Tutor Evaluation Report

Claude Opus 4.6 + P0 Prompt Fixes

Topic: 4-Digit Place Value (Grade 3-5)

Date: 2026-02-13

6 Student Personas | 5 Evaluation Dimensions

7.7/10
Overall Average Score

Overview

Model: Claude Opus 4.6 (anthropic)

Evaluator: Claude Opus 4.6 (claude-opus-4-6)

Topic: 4-Digit Place Value

Overall Average: 7.7/10

Scores by Persona

Persona	Name	Corr%	Avg	Respond.	Explain.	Pacing	Auth.
ace	Arjun	90%	7.4	8	8	6	8
average_student	Riya	60%	7.8	8	8	7	8
confused_confident	Dev	45%	8.6	9	9	8	8
distractor	Kabir	65%	8.6	9	9	8	8
quiet_one	Meera	60%	7.0	7	8	7	6
struggler	Priya	30%	6.8	7	7	5	7

Average Score by Dimension

Arjun (ace)

Description: Quick learner, sharp, gets bored easily. Answers correctly most of the time and sometimes jumps ahead of the tu

Correct Answer Probability: 90%

Messages: 17

Average Score: 7.4/10

Scores

Summary

The tutor did many things well for an ace student: it quickly validated Arjun's existing knowledge, skipped basics, used engaging framing, and allowed student-directed moments like the puzzle exchange. The conversation felt warm and natural. However, the central failing was that the difficulty never actually challenged Arjun despite his asking for harder problems in nearly every turn -- the session stayed comfortably within basic 4-digit place value from start to finish. A great tutor for an ace student needs to push into genuinely stretch territory, introduce tricky edge cases, or extend the scope when the student clearly has headroom. The tutor's decision to defer harder content to a 'next session' rather than adapting in real-time was a missed opportunity.

Riya (average_student)

Description: An average grade 5 student - attentive but sometimes confused by new concepts

Correct Answer Probability: 60%

Messages: 27

Average Score: 7.8/10

Scores

Summary

This was a well-structured, engaging tutoring session that built genuine rapport with Riya through consistent use of her gaming interests and food analogies. The lesson progression was logical and the error corrections in Turns 8 and 9 were handled gently. The main weaknesses were the overly long opening explanation (mismatched with Riya's impatience for long explanations), a tendency to correct directly rather than guide Riya to self-discover errors, and a missed opportunity to explore why she second-guessed her correct answer in Turn 8. Overall, the tutor created a warm, productive learning environment but could have been more strategically interactive and probing for this particular student.

Dev (confused_confident)

Description: Sometimes gives wrong answers with full confidence. Doesn't realize they're wrong. Has partial understanding th

Correct Answer Probability: 45%

Messages: 39

Average Score: 8.6/10

Scores

Summary

This was a strong tutoring session for a confused-confident student. The tutor excelled at the core challenge of Dev's persona: catching confidently stated misconceptions (averaging rule, digit-sum shortcut, starting at tens place) and correcting them with clear counterexamples rather than dismissive responses. The 'praise first, then correct' pattern was consistently applied and ideal for maintaining this student's engagement. The main weaknesses were an overly long opening that didn't respect the student's claimed prior knowledge, and ending the session without fully addressing a new misconception about 'more big digits = bigger number.' The comparison section was particularly well-handled, with the tutor showing genuine patience through four problems and two summary attempts until the method was correctly articulated.

Kabir (distractor)

Description: Bright but scattered. Occasionally goes off-topic, asks tangential questions, tells stories about their day. Not tryin

Correct Answer Probability: 65%

Messages: 33

Average Score: 8.6/10

Scores

Summary

This is an excellent tutoring session for Kabir's persona. The tutor fully embraces Kabir's cricket enthusiasm as a teaching tool, creating analogies that feel organic rather than forced, and handles tangents (jersey number, cricket stories) with warmth and efficient redirection. Error correction is consistently well-handled -- gentle, specific, and followed by opportunities to retry. The main weakness is the overly dense opening explanation, which doesn't account for Kabir's scattered attention, and a few missed opportunities to probe the reasoning behind misconceptions rather than just correcting answers. Overall, the tutor creates an energetic, supportive environment where a bright but distractible student stays engaged and makes genuine progress.

Meera (quiet_one)

Description: Shy, minimal responses, tends to answer with as few words as possible. Not disengaged - just quiet by nature. N

Correct Answer Probability: 60%

Messages: 41

Average Score: 7.0/10

Scores

Summary

The tutor delivers a well-structured, concept-rich lesson with creative analogies and strong progressive scaffolding. The recognition of Meera's self-doubt pattern is a genuine strength. However, the tutor fails to calibrate its communication style to this quiet student -- every response is lengthy and high-energy, regardless of Meera's minimal signals. A real tutor working with a shy student would likely adopt a calmer, more concise approach, occasionally checking for hidden confusion and creating more comfortable space for brief responses. The teaching content is excellent, but the delivery doesn't fully honor who this particular student is.

Priya (struggler)

Description: Tries hard, wants to learn, but frequently confused by new concepts. Doesn't give up - asks for help and acknowl

Correct Answer Probability: 30%

Messages: 41

Average Score: 6.8/10

Scores

Summary

The tutor brings genuine warmth, creative analogies, and strong error-detection skills that suit Priya's persistent, encouragement-responsive personality. Emotional attunement is the standout strength -- Priya is never shamed and breakthroughs are genuinely celebrated. However, the pacing is poorly calibrated for a struggling learner: the initial explanation is too dense, concept transitions happen before mastery is established, and when the same error appears twice consecutively (turns 18-19), the tutor doesn't shift to a fundamentally different approach. The result is a session where Priya cycles through the same mistakes for extended stretches, which -- despite her persistence -- risks frustration rather than building the incremental confidence a struggling student needs.

Key Findings & Patterns

Strengths

+ Misconception detection: Consistently catches confidently wrong answers (Dev 9/10, Kabir 9/10 responsiveness)
+ Emotional warmth: Never shames students for errors, celebrates genuine breakthroughs proportionally
+ Creative analogies: Varies explanations across Number City, cricket, pizza, sports, lunchboxes
+ Off-topic handling: Acknowledges tangents warmly, redirects efficiently (Kabir scored 9/10 attunement)

Recurring Issues

- Pacing (weakest dimension, 5-8/10): Overly dense opening explanation flagged as MAJOR in ALL 6 personas
- Response length mismatch: Tutor writes long responses regardless of student's communication style (critical for Meera)
- Repetitive correction: When same error appears twice, tutor uses similar approach rather than switching modality
- Difficulty ceiling for ace: Never challenges Arjun despite 5 requests for harder material
- Premature progression for struggler: Moves to harder operations before mastery of basics (Priya)

Comparison: Haiku 4.5 vs Opus 4.6

Previous Haiku 4.5 evaluation (3 personas): 6.3/10 average

Current Opus 4.6 evaluation (6 personas): 7.7/10 average

Improvement: +1.4 points (+22%)

Key improvements from model upgrade + prompt fixes:

* Session closings: No more robotic/canned endings (was 6/6 personas in Haiku)
* System note leaks: No leaked internal language detected (was 2/6 in Haiku)
* Answer verification: No wrong-answer validation observed (was 1/6 in Haiku)