the platform for implementing the LLM-based coding workflow. For each examination year, an Excel file containing one comment per row and one category per column was submitted to the LLM, which generated a new file that preserved the original data structure while adding sentiment classifications and thematic category indicators (1 = present, 0 = absent). Coding was performed separately under each of the three evaluation conditions.

## 2.5 Statistical Analysis

Agreement was assessed between the two human raters and between the LLM and each human rater under each evaluation condition. Cohen's kappa coefficient ( $\kappa$ ), which accounts for agreement expected by chance, was calculated for sentiment classification and for each thematic category separately by examination year. Kappa values were then averaged across examination years and other relevant analytical groups (e.g., raters or

categories), as appropriate. Kappa values were interpreted according to the criteria proposed by [Landis and Koch \(1977\)](#), with values of 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.00 indicating fair, moderate, substantial, and almost perfect agreement, respectively.

## 3 Results

Inter-rater agreement between the two human raters was consistently high across sentiment analysis and thematic categorization, with average Cohen's  $\kappa$  ranging from 0.68 to 0.97 across categories ([Table 2](#)). Human-human agreement was highest for sentiment analysis (average  $\kappa = 0.95$ ) and for categories with relatively concrete coding criteria, such as *Not Enough Break Time* (average  $\kappa = 0.97$ ), *Prefer Time by Section* (average  $\kappa = 0.94$ ), and *Poor Media Quality* (average  $\kappa = 0.94$ ). Lower, although still substantial, agreement was observed for more subjective categories, including *Questions Not Reflecting Practice* (average  $\kappa = 0.73$ ) and *Exam Functions Need Improvement* (average  $\kappa = 0.68$ ).

| Category                          | Human-Human<br>$\kappa$ | LLM Baseline<br>$\kappa$ | LLM Thematic<br>Input $\kappa$ | LLM Training<br>$\kappa^\ddagger$ |
|-----------------------------------|-------------------------|--------------------------|--------------------------------|-----------------------------------|
| Sentiment                         | .95                     | .49                      | .46                            | .70                               |
| Prefer Time by Section            | .94                     | .66                      | .56                            | .83                               |
| Not Enough Break Time             | .97                     | .70                      | .51                            | .78                               |
| Not Enough Total Time             | .93                     | .62                      | .47                            | .62                               |
| Poor Media Quality                | .94                     | .66                      | .57                            | .80                               |
| Editorial Errors <sup>†</sup>     | .90                     | .57                      | .43                            | .45                               |
| Poor Question Quality             | .84                     | .50                      | .46                            | .60                               |
| Lack of Content Balance           | .84                     | .38                      | .28                            | .49                               |
| Questions Not Reflecting Practice | .73                     | .41                      | .38                            | .53                               |
| Exam Functions Need Improvement   | .68                     | .42                      | .37                            | .25                               |
| <i>Average across categories</i>  | <i>.87</i>              | <i>.54</i>               | <i>.45</i>                     | <i>.61</i>                        |

Table 2. Average Cohen's kappa coefficients between human raters and the LLM across examination years

\*  $\kappa$  = Cohen's kappa coefficient. 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect agreement.

<sup>†</sup> Values for Editorial Errors were averaged across 2021 and 2022 because no editorial errors were identified by human raters in 2024.

<sup>‡</sup> Only 2022 and 2024 data were used for LLM training calculations because the 2021 coded data were used for training.

Under the baseline condition, agreement between the LLM and human raters was generally moderate ([Table 2](#)). For sentiment analysis, the LLM achieved an average  $\kappa$  of 0.49.

Across thematic categories, agreement ranged from average  $\kappa = 0.38$  (*Lack of Content Balance*) to average  $\kappa = 0.70$  (*Not Enough Break Time*).

Providing the LLM with thematic definitions and coding rules did not improve performance (Table 2). Instead, agreement decreased across all outcomes, with the largest decreases observed for *Not Enough Break Time* (average  $\kappa = 0.70$  to 0.51), *Not Enough Total Time* (average  $\kappa = 0.62$

to 0.47), and *Editorial Errors* (average  $\kappa = 0.57$  to 0.43). The LLM classified a higher percentage of comments as *positive* than did the human raters under the thematic input condition (Table 3). Because comments classified as *positive* were not assigned thematic categories, this tendency may have contributed to lower thematic-category tagging rates relative to the baseline condition (Table 4).

| Rater              | Positive (%) | Negative (%) | Mixed (%) |
|--------------------|--------------|--------------|-----------|
| Human              | 21.0         | 48.0         | 30.8      |
| LLM Baseline       | 45.3         | 30.7         | 24.3      |
| LLM Thematic Input | 48.0         | 29.0         | 23.0      |
| LLM Training*      | 16.5         | 50.5         | 33.0      |

Table 3. Average percentage of comments classified as *positive*, *negative*, and *mixed* by human raters and the LLM across examination years

\* Only 2022 and 2024 data were used because the 2021 consensus data were used for training.

In contrast, the addition of human-coded training examples substantially improved LLM performance for most outcomes (Table 2). Agreement for sentiment analysis increased from average  $\kappa = 0.49$  under baseline condition to average  $\kappa = 0.70$  after training. The highest post-training agreement was observed for *Prefer Time by Section* (average  $\kappa = 0.83$ ), *Poor Media*

*Quality* (average  $\kappa = 0.80$ ), and *Not Enough Break Time* (average  $\kappa = 0.78$ ). Improvements were also observed for *Questions Not Reflecting Practice* (average  $\kappa = 0.41$  to 0.53), *Lack of Content Balance* (average  $\kappa = 0.38$  to 0.49), and *Poor Question Quality* (average  $\kappa = 0.50$  to 0.60).

| Thematic Category                 | Human (%) | LLM Baseline (%) | LLM Thematic Input (%) | LLM Training (%)* |
|-----------------------------------|-----------|------------------|------------------------|-------------------|
| Prefer Time by Section            | 8.7       | 4.3              | 3.7                    | 10.0              |
| Not Enough Break Time             | 4.5       | 2.7              | 1.7                    | 5.0               |
| Not Enough Total Time             | 15.5      | 11.0             | 6.3                    | 20.5              |
| Poor Media Quality                | 23.0      | 15.0             | 14.3                   | 23.5              |
| Editorial Errors                  | 1.5       | 3.3              | 2.3                    | 3.0               |
| Poor Question Quality             | 28.5      | 18.7             | 17.0                   | 42.5              |
| Lack of Content Balance           | 13.5      | 7.3              | 6.3                    | 24.5              |
| Questions Not Reflecting Practice | 18.8      | 7.7              | 6.3                    | 31.0              |
| Exam Functions Need Improvement   | 5.0       | 5.0              | 2.7                    | 17.0              |

Table 4. Average percentage of thematic categories tagged by human raters and LLMs across examination years

\* Only 2022 and 2024 data were used because the 2021 consensus data were used for training.

## 4 Discussion

This study examined the extent to which an LLM could replicate human coding of post-examination comments and whether additional guidance or training examples improved LLM coding performance. Three key findings emerged. First, human raters demonstrated consistently high agreement across sentiment analysis and thematic categorization, supporting the reliability of the coding framework. Second, the LLM achieved moderate agreement under the baseline prompting condition, with performance varying substantially across categories. Third, providing category definitions and coding rules that introduced task dependency did not improve performance, whereas the addition of human-coded examples resulted in substantial gains for sentiment analysis and most thematic categories.

Both human raters and the LLM exhibited higher agreement for concrete and observable categories such as *Not Enough Break Time*, *Prefer Time by Section*, and *Poor Media Quality*, and lower agreement for categories requiring subjective interpretation, including *Questions not Reflecting Practice* and *Exam Functions Need Improvement*. Consistent with the findings of [Liu et al. \(2025\)](#) using GPT-4, these parallel difficulty patterns observed for human raters and the LLM suggest that lower agreement in certain categories may reflect the complexity of the coding task itself rather than model performance alone.

A particularly noteworthy finding was that providing thematic definitions and coding rules actually reduced agreement in most thematic categories. One potential explanation is that the thematic-input condition introduced task dependency between sentiment and category assignments (i.e., not assigning thematic-category tags to comments classified as *positive*), which unintentionally amplified LLM's potential optimism bias. In contrast, the addition of human-coded examples substantially improved agreement across most outcomes, especially in more clearly defined domains.

Several limitations should be noted. The study used comments from one subspecialty examination, a predefined set of thematic categories, and only one set of thematic input and coding rules, and a single training dataset. Other contexts or iterative prompt-development approaches may yield different results. In

particular, thematic category definitions and coding rules should have been separate prompts.

In conclusion, this study suggests that LLMs may be useful for preliminary categorization and prioritization of examination comments for further review. However, human oversight remains important for comments involving ambiguity, complexity, contextual reasoning, and subjective judgment. Future studies may investigate the reasoning that LLM provides after coding execution and explore creative approaches to improve LLM coding performance.

## References

- Issam Bennis and Safwane Mouwafaq. 2025. Advancing AI-driven thematic analysis in qualitative research: A comparative study of nine generative models on cutaneous leishmaniasis data. *BMC Medical Informatics and Decision Making*, 25, 124. <https://link.springer.com/content/pdf/10.1186/s12911-025-02961-5.pdf>
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. <https://www.jstor.org/stable/2529310>
- Xiner Liu, Andres Felipe Zambrano, Ryan S. Baker, Amanda Barany, Jaclyn Ocumpaugh, Jiayi Zhang, Maciej Pankiewicz, Nidhi Nasir, and Zhanlan Wei. 2025. Qualitative coding with GPT-4: Where it works better. *Journal of Learning Analytics*, 12(1), 169–185. <https://files.eric.ed.gov/fulltext/EJ1465623.pdf>
- Angelina Parfenova, Alexander Denzler, and Jörgen Pfeffer. 2024. Automating qualitative data analysis with large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Student Research Workshop)*, 83–91. <https://aclanthology.org/2024.acl-srw.17/>
- Johnny Saldaña. 2025. *The Coding Manual for Qualitative Researchers* (5th ed.). Sage Publications.