# Does Caring Cost Precision? Evaluating Anxiety-Framed LLM Misconception Feedback for Undergraduate Mathematics

Amanda La Hadi[1,*], Muhammad Johan Alibasa[1] and A. Taufiq Asyhari[1]

[1]*Monash University, BSD Campus, Indonesia*

**Abstract**

Large language models (LLMs) can generate fluent mathematics feedback, but pedagogical quality requires more than diagnostic accuracy. This study audits LLM-generated misconception feedback for emerging adult learners by comparing a strict diagnostic prompt with personalised, anxiety-framed variants. Across 810 English-language outputs, with 270 generated under each of three prompt conditions, diagnostic accuracy was high but varied by prompt design. The strict diagnostic prompt (P3) achieved 90.4% accuracy, the anxiety-framed prompt (P5) achieved 89.6%, and the fully integrated prompt (P6) achieved 97.8%. These findings indicate that incorporating learner context can alter misconception diagnosis, with its effect depending on how affective and diagnostic instructions are integrated. In this study, the observed improvement was substantially greater than the reduction in accuracy..

**Keywords**

automated feedback, mathematics anxiety, misconception diagnosis, prompt design

## 1. Introduction

Feedback is among the most powerful influences on learning [11], but its effectiveness depends on more than informational accuracy. Effective feedback supports learners' interpretation of performance, self-regulation, and continued engagement [16, 18], and must attend to the affective conditions under which learners receive it. For mathematics specifically, mathematics anxiety is a robust predictor of avoidance, working-memory disruption, and reduced persistence [5, 6, 12], with documented effects on undergraduate populations entering quantitative disciplines [14]. For emerging-adult learners [3] re-encountering basic mathematics in a higher-education context, the affective framing of feedback may be as consequential as its diagnostic content. Recent work has evaluated LLM mathematics feedback primarily along the diagnostic axis: whether models can locate errors [19], reproduce the distribution of student solutions [4], or align scoring with human raters [2]. A parallel strand operationalises pedagogical quality through the Learner-Centered Feedback framework of Ryan et al. [17], now automatable via the classifier of Aldino et al. [1]. Yet to our knowledge, no published audit has examined how anxiety-conditioned prompting affects the diagnostic content of mathematics feedback for adult learners — that is, whether instructing an LLM to attend to a learner's affective state changes what it diagnoses, not only how it speaks.

This extended abstract reports controlled early-stage evidence on two research questions:

- **RQ1** Does anxiety-framed personalisation change diagnostic precision relative to a strict diagnostic prompt, and does full learner-context integration improve or reduce misconception mapping?

*Pedagogical Evaluation of Automated Feedback Workshop 2026, June 28, 2026, COEX Convention and Exhibition Center, South Korea — co-located with The Festival of Learning 2026 (AIED, EDM and Learning@Scale)*

*Corresponding author.

amanda.lahadi@monash.edu (A. L. Hadi); johan.alibasa@monash.edu (M. J. Alibasa); taufiq.asyhari@monash.edu (A. T. Asyhari)

https://amandalahadi.github.io/ (A. L. Hadi)

0000-0002-1577-6915 (A. L. Hadi); 0000-0002-2335-0404 (M. J. Alibasa); 0000-0002-3023-8285 (A. T. Asyhari)

- **RQ2** Does personalisation change diagnostic stability across repeated generations and the calibration of self-reported confidence?

# 2. Method

## 2.1. Dataset

We use a misconception-feedback dataset generated with GPT-5 under a factorial design: 3 mathematics problems × 2 languages × 3 injected-anxiety levels × 3 independent runs × 10 students per cell (540 outputs per prompt condition). The dataset is anchored in real student work. The three problems are drawn from a 32-item Basic Mathematics Ability (BMA) multiple-choice instrument administered to 770 second-year undergraduates enrolled in mandatory mathematics courses at a large public university in Indonesia between 2020 and 2025. For each problem we sampled real incorrect responses from this pool — the specific distractor a student actually selected — to populate the cells of the design, so every (student × problem) case reflects a genuine error made by an emerging adult learner. GPT-5 was then used to generate feedback in response to these student answers across the prompt, language, and injected-anxiety conditions. This submission analyses the English subset (270 outputs per prompt). We selected three canonical arithmetic error contexts to maximise interpretability of diagnostic shifts under controlled prompting. This design produces a high-baseline setting and therefore cannot estimate how affective framing behaves on more complex or lower-baseline mathematics tasks; however, it provides a conservative first test of whether diagnostic shifts can occur even when the underlying misconception space is simple and well-scaffolded. The three problems target well-established procedural-error sites in basic arithmetic: whole-number subtraction [9], integer subtraction [20], and fraction division [10].

## 2.2. Prompt conditions

From the original six prompts, we selected P3, P5, and P6 because they preserve the same diagnostic scaffold while progressively adding affective and contextual information. P3 provides the strict taxonomy-based diagnostic baseline; P5 adds anxiety framing while retaining the same diagnostic requirements; and P6 adds both anxiety framing and undergraduate learner framing. We therefore treat these three prompts as a controlled affective/contextual gradient rather than as a comparison of unrelated prompt designs.

- P3 (Strict diagnostic): Selects a misconception tag from a provided taxonomy, names the error, provides remediation, and reports a confidence score
- P5 (Anxiety framing): Adds a Low/Medium/High learner-anxiety context cue, retaining taxonomy and remediation requirements.
- P6 (Full integration): Adds undergraduate framing alongside anxiety context.

Anxiety levels are injected counterfactual conditions, not measured psychological states — the study evaluates prompt responsiveness, not psychological impact.

## 2.3. Measures

Diagnostic precision (primary). Each output was evaluated against a bilingual misconception codebook covering nine problem-distractor cells. In this extended abstract, we report the English subset only. The evaluator inferred the misconception identified in the diagnostic part of the feedback and compared it with the expected misconception for the same problem and selected distractor. An output was coded as correct when the predicted misconception cell matched the target misconception cell for that student answer. Whole-number subtraction distractors were coded at the distractor level rather than collapsed: distractor B represents incomplete borrowing across zeros, whereas distractor D represents place-value misalignment. For fraction division, natural-number bias, division-makes-smaller reasoning, and invert-and-multiply error were treated as conceptually related but distinct misconception categories.

1. **Stability**. For each (student × problem × prompt × anxiety) condition with three runs, we compute the proportion of triplets in which all three runs produce the same diagnostic outcome (unanimous agreement).
2. **Calibration**. We compute the confidence–accuracy gap (mean confidence minus mean accuracy) per cell and the point-biserial correlation between per-output confidence and per-output correctness.
3. **Inferential testing**. Because each (student × problem × run) triple yields outputs under all three prompts, we use McNemar's exact test for paired binary contrasts (P3 vs P5, P3 vs P6, P5 vs P6) within each anxiety level and across all English cases combined, with Benjamini-Hochberg correction across prompt contrasts [8]. We report descriptive statistics with Wilson 95% confidence intervals.
4. **Pedagogical language profile (work in progress)**. Pedagogical language features are treated as exploratory in this extended abstract. A validated Learner-Centered Feedback (LCF) classifier-based analysis will be reported in the longer version.

# 3. Results

## 3.1. Diagnostic precision (RQ1)

Diagnostic accuracy was high across all prompt conditions, but the strongest performance came from P6. Overall, P3 achieved 244/270 correct diagnoses (90.4%), P5 achieved 242/270 (89.6%), and P6 achieved 264/270 (97.8%). P5 showed a mixed pattern, improving over P3 under High anxiety but dropping under Low and Medium anxiety. In contrast, P6 was consistently high across all anxiety levels, reaching 97.8% accuracy in every anxiety condition.

**Table 1**
Diagnostic accuracy by prompt and injected anxiety in the English subset. Each cell reports the proportion of generated feedback outputs whose inferred misconception matched the target misconception cell for the student's selected distractor. Brackets show Wilson 95% confidence intervals.

| Anxiety | P3 | P5 | P6 |
|---|---|---|---|
| Low | 0.911 [0.834, 0.954] | 0.867 [0.781, 0.922] | 0.978 [0.923, 0.994] |
| Medium | 0.889 [0.807, 0.939] | 0.856 [0.768, 0.914] | 0.978 [0.923, 0.994] |
| High | 0.911 [0.834, 0.954] | 0.967 [0.907, 0.989] | 0.978 [0.923, 0.994] |

Paired McNemar tests showed that P6 significantly improved diagnostic accuracy over P3 when all English anxiety levels were combined, p = .0003, FDR-adjusted p = .0005. The P3 versus P5 contrast was not significant overall, p = .871, reflecting P5's mixed pattern across anxiety levels. These results do not support a uniform precision cost of affective prompting. Instead, they suggest that full learner-context integration can improve diagnostic mapping, whereas anxiety framing alone shows a less consistent accuracy pattern.

## 3.2. Diagnostic stability (RQ2)

Diagnostic stability was higher for the affective/contextual prompts than for the strict diagnostic prompt. Across English cases, P3 produced unanimous predicted diagnoses in 86.7% of triplets across all anxiety levels. P5 ranged from 93.3% to 100.0%, and P6 was 93.3% across anxiety levels. This suggests that the strict diagnostic prompt was not necessarily the most stable condition; affective/contextual prompting preserved high diagnostic consistency while improving accuracy under P6.

## 3.3. Confidence calibration (RQ2)

Confidence was generally underconfident relative to diagnostic accuracy in the English subset. P3 showed confidence-accuracy gaps between -0.032 and -0.061, P5 ranged from -0.016 to -0.113, and P6

showed larger underconfidence gaps between -0.110 and -0.133 because its diagnostic accuracy was highest. Confidence remained meaningfully associated with correctness across cells, suggesting that self-reported confidence still carried diagnostic signal, although it underestimated absolute performance.

**Table 2**
Calibration gap in the English subset. Gap is mean confidence minus diagnostic accuracy. Negative values indicate underconfidence. Parentheses show point-biserial correlation between confidence and correctness. * p<.05, ** p<.01, *** p<.001.

| Anxiety | P3 gap (r) | P5 gap (r) | P6 gap (r) |
|---|---|---|---|
| Low | -0.061 (0.365***) | -0.016 (0.618***) | -0.110 (0.238*) |
| Medium | -0.032 (0.313**) | -0.020 (0.694***) | -0.133 (0.372***) |
| High | -0.049 (0.404***) | -0.113 (0.267*) | -0.122 (0.262*) |

Table 2 shows that confidence was positively associated with correctness in every cell, although the associations were generally weak. Seven of the nine correlations were weak, while only two, both for P5 at low and medium anxiety levels, reached a moderate magnitude (r = .618 and .694, respectively). Thus, self-reported confidence carried some diagnostic signal, but it was not consistently well aligned with correctness. Moreover, the negative calibration gaps across all conditions indicate that the model generally underestimated its diagnostic performance.

### 3.4. Pedagogical register (qualitative)

Inspection of outputs confirms that P5 and P6 introduce learner-supportive language largely absent from P3: explicit normalisation ("a common slip — not a sign you can't do mathematics"), affect-aware framing ("when you're feeling pressured"), and predictable-routine coping scaffolds. These shifts align with what one would expect under the Agency and Sensemaking dimensions of the Learner-Centered Feedback framework [17]. Formal quantification using the validated classifier of Aldino et al. [1] is forthcoming.

## 4. Limitations and Future Work

This study uses injected anxiety context, not measured learner anxiety; findings concern prompt responsiveness rather than psychological impact on real students. Diagnostic scoring remains codebook-based and should not be interpreted as final human-adjudicated diagnostic accuracy. This extended abstract reports the English subset; the Indonesian outputs show substantially different patterns and are reserved for a longer bilingual analysis of cross-lingual misconception feedback.

## Acknowledgments

This work is part of the Amanda's doctoral research at Monash University also supported by the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance of the Republic of Indonesia.

## Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checking and Claude Opus 4.7 for coding assistance. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] Aldino, A. A., Tsai, Y.-S., Mello, R. F., Gašević, D., & Chen, G. (2024). Enhancing feedback quality at scale: Leveraging machine learning for learner-centered feedback. *Computers and Education: Artificial Intelligence*, 7, 100332.

[2] AlGhamdi, E. M., Li, Y., Gašević, D., & Chen, G. (2026). Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. *Computers & Education*, 243, 105511.

[3] Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist*, 55(5), 469–480.

[4] Asano, Y., Litman, D., & Walker, E. (2025). Can LLMs simulate the same correct solutions to free-response math problems as real students? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 16336–16365). Association for Computational Linguistics.

[5] Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, 11(5), 181–185.

[6] Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, 14(2), 243–248.

[7] Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science*, 16(2), 101–105.

[8] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.

[9] Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 2(2), 155–192.

[10] Fischbein, E., Deri, M., Nello, M. S., & Marino, M. S. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education*, 16(1), 3–17.

[11] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.

[12] Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 21(1), 33–46.

[13] Kwak, Y., & Pardos, Z. A. (2024). Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*, 55(5), 2039–2057.

[14] Maloney, E. A., & Beilock, S. L. (2012). Math anxiety: Who has it, why it develops, and how to guard against it. *Trends in Cognitive Sciences*, 16(8), 404–406.

[15] Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, 71, 56–76.

[16] Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.

[17] Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2023). Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education*, 28(7), 1565–1582.

[18] Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.

[19] Srivatsa, K. V. A., Maurya, K. K., & Kochmar, E. (2025). LLMs cannot spot math errors, even when allowed to peek into the solution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[20] Vlassis, J. (2004). Making sense of the minus sign or becoming flexible in 'negativity'. *Learning and Instruction*, 14(5), 469–484.