A Example of Problem Authoring

For completeness, we show a sample problem specification:

```
((private-name "addition")
(synthesized-name a)
(public-test-suite
 ((check-equal? 5 2 3)
  (check-equal? -3 -1 -2)
  (check-equal? 5 7 -2)
  (check-equal? 0 -2 2)))
(private-test-suite
 ((check-equal? 5 1 4)
  (check-equal? 5 5 0)
  (check-equal? -7 -1 -6)
  (check-equal? -14 -7 -7)))
(bad-impl
"(define (a x y)
 (- x y))"))
```

B Details and Analysis of Quantitative Results

In this appendix, we provide additional information and detailed analyses of the responses to the research questions. The summary answers are given in the main paper in §7.4.

All statistical analyses were done with `IBM SPSS Statistics 29` using $\alpha = 0.05$ for significance testing.

B.1 Institution and Students

In addition to the previous study, we also sought not only to capture performance data but also to survey demographic attributes, emotions, and attitudes. IRB approval for this study thus was applied for and approved by the local institutional review board under number 2023-F10-32 (November 5, 2023). To ensure GDPR compliance, we deployed Porpoise on a locally hosted Kubernetes cluster, used a locally hosted LimeSurvey instance for all surveys, and restricted access to the institution's virtual private network. Participants were informed that their text would be sent to an externally hosted LLM and reminded to not include personally identifying information.

In the course, students usually work in groups of three on the weekly homework assignments. This study was conducted in two rounds. Both times, one of the assignments gave students the option of participating in the Porpoise study. Students who declined to opt-in were given a programming assignment deemed to be of similar effort. For either round, less than 12% (Round 1: 39/335; Round 2: 30/331) declined to opt in. The problems and solutions to the non-opt-in assignments were made public after each round.

Through the learning management system, each student was assigned a unique, alphanumeric identifier which served as an authentication token for both Porpoise and the local LimeSurvey installation; these identifiers guaranteed that student data was kept in sync between Porpoise and LimeSurvey. Upon logging into either system, the students had to confirm that they were aware of the protocol and consented to participate and have their data analyzed. The log data extracted from these systems was then merged with the data from the learning management system to award credit for completion; an exercise was considered "completed" if either all tests had been passed (for non-opt-in) or more than one meaningful interaction had taken place (for opt-ins). Cursory inspection revealed that students indeed tried to interact with Porpoise in a meaningful way: we found no instances of students submitting the same text over and over, or submitting irrelevant text.

B.2 RQ-Can't-Code

Though for diverse reasons, all four problems that students would not have been able to solve with their programming knowledge (rem, piglet, table-sum, average-3) are the problems on which they did poorest with Porpoise. Therefore, this process did not showcase how an LLM could produce solutions that they could specify but not code. This pattern is intriguing and warrants further analysis. As a starting point, we hypothesize that the problems in Round 2 required more algorithmic abstraction than could have been expected from students in their first semester.

B.3 RQ-Perception

Students were asked to rate their perceived usefulness of *PS*s before and after using Porpoise on a scale from 1 ("not useful at all") to 10 ("very useful"). We hypothesized that students would find the tool would significantly increase perceived usefulness, because it would show that one could obtain a program directly from it without having to write code. However, student ratings were very high even before the first round, with a median value of 8/10. Comparing pre- and post-intervention ratings with a Related-Samples Wilcoxon Signed Rank test, we saw no significant change in Round 1 but statistically significant increase in Round 2. Of the 277 participants in Round 2, using Porpoise increased their rating for 56 participants while it decreased the rating for 29; 192 participants reported unchanged ratings. Using Porpoise elicited a statistically significant median increase, $z = 2.174$, $p = 0.030$. Over the whole interventions, however, the changes were not statistically significant, with each survey showing a median rating of 8/10.

B.4 RQ-Prompt

Unpaired-t-tests showed that, after Bonferroni correction for multiple testing, there was no significant difference favoring either showing on not showing a buggy program. Not showing a buggy program resulted in significantly fewer attempts for divs (Round 1), with a difference of 1.6 attempts ($p_{\text{adj}} = 0.036, t = 2.781, df = 694, CI = [0.470, 2.730]$); the effect size, however, was small ($d = 0.22$). Notably, there was

	Round 1 ($n = 274$)	Round 2 ($n = 281$)
Mental Demand	4.48 ± 1.359	4.88 ± 1.349
Physical Demand	1.96 ± 1.456	1.87 ± 1.239
Temporal Demand	3.49 ± 1.704	3.65 ± 1.563
Performance	3.75 ± 1.497	3.78 ± 1.447
Effort	4.42 ± 1.330	4.44 ± 1.284

Figure 7. Intervention 2: NASA-TLX results. Data on a scale from 1 (low) to 7 (high) shown as mean ± standard deviation.

Round 1	n	Mean	z_{skewness}	z_{kurtosis}
PROUD	273	2.81 ± 0.099	5.558†	0.565
FRUSTRATED	274	4.03 ± 0.113	−0.095	−3.853†
STUPID	259	3.54 ± 0.099	0.596	−1.886

Round 2	n	Mean	z_{skewness}	z_{kurtosis}
PROUD	280	3.18 ± 0.107	4.130†	−1.686
FRUSTRATED	279	3.89 ± 0.106	0.657	−3.234†
STUPID	275	3.59 ± 0.097	2.136†	−1.761

Figure 8. Intervention 2: Means for the "emotional response" surveys; value ± standard deviation. †: z-scores outside the $[-1.96, 1.96]$ 95% CI, i.e., indicating violation of normality.

no appreciable difference on the problems where students did *especially poorly*.

In student comments (see §7.5), some said they did not even bother looking at the buggy program. If many students followed this practice, it would explain why there were no notable differences. (An interface that asked students to click a button to see the buggy program would have helped know how many bothered, but the very presence of the button may create a friction that reduces seeing them.) Also, while there was no "global" benefit (or notable harm), some individuals may have found value to it: a handful of student comments did indicate that it helped them get started.

B.5 RQ-Complexity

We administered the NASA Task Load Index (NASA-TLX) [17] to examine how Porpoise was perceived. For ecological efficiency, we moved from the TLX's 21-point scales to seven points. Since we assessed emotions using a separate survey (appendix B.6), we also removed the "frustration" scale from the instrument.

Figure 7 presents the results of administering the TLX immediately after Round 1 and Round 2. Paired-samples t-test analyses revealed that only the "Mental Demand" subscale showed significant differences: The $n = 237$ participants responding to both surveys reported a statistically significant increase of 0.371 ± 1.664 units (95% CI, 0.158 to 0.584) with respect to mental demand from Round 1 (4.51 ± 1.317) to Round 2 (4.88 ± 1.343), $t(236) = 3.437$, $p < 0.001$, $d = 0.22$). As no other measure changed in a statistically significant way, we interpret this change as being due to the increased intellectual demand from Problem Set 1 to Problem Set 2.

Inspecting data for the other scales, we observe the following for the data from Round 1. The "Physical Demand" (*How physically demanding was the task?*) is very low, as expected. The "Temporal Demand" (*How hurried or rushed was the pace of the task?*) is centered, which suggests students did not feel too rushed. For "Performance" (*How successful were you in accomplishing what you were asked to do?*), the answers were not found to be normally distributed (Kolmogorov–Smirnov test, $p < 0.001$). We see a positive skewness (0.035 ± 0.147, $z = 0.238$) and a negative kurtosis (-0.611 ± 0.293, $z = -2.085$, thus violating normality), which indicate that the data was slightly shifted to the "unsuccessful" side of the scale with significantly more heavy tails than a normal distribution. For "Effort" (*How hard did you have to work to accomplish your level of performance?*), the answers were not found to be normally distributed (Kolmogorov–Smirnov test, $p < 0.001$). We see a negative skewness (-0.323 ± 0.147, $z = 2.192$, thus violating normality) and a positive kurtosis (0.175 ± 0.293, $z = 0.597$), which indicate that the data was significantly shifted towards the "working hard" side with slightly less heavy tails than a normal distribution.

For Round 2, we obtained similar results for "Physical Demand", "Temporal Demand", and "Effort" relative to Round 1 (see Figure 7). The "Performance" data was not found to be normally distributed (Kolmogorov–Smirnov test, $p < 0.001$). Here, we saw a positive skewness (-0.079 ± 0.145, $z = 0.544$), so compared to Round 1, the responses had slightly moved to the "successful" side of the scale, and, again, a negative kurtosis (-0.588 ± 0.290, $z = 2.027$, thus violating normality). This seems to indicate that the participants had gained some traction to move from perceived slight underperforming to perceived slight overperforming. As we do not have any other data to triangulate with, we cannot do more than speculate that this might be due to some habituation effect.

B.6 RQ-Emotion

In post-intervention surveys, we also administered a three-item survey to capture the students' emotional responses [22].[6] It asked students to compare their experience with Porpoise against their experience with "traditional" homework assignments, on a scale from 1 ("Much more true of traditional assignments") to 7 ("Much more true of Porpoise"):

- Upon completing the assignment, I felt proud/accomplished [PROUD].
- While working on the assignment, I often felt frustrated/annoyed [FRUSTRATED].
- While working on the assignment, I felt dictionte/stupid [STUPID].

[6] The original survey [22] contains a fourth item directed at self-efficacy, which was not the focus of our evaluation.

Figure 8 summarizes the results. Test for normality using the Kolmogorov–Smirnov test showed that for none of the questions the responses in Round 1 were normally distributed—which would have indicated that, by and large, traditional assignments evoked the same emotions as the assignments using Porpoise. Instead, the positive skewness together with the z-score for proud indicates that students felt significantly more proud about their achievements when working on traditional assignments. On the other hand, there was a slight tendency to feel more frustrated when working with Porpoise and a much broader range of "frustrating experiences" in both conditions (frustrated). The data for stupid was not normally distributed but did not show a clear tendency.

This general trend was confirmed in Round 2: Students felt more proud about their accomplishments in traditional assignments and had a very broad variety of where they experienced frustration. In this round, it became more clear, though, that working with Porpoise induced significantly less feelings of stupidity or inadequacy: The skewness towards "more true for traditional homework assignments" was statistically significant.

Received 2025-06-30; accepted 2025-08-01