Journal Pre-proofs

Generative AI for scalable feedback to multimodal exercises

Lukas Jürgensmeier, Bernd Skiera

PII: S0167-8116(24)00043-0
DOI: https://doi.org/10.1016/j.ijresmar.2024.05.005
Reference: IJRM 1581

To appear in: *International Journal of Research in Marketing*

Received Date: 12 February 2024

Please cite this article as: L. Jürgensmeier, B. Skiera, Generative AI for scalable feedback to multimodal exercises, *International Journal of Research in Marketing* (2024), doi: https://doi.org/10.1016/j.ijresmar.2024.05.005

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 Published by Elsevier B.V.

Generative AI for Scalable Feedback to Multimodal Exercises

Lukas Jürgensmeier & Bernd Skiera

==========

ARTICLE INFO

Article history

First received on 12 February 2024 and was under review for 2 months.

Area Editor: Michael Haenlein

Lukas Jürgensmeier, Ph.D. Student, Faculty of Business and Economics, Goethe University Frankfurt am Main, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany, email: juergensmeier@wiwi.uni-frankfurt.de.

Bernd Skiera, Full Professor, Chair of Electronic Commerce, Faculty of Business and Economics, Goethe University Frankfurt am Main, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany, email: skiera@wiwi.uni-frankfurt.de. Corresponding author.

Bernd Skiera is also a Professorial Research Fellow at Deakin Business School, 221 Burwood Highway, Burwood, VIC 3125, Australia.

Declarations of interest: none.

Acknowledgments

We thank the IJRM co-editor, David Schweidel, the Area Editor, and three anonymous reviewers for excellent feedback. Additionally, we thank Karim Zibo, Matti Vennen, and David Miesner from Zavi AI for their support in developing the web application, which is available through StudyLabs.ai. We also thank Jan Bischoff, Jost Kaufmann, and Ryan Grabowski for their excellent support of the project.

Funding

This work benefitted from funding by the "efl – the Data Science Institute" and the "Digital Teaching and Learning Lab" (DigiTeLL), a project at Goethe University Frankfurt funded through Stiftung Innovation in der Hochschullehre (Foundation for Innovation in Higher Education).

Journal Pre-proofs

Accepting Editor: David Schweidel

==========

Journal Pre-proofs

Generative AI for Scalable Feedback to Multimodal Exercises

Detailed feedback on exercises helps learners become proficient but is time-consuming for educators and, thus, hardly scalable. This manuscript evaluates how well Generative Artificial Intelligence (AI) provides automated feedback on complex multimodal exercises requiring coding, statistics, and economic reasoning. Besides providing this technology through an easily accessible web application, this article evaluates the technology's performance by comparing the quantitative feedback (i.e., points achieved) from Generative AI models with human expert feedback for 4,349 solutions to marketing analytics exercises. The results show that automated feedback produced by Generative AI (GPT-4) provides almost unbiased evaluations while correlating highly with (r = .94) and deviating only 6% from human evaluations. GPT-4 performs best among seven Generative AI models, albeit at the highest cost. Comparing the models' performance with costs shows that GPT-4, Mistral Large, Claude 3 Opus, and Gemini 1.0 Pro dominate three other Generative AI models (Claude 3 Sonnet, GPT-3.5, and Gemini 1.5 Pro). Expert assessment of the qualitative feedback (i.e., the AI's textual response) indicates that it is mostly correct, sufficient, and appropriate for learners. A survey of marketing analytics learners shows that they highly recommend the app and its Generative AI feedback. An advantage of the app is its subject-agnosticism—it does not require any subject- or exercise-specific training. Thus, it is immediately usable for new exercises in marketing analytics and other subjects.

Keywords: Generative AI, Automated Feedback, Marketing Analytics, Learning, Education

Journal Pre-proofs

1. Introduction

The saying "data is the new oil" reflects the notion that data has become an immensely valuable resource in today's digital era—much like oil was during the Industrial Age. In marketing, the advent of innovative data sources—such as social media platforms, online transaction records, and sensory technologies—has transformed the field into a data-rich domain. Similar to oil, however, data requires extraction, refinement, and effective utilization to unlock its value. Consequently, marketing analytics, also called "data science in marketing" or "quantitative marketing," has become crucial for informing management decisions and improving firm performance (Germann, Lilien, and Rangaswamy 2013; McAfee and Brynjolfsson 2012).

By virtue of its data-intensive nature, marketing analytics often involves coding statistical models and interpreting the results to provide marketing insights. Hence, applying such quantitative methods in firms and research requires a solid quantitative education of current learners (i.e., students), future decision-makers, and researchers. Marketing analytics course syllabi and the expectations of industry executives towards new graduates reflect that such quantitative skills are in high demand (Liu and Burns 2018).

Educators frequently succeed in their teaching methods by engaging learners in problem-solving exercises and providing constructive feedback. This interactive approach —also referred to as formative feedback—reinforces understanding and fosters an environment where feedback becomes a powerful tool for learning success (Dixson and Worrell 2016).

However, providing feedback on exercises is a significant challenge in teaching marketing analytics and other disciplines. As these exercises become complex, delivering feedback becomes increasingly time-consuming and, thus, costly. However, neglecting to provide ample feedback to learners risks stunting their learning progression.

A potential solution to this problem is providing learners with scalable feedback through Generative AI. One prominent branch of Generative AI includes Large Language Models (LLMs) based on the breakthrough Generative Pre-Trained Transformer (GPT) framework (Radford et al. 2018). Now popularized through products such as ChatGPT, the underlying technology is capable of general-purpose language generation because the models learn the statistical relationship of words in natural language based on vast amounts of text and a subsequent training process (Radford et al. 2019).

Because Generative AI is a general-purpose technology that can generate coherent text independent of the specific subject, its applications are far-ranging across diverse disciplines, including healthcare, manufacturing, human resources, IT management, and the two subjects relevant to this article—marketing and education (Ooi et al. 2023).

Generative AI can generate text that mirrors what humans write or say, such that this technology can respond to user-defined prompts, including questions or more complex assignments. Essentially, responding to questions is the role that educators typically take on when teaching: Learners ask questions, and the educator responds with answers that foster the learners' understanding of the subject. Hence, we expect that Generative AI will be able to mirror what educators do—provide feedback to learners regarding the correctness of their solution and suggest ways how to improve.

Generative AI has already been subject to a wide-ranging debate in higher education, but also in marketing education specifically (Peres et al. 2023). While much of this discussion focuses on risks, such as ensuring that students do not cheat with Generative AI, this article highlights

Journal Pre-proofs

Generative AI's capability to provide highly accurate feedback at scale as a valuable opportunity for educators and learners.

Today, educators in marketing analytics and other domains typically use human resources to provide such feedback. For example, professors and teaching assistants answer questions about exercises and case studies in class, score take-home assignments, or guide learners during office hours. Some eLearning platforms (e.g., DataCamp) offer instant feedback on whether a learner's code produces the correct result. However, this approach focuses on an environment in which educators or software providers foresee the incorrect solutions a learner might provide and generate the respective feedback on this basis (e.g., by suggesting a particular way to correct a mistake). Such feedback works well for smaller exercises but likely reaches a limit to its feasibility when the exercises are large and complex, requiring personalized feedback for learners' specific solutions. In addition, even if an educator is capable of generating a wide range of feedback addressing a large number of specific solutions, it does not scale in the sense that it does not help to provide feedback to other exercises. Therefore, generating appropriate feedback for new exercises remains a cumbersome task.

We aim to address these challenges and enhance marketing education by introducing and evaluating an easy-to-use feedback web app based on Generative AI. With this purpose in mind, this manuscript describes and provides access to such an app, which educators and learners can use to generate personalized feedback for complex exercises. We then examine how well Generative AI can provide scalable feedback for exercises in marketing analytics, a subject that typically features multimodal exercises requiring coding, statistics, and economic reasoning.

More specifically, we assess how well our app can use Generative AI to provide feedback to learners through three studies that evaluate two types of feedback and their usefulness for learners. Study A evaluates the Generative AI app's quantitative feedback, which resembles the points awarded in an exam setting. Study B looks at qualitative feedback, which resembles the textual or verbal feedback an educator would provide. Study C then evaluates the app's implementation in a classroom setting by surveying learners about their usage experience.

Overall, the results show that the app and underlying technology are promising. Study A evaluates the feedback's quantitative performance by comparing the points awarded to the submitted solution, automatically determined in the app with Generative AI, with human expert assessments, referred to as human scores. We let the app generate feedback to 4,349 answers from 243 learners to 36 marketing analytics questions of varying difficulty and complexity and compare the AI score (i.e., points achieved on each exercise) to human scores. We do so in four different settings to investigate how different design choices—which Generative AI model is used and whether it has access to the exercises' correct solutions—impact the feedback's accuracy.

In the best-performing setting (GPT-4 with correct solution), the AI scores (per learner) correlate strongly with the human scores ($r = .94$), provide almost unbiased overall scores (mean error = - 2.4 points out of 90 achievable points), and produce a mean absolute error of 5.7 points (6% of achievable points). Sixty-three percent—2,766 of the 4,349 sub-exercises—received identical human and AI scores.

Furthermore, we assess the performance and cost of alternative Generative AI models. We compare cost and performance by evaluating for each focal model whether an alternative model exhibits either i) the same or better performance at a lower cost or ii) better performance at a lower or the same cost. This analysis shows that in our setting, educators

Journal Pre-proofs
should use either GPT-4, Mistral Large, Claude 3 Opus, or Gemini 1.0 Pro because these models' combination of costs and performance dominate three other Generative AI models (Claude 3 Sonnet, GPT-3.5, and Gemini 1.5 Pro). The choice among these models then depends on the available budget and the acceptable level of performance.

Beyond receiving feedback in the form of achieved points, learners can benefit from more substantiated qualitative feedback. Such textual feedback explains the quantitative evaluation to learners and suggests ways to improve the solution to obtain a fully correct answer. Thus, we investigate the quality of such qualitative feedback in Study B. In this study, we assess the Generative AI's qualitative feedback in three dimensions—correctness, sufficiency, and appropriateness—and find that most of the feedback is of high quality in all dimensions.

Beyond assessing the quantitative and qualitative feedback, Study C provides evidence of how useful learners find the app and whether it is a worthwhile addition to traditional exam preparation—in our case, an in-class discussion of a mock exam. Surveying 43 undergraduate learners in marketing analytics who used the app to receive feedback on their mock exam solutions, we found standardized usability and technology acceptance scores in the top percentiles. 93% of learners would use the app again, 84% would recommend or highly recommend it to their peers, and 76% find it helpful or very helpful. Almost every surveyed learner (93%) stated that the app adds some (26%) or very much value (67%) to the traditional teaching method of discussing the mock exam in class.

These three studies show that Generative AI can enhance education by providing reasonably accurate feedback to learners without human intervention. While we consider marketing analytics exercises in our empirical studies, a benefit of the app's design and underlying Generative AI models is its subject-agnosticism. These models do not require additional training or fine-tuning for different exercises—neither for marketing analytics nor other subjects' exercises. In combination with the highly usable web interface, educators can adapt the app to their specific educational setting without any technical expertise in Generative AI. Supplying the app with an exercise and (optionally) a correct solution and grading rubric suffices to use the app in other educational domains.

The remainder of this article follows this structure: Section 2 provides an overview of previous research using automated feedback in educational settings. In Section 3, we introduce the app for automated feedback and its underlying technology, based on Generative AI. The subsequent three studies in Sections 4–6 compare the app's quantitative feedback with human grading (Study A), assess the qualitative feedback (Study B), and present survey results from the app's in-class deployment (Study C). We end in Section 7 by discussing the conclusions and implications for educators and learners in marketing and beyond.

2. Previous Research on Automated Feedback and Generative AI in Marketing

Providing learners with personalized formative feedback during the learning process and summative feedback through exam grades is a defining feature of higher education curricula. A likely reason for educators' frequent use of feedback is its potential positive impact on learning success (Gibbs and Simpson 2005). Feedback is essential because it allows learners to assess their current understanding and suggests areas to focus on in order to master the subject (Chickering and Gamson 1987).

Through formative feedback, learners solve exercises, receive personalized feedback from educators, and improve. After such a learning process, educators commonly use summative feedback (or summative assessments) to evaluate a learner's progress through exams (Heron 2011). The use of these two feedback types motivates the decision that our Generative AI app

Journal Pre-proofs
should provide quantitative (i.e., summative) feedback that scores how well the learner performed, as well as qualitative (i.e., formative) feedback that explains the quantitative feedback and suggests how to improve.

While essential, providing personalized feedback to learners is also time-consuming for educators. Nevertheless, automating feedback generation has seen little attention in the marketing context. One notable exception is Czaplewski (2009), who discusses how grading rubrics can support educators in their evaluations through the use of a computer-assisted approach that promises efficiency gains for manual feedback generation.

In higher education, educators commonly use single- or multiple-choice questions along with systems that can automatically score those questions, which is especially useful for large-scale courses with hundreds of learners (Brown and Abdulnabi 2017). Thus, in marketing education, automated feedback has mostly been confined to software such as EvaExam or GradeYourTest, which can assess simple exercise types, such as multiple-choice questions or numerical problems with a well-defined correct answer.

So far, established automated grading systems work well with exercises with one correct answer or a specific answer pattern, as they can compare the learner's response with the correct answer or pattern. However, they may fall short with evaluation tasks that require higher-order cognitive skills, such as interpretation, reasoning, and decision-making. Indeed, we are unaware of automated grading systems in marketing (or other business fields) that focus on exercises of similar complexity to the ones we use in our empirical study.

Beyond marketing education, however, automating feedback is an active research strain, with more than 100 academic articles across many domains studying Intelligent Tutoring Systems between 2008 and 2018 (Deeva et al. 2021). However, these tools often require significant subject expertise, subject-specific training data, or explicit rules. The tools are cumbersome to transfer to other settings (e.g., a different type of exercise) or other domains, as illustrated by the following examples from the literature.

Barnett et al. (1978) describe one of the earlier examples of automating feedback in medicine: assuring patient care quality and auditing the actions of medical personnel through an early computer system. In programming, Singh, Gulwani, and Solar-Lezama (2013) developed a system to provide automated feedback to students in introductory programming classes that spots 64% of errors in students' assignments. However, this system requires a detailed, correct solution and an error model that includes anticipated mistakes and matching corrections.

With the continuous improvement of computer-based feedback systems, the linguist Koltovskaia (2020) assesses how writers interact with automated grammar and spelling correction tools. Zhang and Hyland (2018) compare human writing evaluations with automated ones. While both studies provide important insight into how humans interact with automated feedback, they rest on the limited observations from two students.

Before the recent advances in Generative AI, scholars had been using natural language processing to improve automated feedback systems. In mathematics, Botelho et al. (2023) used Google's BERT to train a model on 150,000 historic exercise answers which then provided automated feedback on open-ended math problems. However, this model is narrowly trained on specific math problems. As a result, it is not readily transferrable to other domains without massive training data.

Journal Pre-proofs

The recent advances in Generative AI promise to solve these challenges because researchers feed massive amounts of data—representing a significant amount of all available knowledge—into training the models. As a result, GPTs appear to have "general knowledge" consisting of significant subject knowledge across all domains (Radford et al. 2019). Thus, such models can rely on a foundational subject expertise without additional training. With the provision of additional context to the model through prompts, such models can perform specialized tasks across domains. Hence, GPT-based feedback systems promise to provide feedback on questions in all domains and could be useful without extensive training.

Early research results suggest that Generative AI explanations can improve learning outcomes (Kumar et al. 2023) and teaching learners how to use Generative AI will better prepare them for marketing jobs that require Generative AI usage (Guha, Grewal, and Atlas 2024). Beyond the educational sphere, research on Generative AI in marketing suggests great potential, for example, in the form of automating copywriting for search engine optimization (Reisenbichler et al. 2022), market research (Brand, Israeli, and Ngwe 2023; Goli and Singh 2024; Li et al. 2024), identification of marketing constructs (Ringel 2023), advertising image generation (Jansen et al. 2023), and emotional customer support (Huang and Rust 2023).

While the initial hype and excitement regarding GPT's capabilities and its widespread adoption promise a paradigm shift in the underlying technology of automated feedback systems, our understanding of how well such systems work under varying conditions remains limited. We aim to address this limitation by developing and evaluating a web-based app using GPT's capabilities and providing guidance for educators on the usefulness of GPT-based automated feedback systems in marketing and beyond.

3. Underlying Technology and Implementation in an App

3.1. Aim of the App

Beyond evaluating how well Generative AI can provide feedback in an educational setting, our contribution is to make this technology widely available through a web app for educators and learners. The app aims to provide feedback to learners' solutions of multimodal exercises in marketing analytics—exercises that involve analyzing data, interpreting the results, and drawing economic conclusions. Hence, the exercises should go far beyond simple multiple-choice tasks and instead offer the possibility of qualitative feedback (*formative feedback*, i.e., advice on how to improve) and quantitative feedback (*summative feedback*, i.e., a score reflecting how well the learner did).

Figure 1 illustrates the app's basic idea by visualizing the process from i) educators defining an exercise to ii) learners supplying a solution and iii) the app generating feedback based on Generative AI. Educators first insert an exercise into the web app, which could be an individual exercise or a set of several, such as a full exam. For example, educators can ask learners to compute the optimal price of a product given historical prices and quantities sold (see Exam A in the later section that features a similar question). Learners then solve this rather complex exercise by setting up a demand function, estimating it with statistical software and supplied data, setting up a profit function based on the estimated demand function, taking its first derivative, and then solving for the optimal price.

Evaluating this multi-step and multimodal exercise requires coding, calculus, and logical reasoning skills. Such complex exercises often have several correct solutions. In addition, the degree of correctness differs for erroneous results (e.g., a minor calculation error versus a wholly incorrect solution). Hence, solely giving feedback on whether the derived optimal price is correct—which would be easy to automate, e.g., through single-choice answer

Journal Pre-proofs
options—is insufficient for learning and assessment purposes. Therefore, we need a more sophisticated approach to providing more granular feedback. In our app, learners simply insert their full answer, which includes the described multimodal input. The app then combines the learner's solution with the exercise and evaluates how correct the learner's solution was. Ultimately, learners receive this feedback through the app.

Figure 1 INSERT APPROXIMATELY HERE

3.2. Access to the App

Readers can test the app by registering at studylabs.ai/en/invite/ijrm. The app enables two user roles, mirrored in two account types: *educator* and *learner* accounts. First, an educator account serves as the administrator. The educator account can insert an exercise into the app, supply the relevant metadata, administer the learner accounts, and see the solutions learners submitted, together with the app's feedback. Learner accounts only see the exercise and can insert their solutions into the app for feedback generation.

The empirical study introduces two marketing analytics exams to test the application. We pre-loaded both exams through the educator accounts. Hence, these exercises are already visible in each learner's account. We encourage app testers to define new exercises and assess how well the app generates feedback.

3.3. Approaches to Providing Feedback

Multiple approaches and design choices are available when designing an app with underlying Generative AI technology that can provide feedback to learners. This manuscript restricts itself to two relevant dimensions, yielding two by two, i.e., four different settings. First, we distinguish the degree of input supplied to the app—with the exercise's correct solution or without it. Second, we distinguish the type of Generative AI model that powers the app and generates the feedback—Open AI's GPT-3.5 (16k) and GPT-4 (8k). The last value in the model description (here: "16k" and "8k") refers to the combined maximum length for model input and output. This value's unit is *tokens*, where 100 tokens correspond to approximately 75 words. Hence, for GPT-3.5-turbo-16k with 16,385 tokens, the input and output combined can consist of approximately 12,000 words (e.g., 8,000 words of input leaves 4,000 words for the output or vice versa).

While there are more possible design choices (e.g., prompt template, reinforcement learning, and few-shot training), we focus on these two dimensions because they allow us to investigate the tradeoffs between flexibility (how much information the educator needs to supply), cost to run the app based on the underlying model, and the resulting feedback accuracy, as outlined in Table 1.

Table 1 INSERT APPROXIMATELY HERE

3.3.1. Input: Correct Solutions vs. No Correct Solution

In Table 1's first row, the app provides feedback on the minimum possible information required—the exercise and the learner's solution to it. This setting is particularly appealing since the app can provide feedback *without* having access to the correct solution. In this setting, the app can provide feedback on any (new) exercise. Users of the app only have to

supply the existing exercise and the learner's solution, which is particularly helpful for learners without access to correct solutions.

In this setting, the feedback on the learner's solution stems from the Generative AI model's comparison of the learner's solution with its understanding of a correct solution to the supplied exercise. Hence, the model's attempt at answering the question rests solely on its "common knowledge," i.e., its general understanding of the world based on the context-unspecific training data. We expect some misalignment between the model's understanding of the exercise (without any information about the context beyond the supplied exercise) and the educator-supplied correct solution. Nevertheless, we expect the model to know the correct answer to some extent—based on the chance that the model has coincidentally "seen" the knowledge required for answering the question during its training. So far, however, we do not know to which degree the model can understand the exercise and assess the learner's solution without having access to the correct solution.

Table 1's second row introduces the setting with the most context: Besides the exercise and the learner's solution, we provide a rubric and a correct solution to the model. Hence, the model can compare the learner's solution with the correct solution, integrating a detailed scoring rubric (in addition to the model's "general knowledge"). Hence, we expect the feedback in this setting to be more accurate and have a lower variance than in the first condition with no solutions.

3.3.2. Generative AI Model: GPT-3.5 (16k) vs. GPT-4 (8k)

The second dimension of design choices in Table 1's columns distinguishes the underlying Generative AI model supplying the feedback. In the first part, this study focuses on two popular models available as of September 2023. First, we use GPT-3.5 (more specifically, GPT-3.5-turbo-16k). According to OpenAI (2023b), this model is currently the "most cost-effective" model in their portfolio. Second, we implement OpenAI's most capable model currently available, GPT-4 (more specifically, GPT-4-8k). This model promises higher accuracy, broader knowledge, and more advanced reasoning capabilities than the provider's other offerings (OpenAI 2023b).

While we expect greater feedback accuracy from GPT-4 compared to GPT-3.5, this improved accuracy comes at a higher cost. Our implemented GPT-4 variant is currently one order of magnitude more expensive than the implemented GPT-3.5 variant. Hence, there is a tradeoff between improved accuracy and cost, which we will quantify in the empirical study.

In the second part, we compare the performance of those models to competing models that have emerged recently. Specifically, we integrate Anthropic's Claude 3 Opus and Claude 3 Sonnet (Anthropic 2024), Google's Gemini 1.5 Pro (Google Gemini Team 2024a) and Gemini 1.0 Pro (Google Gemini Team 2024b), and Mistral AI's Mistral Large (Mistral AI 2024). As of April 2024, these models represent the various firms' most capable available models and less expensive alternative models.

3.4. Back End: From Input via API Request to Output

While the previous subsection described the user-facing front end of the app, this subsection augments this description by detailing the underlying processes in the back end. Figure 2 illustrates the app's back end and traces the process from the user-supplied input to the generated output with feedback.

Journal Pre-proofs
Figure 2 INSERT APPROXIMATELY HERE

The educator supplies the full assignment, which can consist of one or multiple exercises, by inserting the exercise texts and the number of achievable points. Additionally, educators can provide a rubric that assigns points to individual parts of the exercise and the correct solution. After educators set up an assignment, learners are able to see it and can supply their answers to the sub-exercises through the app's front end.

The app then splits the full assignment into *N* subtasks, where *N* is the number of sub-exercises. For every sub-exercise, the app merges all input information with a pre-defined but customizable prompt template ("system prompt") that provides detailed instructions to the Generative AI model. For this study, we provide information on the AI's role (i.e., feedback assistant), the task (generating feedback for learners), and the format (a numeric value for "points_achieved" and a string for constructive feedback). We show this prompt in the Web Appendix.

After merging the user-provided input with the prompt template, the app sends *N* requests to the respective API, which queries a response from the specified Generative AI model. The app then receives the API's response and displays the generated feedback by sub-exercise. This feedback contains the exercise, the learner's solution, and the quantitative and qualitative feedback.

3.5. Multilingual Support

One major advantage over traditional approaches to providing automated feedback is that a Generative AI model naturally supports all major languages if trained on them. For example, OpenAI (2023a) shows that GPT-4 performs best in English but almost equally well in Italian, Afrikaans, Spanish, German, French, and many more. Hence, our app can, by default, accept exercises and solutions in languages other than English and provide feedback in the same language. While the underlying Generative AI model is multilingual by default, our app front end is bilingual in English and German and easily adaptable to other languages. Since the corresponding marketing analytics course is in German, we used German in our empirical studies and present English translations.

In the following three empirical studies, we use the described app and its underlying technology to evaluate how well both can provide feedback in multiple dimensions (Studies A and B) and assess how usable and helpful the app is for learners (Study C).

4. Study A: Performance Comparison of Quantitative Feedback of AI and Humans

4.1. Aim: Evaluating Quantitative Feedback—How Accurately Does the AI Allocate Points?

Study A starts by answering the research question of how well our app provides feedback to marketing analytics learners by comparing the *quantitative* performance of our app's Generative AI feedback with human feedback (in the following abbreviated "AI vs. human"). Quantitative feedback refers to the number of points achieved out of the maximum achievable points per exam or sub-exercise. Hence, we compare AI with human evaluations of learner performance.

Journal Pre-proofs

In this study, we first evaluate the feedback generated in four settings, introduced in Table 1, and identify the setting yielding the best feedback accuracy (AI vs. human scores). Second, we assess the cost–performance tradeoff among seven Generative AI models.

4.2. Context: Two Marketing Analytics Exams—Exercises and Learners' Solutions

Our exercises and corresponding answers stem from two different in-person electronic exams in two consecutive cohorts of a marketing analytics class for second-year undergraduate students ("learners") at a large European university. Both exams contain a variety of levels of difficulty and, crucially for our evaluation, are complex in the sense that they require coding, statistics, and economic reasoning to solve. While this section broadly describes the characteristics and contents of both exams, we refer to this repository (https://github.com/lukas-jue/marketing-analytics-exams) for both full exams. For each exam, the corresponding folders include the exercises, correct solutions, and grading rubrics. The first 14 sub-exercises from Exam A mostly correspond to the case study from Skiera and Jürgensmeier (2024), to which we refer for a detailed discussion of the exam type and the associated course's pedagogical approach.

After administering the exam, the professor and the teaching assistants (Ph.D. students) graded the exams without any support from automated grading systems, yielding the final grades of the class. All learners could review their grading and submit complaints if the grading seemed incorrect. This review led to minor changes to the points achieved in around 1% of the exams. Hence, we consider human grading (after the review and the resulting minor changes) as the benchmark—an expert score reviewed and undisputed by learners. This expert score after the learner review is the closest available reflection of the "ground truth," but it might not be flawless. For example, the person grading the exam is not always the person teaching the course—introducing imprecision into the expert score—or the score for a given answer can be subject to interpretation. Still, we use expert scoring as the most relevant benchmark for feedback through Generative AI because it is, so far, the predominant method of assessing learners' solutions.

Table 2 presents the characteristics and key summary statistics of the two exams' human scores. Across both exams, our corpus contains 243 submissions with 4,349 solutions to sub-exercises.

Table 2 INSERT APPROXIMATELY HERE

Figure 3 visualizes the distribution of the human expert points by exam, showing that out of the 90 achievable points, the submitted solutions cover the entire range of possible outcomes with a few poor, many medium, and some high-quality solutions.

Figure 3 INSERT APPROXIMATELY HERE

Both marketing analytics exams contained a wide range of questions—in terms of required statistical methods, substantive marketing knowledge, and exercise complexity. The exams especially aimed to test learners' understanding of how to use marketing research methods to derive managerially relevant and actionable decision support (Albers 2012). Using the R programming language, learners wrote code, generated outputs, and interpreted the outputs in an economically meaningful way. From low to high exercise complexity, in Exam A, learners

- answered two single-choice questions,
- implemented code to generate simple exploratory data analyses through visualizations and summary statistics to answer basic questions about the variables included in the data set,
- set up an appropriate linear regression to estimate demand curves from the supplied price- and sales data
- interpret the linear regression's estimated coefficients,
- identify and avoid typical problems when working with linear regressions, such as multicollinearity or omitted variable bias,
- use logistic regression to estimate churn probabilities and compute Customer Lifetime Values (CLV) based on customer characteristics,
- use the linear regression's estimates to derive optimal prices and determine how changes in decision variables impact demand and, thereby, the optimal price.

This high variance in exercise complexity and marketing topics enables us to evaluate how well the app can evaluate learners' solutions depending on those characteristics. Exam B featured a similar variety of exercise types and complexities but slightly different topics. The first large exercise in Exam B requires learners to evaluate the effectiveness of TV advertising through an analysis resembling a before-and-after regression analysis based on Eisenbeiss and Bleier (2020). The second exercise requires learners to conduct a conjoint analysis of a streaming service's customers and estimate their preferences for different product types.

Figure 4 shows one sub-exercise of Exam A that illustrates how human experts and the app should evaluate the learner's solution and allocate points to small steps of the rubric.

Figure 4 INSERT APPROXIMATELY HERE

4.3. Method: Performance Metrics to Compare Human vs. AI Evaluation

We evaluate the quantitative feedback through four metrics: mean error, mean absolute error, their standard deviations (which are identical), and costs to generate the feedback. The first metric answers how well the overall AI score aligns with human scores and whether there is a systematic error. For each of the four settings s, i.e., each combination of the Generative AI model and degree of supplied context, we define the *mean error* as

$$\bar{\epsilon}_s = \frac{\sum_{i=1}^{n} (\text{AI Points}_{is} - \text{Human Points}_{is})}{n}, \quad (1)$$

where i refers to a learner's exam submission. The mean error measures the error for the full exam and indicates whether the app's score in setting s is biased. A low value, however, could also occur because the errors across exams cancel each other out. That is why we additionally compute the mean absolute error, MAE_s. The mean absolute error is always larger than or equal to the mean error. Third, we want to quantify the error's distribution by computing δ, the sample standard deviation of the error ($\epsilon_{is} = \text{AI Points}_{is} - \text{Human Points}_{is}$). Fourth, we

Journal Pre-proofs
measure the cost per exam evaluation by recording the number of tokens consumed by each exam request, translating their value into US dollars, and comparing these to the cost of human evaluators.

4.4. Results: Quantitative Performance Evaluation

We start by evaluating the performance for the full exams. This assessment answers the question of whether the feedback of Generative AI provides learners with an accurate overall evaluation of the number of achieved points. Second, we zoom in on the best-performing app setting (i.e., which model and degree of context) and provide performance evaluations on a sub-exercise level. Through this more granular analysis, we aim to discover how well the app performs for various exercise types.

4.4.1. Feedback Accuracy by App Setup

Figure 5 compares the app's scores to those of the human expert for all 243 exams. We run four app settings, varying by the Generative AI model and the degree of supplied exercise context as input. If the app's exam scores were identical to the human expert score, all points would lie on the identity line (AI Points = Human Points). The more the scores deviate from this line, the worse the AI performed.

As expected, the evaluation of the simplest setting—GPT-3.5 without the correct solutions—correlates least with the human score among the four different settings, albeit still relatively highly at r = .75. Notably, the correlation is very similar irrespective of the input context and only increases marginally to r = .76 when supplying the exercises' solutions as additional context. Nevertheless, the scatterplot reveals that the evaluation without correct solutions suffers from a large positive bias of 13.2 points (out of 90 points), which decreases substantially to 3.5 points with the correct solutions as context. While the mean error measures bias, the mean absolute error quantifies how much the AI's score deviates from the human score, irrespective of the direction. The mean absolute error is highest for the simplest setting (GPT-3.5 without solutions) at 15.5 points and decreases by 40% to 9.3 points when we also provide the solution.

A drawback of generating solutions with GPT-3.5 is that it fails in approximately 24% of sub-exercises to generate the allocated points in the defined format. For example, the textual output is "Points: 2" or "Two points" instead of the easily parsable "⌘2⌘" (see the Web Appendix for the detailed prompt that defines this format). While this incorrect formatting is less of a problem in the app because users still see the allocated points, it requires extensive manual parsing for our performance evaluation. Because spot-checks reveal no systematic pattern regarding the exercises for which GPT-3.5 fails to generate the correctly formatted points, we omit those sub-exercises for our performance evaluation.

Figure 5 INSERT APPROXIMATELY HERE

Conversely, the most advanced model (GPT-4) combined with most context (i.e., with correct solutions and a rubric) performs best across the three performance metrics that quantify how accurately the setting evaluates the 90-point exams. This setting exhibits the lowest mean error (- 2.5 points), the lowest mean absolute error (5.7 points), and the lowest error standard deviation (7.1 points). However, this performance comes at the highest cost among the four settings (US$0.79 per exam).

Journal Pre-proofs

An order of magnitude cheaper than the best-performing setting is the one with solution and GPT-3.5. However, this tenfold decrease in cost does not correspond to one-tenth of the performance compared to the top-performing GPT-4 setting. Instead, the performance decrease is 40–63 %, depending on the performance metric (40% higher mean error at 3.5 points, 63% higher mean absolute error at 9.3 points, and 61% higher standard deviation at 11.4 points). Hence, the marginal returns to increased cost seem to diminish.

While the cost of evaluating a whole class with the best-performing model at US$0.79 per exam can be non-trivial, those costs are still an order of magnitude lower than human expert evaluation. For the two considered exams, human expert evaluators (Ph.D. students) took approximately 20 minutes to grade each exam in our setting. At our large European public university, this time corresponds to gross labor costs of approximately US$12—fifteen times higher than the best-performing automated evaluation setting (GPT-4 with solutions) and 170 times higher than the second-best setting (GPT-3.5 with solutions). Additionally, evaluating a single exam automatically through the app takes less than one minute, compared to around 20 minutes by human experts.

We use the above costs to estimate (although with great uncertainty) the cost savings for educational institutions. For each feedback generation on a learner's answers (here: answers to a full exam or a similarly comprehensive case study), the use of Generative AI in our setting saves approximately US$11.21 ($12 - $0.79), representing a 93% cost reduction. Let us further assume that a medium-sized university has 10,000 learners (i.e., students), each student takes five courses per semester, and each student receives personalized feedback only once per course per semester from a teaching assistant paid US$12 for generating such feedback. Generative AI feedback could then save *10,000 students × US$11.21 × 5 courses per semester × 2 semesters per year = US$1.1 million* per year. If we extrapolate these values to the approximately 19 million enrolled students in the US in 2024 (National Center for Education Statistics 2023), the potential cost savings across the US would be more than US$2.1 billion, equivalent to more than 100,000 scholarships at US$20,000 per year.

These calculations assume that educators substitute their human feedback with Generative AI feedback. In addition, it also assumes that students only received personalized feedback once per course. Generative AI, however, could also enable a much higher volume of personalized feedback to learners. Hence, Generative AI enables educators to provide personalized feedback on every submission at a comparatively low cost, which might be particularly relevant for courses featuring many case studies or mock exams. Hence, educators can greatly scale the feedback quantity.

4.4.2. Feedback Accuracy by Alternative Models

So far, we have compared the performance of two very popular Generative AI models from OpenAI. However, other firms have recently started offering alternative Generative AI models. Therefore, we proceed by comparing the performance and costs of the two models from OpenAI with five additional competing Generative AI models—Anthropic's Claude 3 Opus and Claude 3 Sonnet, Google's Gemini 1.5 Pro and Gemini 1.0 Pro, and Mistral AI's Mistral Large.[1] We focus in this comparison on the setting that provides the correct solution (see Table 1) because it performed best in the previous assessment.

---

[1] The app uses the following exact models and integration methods:
Claude 3 Opus: claude-3-opus-20240229 via Anthropic API;
Claude 3 Sonnet: claude-3-sonnet@20240229 via Google Cloud;
Gemini 1.5 Pro: gemini-1.5-pro-preview-0409 via Google Cloud;

Journal Pre-proofs

Figure 6 summarizes and compares the performance of the seven Generative AI models in terms of the previously introduced metrics. GPT-4 features the highest accuracy of feedback scores as measured by the correlation with human feedback, the error's standard deviation, and the mean absolute error. While Claude 3 Opus scores worse in these metrics, its mean error (i.e., the bias on a full exam level) is closest to zero at -0.2 points out of 90 achievable points.

Thus, our app performs best across most performance metrics when using GPT-4 out of the seven considered models. However, users choosing between the seven tested models face the same cost–performance tradeoff as when choosing between the lower-performance and cheaper GPT-3.5 and the higher-performance but comparatively expensive GPT-4.

Figure 6 INSERT APPROXIMATELY HERE

Relating the cost and performance of Generative AI models to each other enables us to determine which models are efficient. In our scenario, a model is efficient (i.e., no other model dominates such model) if none of the other models exhibits i) a better performance at the same or lower cost or ii) at least the same performance at a lower cost. Conversely, a model is inefficient if another model dominates it. Thus, we should only select one of the efficient models for generating feedback.

Figure 7 visualizes the relationship between the cost of the models' feedback and their performance in terms of the four previously introduced performance metrics. Additionally, Figure 7 draws the efficiency frontier. Models positioned in a corner of this frontier are efficient. We should not choose a focal model below or above this frontier (depending on whether a low or high value of the performance metric indicates good performance) because there exists an alternative model with either i) or ii)—in other words, an alternative model dominates the focal one. The same holds for models on the efficiency frontier but not on one of its corners. For example, in the first panel of Figure 7, the feedback generated by Claude 3 Sonnet correlates with human feedback to the same degree as Gemini 1.0 Pro, but at a higher cost. Thus, using Claude 3 Sonnet would be inefficient, and we should favor Gemini 1.0 Pro.

Further analyzing Figure 7 suggests that the models GPT-4, Mistral Large, Claude 3 Opus, and Gemini 1.0 Pro are efficient if one considers all four performance metrics jointly. Conversely, the remaining models—Claude 3 Sonnet, GPT-3.5, and Gemini 1.5 Pro—are dominated by alternative models. Hence, we should not implement these dominated models if the cost and performance of providing quantitative feedback are the only evaluation criteria. Nevertheless, there might be other criteria relevant to such a decision that remain unaccounted for in this analysis (e.g., the *qualitative* feedback performance; see Study B).

Figure 7 INSERT APPROXIMATELY HERE

4.4.3. Determinants of Feedback Accuracy

The previous section discussed the four settings' performance metrics and the tradeoff between cost and performance on the aggregate exam level. This section now focuses on the

---

Gemini 1.0 Pro: gemini-1.0-pro via Google Cloud;
Mistral Large: mistral-large-2402 via Microsoft Azure.

Journal Pre-proofs

best-performing setting (GPT-4 with solutions) and presents its performance depending on exercise complexity. This detailed examination indicates which type of sub-exercises the app can evaluate and how well. We use the maximum achievable points per sub-exercise as a proxy for exercise complexity. Combined, our two exams feature 17 + 19 = 37 different sub-exercises with between 2 and 20 achievable points.

Because this evaluation now requires comparing exercises with different achievable maximum points, we cannot use the previously introduced performance metrics based on absolute errors. Hence, we define the error for learner i's solution to sub-exercise j, relative to the maximum achievable points per sub-exercise j, as:

$$Relative\ Error_{ij} = \frac{AI\ Points_{ij} - Human\ Points_{ij}}{Achievable\ Points_j} \quad (2)$$

This change from exam (N = 243) to the sub-exercise level increases the number of observations to M = 4,349. The app evaluated 2,766 of the 4,349 exercises (63%) without any error, and 83% (3,605 out of 4,349) of the sub-exercises had up to a 25% error relative to the maximum number of achievable points. Figure 8 visualizes this performance distribution across all sub-exercises.

Figure 8 INSERT APPROXIMATELY HERE

While Figure 8 visualizes the sub-exercise evaluation performance across all exams, Figure 9 visualizes the performance distributions by maximum achievable points per exercise. This more granular view enables us to assess how heterogeneous the app's performance is, i.e., how well the app performs depending on exercise complexity.

Noteworthy is that all ten categories of varying achievable points feature a median relative error of 0%. For the low-complexity exercises with 2–4 achievable points, the 25th and 75th percentiles additionally lie at 0% relative error. The relative error distribution widens slightly for more complex sub-exercises, although the distribution remains relatively narrow. This heterogeneity analysis suggests that the app's performance is slightly more variable in the case of more complex exercises but remains relatively high even for complex sub-exercises.

Figure 9 INSERT APPROXIMATELY HERE

4.4.4. Feedback Consistency

Because Generative AI systems are non-deterministic and generate different responses to identical requests, feedback accuracy over multiple requests might deviate. This non-deterministic characteristic of Generative AI reflects the same non-deterministic characteristic of humans providing feedback—the same human can grade an exercise differently at two points in time, or two humans can grade the same exercise differently. Therefore, we compare the consistency of two identical Generative AI feedback requests at two points in time.

Specifically, we re-run the best-performing setting (GPT-4 with solutions) in April 2024 and compare it to the initial results discussed above (from September 2023). With r = 0.92, this

Journal Pre-proofs
second round of feedback generation correlates highly, but not perfectly, with the results from the first. When comparing the quantitative feedback of humans vs. the Generative AI, the second round of feedback generation is less accurate than the first (MAE = 7.3 vs. 5.7). This comparison of the feedback's consistency highlights the non-deterministic nature of Generative AI models, yielding high but not perfect consistency.

5. Study B: Expert Performance Evaluation of Qualitative AI Feedback

5.1. Aim: Evaluating Qualitative Feedback—How Do Experts Assess the AI's Textual Feedback?

While Study A assesses the quantitative feedback, Study B turns toward the qualitative feedback evaluation. Hence, this study aims to assess how well our app and the underlying Generative AI provide textual feedback to learners.

5.2. Method: Expert Evaluation of Qualitative AI Feedback

5.2.1. Theoretical Background: Assessment Dimensions of Qualitative Feedback

Through Study B, we aim to assess the qualitative—i.e., the textual—feedback. To achieve this aim, we first derive relevant dimensions for a human expert to evaluate the texts. We derive these by considering which dimensions characterize good feedback.

Specifically, we build upon Gibbs and Simpson's (2005) ten conditions under which feedback supports learning. From those ten conditions, we identify three dimensions relevant to our study and define them, as detailed in Table 3. First, we let the expert evaluate the qualitative feedback's *correctness,* meaning the degree to which it is factually correct. Second, we assess the feedback's *sufficiency*, i.e., whether the text justifies how the Generative AI arrived at a given quantitative feedback. Third, we are interested in the *appropriateness* of the feedback in that it can guide learners toward the correct solutions through constructive advice on how to improve.

The remaining seven conditions from Gibbs and Simpson (2005) are less relevant to assessing our app because they are independent of the feedback's quality. For example, these conditions concern the underlying exercise's quality (whose assessment is beyond this study's scope) or how quickly learners receive the feedback (which occurs almost instantly through our app).

Table 3 INSERT APPROXIMATELY HERE

5.2.2. Expert Evaluation of Qualitative Feedback

We let one expert assess the generated feedback. This expert also graded both exams. Hence, this expert is familiar with the exercises and their solutions and can capably assess the quality of the automatically generated feedback. This dual role ensures consistency when assessing the extent to which the learner's solution is correct, enabling a valid assessment of the Generative AI's feedback. Still, there might be some drawbacks because, for example, errors in grading might be persistent.

To evaluate the Generative AI's textual feedback, the expert compares the learner's solution to an individual sub-exercise with the Generative AI's feedback along the three dimensions from Table 3. Considering the sample solution and the grading rubric, the expert then assigns

Journal Pre-proofs
a score to each dimension on a five-point Likert scale, ranging from "very low" to "low," "medium," high," and "very high."

5.3. Context: Sample of Best-Performing AI Feedback

Because letting human experts assess the app's qualitative feedback takes approximately 20 minutes per exam, we face capacity constraints that prevent us from evaluating the full sample from Study A in four different settings. Hence, this study restricts itself to a random subsample of 20 submissions from Exam A, comprising 340 sub-exercises. Furthermore, we only use the feedback generated by the best-performing setting using GPT-4 and the correct solutions.

5.4. Results: Expert Evaluations

Figure 10 and Figure 11 summarize Study B's results. Figure 10 shows the distribution of expert evaluations across all considered sub-exercises for each of the three feedback dimensions: sufficiency, correctness, and appropriateness. This distribution of expert evaluations shows that in most evaluated sub-exercises, the qualitative feedback of our app is of very high quality across all three dimensions.

According to the expert, the correctness of the app's qualitative feedback was "very high" ("high") in 66% (21%) of cases. Less than 13% of feedback instances were of medium, low, or very low correctness. This finding implies that the high degree of qualitative feedback accuracy (as portrayed in Study A) stems from a factually correct assessment of the solution's merit.

In 90% of sub-exercises, the expert deemed the feedback of "very high" sufficiency, with less than 2% of medium or (very) insufficient feedback. In a small minority of cases, the Generative AI fails to accurately justify why it awarded the points for a sub-exercise.

Similarly, 80% of the considered feedback was characterized as very highly appropriate, and the expert evaluated less than 6% of the feedback as medium or (very) inappropriate in guiding the learners to the correct solution. Hence, the vast majority of feedback enables learners to understand how to improve their understanding. While the expert evaluations of the qualitative feedback are, overall, very positive across all three dimensions, the highest potential for improvement is in the *correctness* dimension.

Figure 10 INSERT APPROXIMATELY HERE

Figure 11 then drills down the expert scores from Figure 10 by the number of achievable points for the assessed exercise. Analyzing the expert scores by the number of achievable points enables us to cautiously assess the relationship between exercise complexity (proxied by the number of achievable points) and the three expert dimensions. The results show differences across varying numbers of achievable points for each sub-exercise. Some of the more complex exercises (with more achievable points) score slightly worse than the less complex exercises, although this difference does not hold consistently across all dimensions.

Figure 11 INSERT APPROXIMATELY HERE

Journal Pre-proofs

Journal Pre-proofs

6. Study C: Learners' Evaluation of App and AI Feedback

6.1. Aim: Evaluating Usage In-Class—How Do Learners Interact with the App and AI Feedback?

While the first two studies evaluated the feedback accuracy—quantitative feedback in Study A and qualitative feedback in Study B—this third Study C aims to evaluate how learners interact with the app and thereby determine the perceived usefulness of Generative AI feedback.

6.2. Context: Learners Using the App to Receive Feedback on a Mock Exam

We achieve this aim by letting learners in an intermediate undergraduate marketing analytics class (at the same university as in Study A) test the app. In the middle of the semester (December 2023), learners received a mock exam to self-assess their acquired skills and familiarize themselves with the exam format. In previous semesters, learners had not received any personalized feedback on their solutions because the required workload for educators would have been too high. Hence, learners only received general feedback by comparing their solutions with the sample solution and by attending a discussion by the professor, explaining some of the challenging exam parts.

Learners had one week between two class sessions to solve the mock exam. In the exercise, learners econometrically analyze a data set with sales quantities, prices, and advertising budget variables and answer substantive marketing questions, such as setting the optimal prices. This mock exam corresponds to the first exercise of Exam A, accessible through the accompanying repository.

After learners had the chance to work on the exam questions for one week, the professor discussed the exercise and its key challenges with the class for approximately 45 minutes. After that, learners received one-time login credentials for our app. Learners then uploaded their own mock exam solutions to the app and assessed the personalized feedback they received.

6.3. Method: Survey Measuring the App's Standardized Usability and Usefulness

Subsequently, and still during the class, we directed learners to a survey asking about their experience with the app. We designed the study to assess how learners interact with the app and how useful they perceive its personalized feedback to be.

The survey sample consists of 43 undergraduate learners attending the marketing analytics class. The median survey participant is in the 4th semester (the scheduled completion time for an undergraduate degree is six semesters). 86% took this course as part of the bachelor's program "Business and Economics," with the remainder taking it as part of a business minor from other degrees. 84% of the respondents were taking the course for the first time. Only two of the 43 learners had never used Generative AI before, and 77% use it at least monthly. While we did not specifically ask about the learners' age and gender in this survey, the (separate) course evaluation from the respective semester suggests that the median age of course evaluators is between 21 and 22 years. Approximately 53% of the course's evaluators are female.

The survey evaluates two core aspects of the app—the user experience and the feedback quality—in three survey parts. For the user experience, we measure the constructs of two commonly used frameworks measuring how learners interact with the app. First, we measure the System Usability Score (SUS) to assess how usable the app is (Brooke 1996). We present

Journal Pre-proofs

the survey items and SUS formula in the Web Appendix. Second, we use the Technology Acceptance Model (TAM) to measure the app's perceived usefulness, perceived ease of use, and learners' behavioral intent (Davis 1989). We detail the survey items and computation of the three TAM metrics in the Web Appendix.

Third, we evaluate the feedback quality by asking survey participants specific questions about the feedback's perceived correctness, completeness, comprehensibility, and helpfulness for exam preparation. Additionally, we asked participants whether the app adds value and whether they would recommend it to fellow learners (refer to Table 5 for the definition of these items).

6.4. Results: Learners' Evaluation

6.4.1. System Usability Score

Our survey yields a mean System Usability Score (SUS) of 96.0 (SD = 7.3) on a scale from zero to 100. This metric is not readily interpretable, so we compare it with the score's percentiles as described by Sauro and Lewis (2016). Based on this classification, our app falls into the 96–100 percentile range of System Usability Scores. According to another benchmark developed by Bangor, Kortum, and Miller (2008), our app's score is in the fourth quartile of SUS scores from the literature. It falls between excellent and the best imaginable scores. Hence, learners consider the app highly usable. We interpret this result as evidence that the user interface fulfills its purpose well and is highly suitable for providing feedback.

6.4.2. Technology Acceptance Model

Next, we measure the core metrics of the initial Technology Acceptance Model (TAM) (Davis 1989) to provide additional standardized measures of how users perceive the app. Table 4 displays the mean and standard deviation of the three constructs: perceived usefulness, perceived ease-of-use, and behavioral intent (i.e., whether the participants plan to use the app again). We measure all items of the constructs on a five-point Likert scale from "strongly disagree" (= 1) to "strongly agree" (= 5) and provide a measure of the constructs' internal consistency (Cronbach's α).

Table 4 INSERT APPROXIMATELY HERE

Beyond the constructs' high mean values (all above four on a five-point Likert scale), 77% of survey participants "strongly agree" or "agree" that the app is useful (perceived usefulness), 95% (strongly) agree that the app is easy to use (perceived ease-of-use), and 88% (strongly) agree that they will use it again (behavioral intent).

6.4.3. Granular Feedback Assessment

While the System Usability Score and the measurements from the Technology Acceptance Model's constructs provide standardized measures for comparison with other apps, we further want to assess the app's merit in the specific use case—providing feedback to marketing analytics learners in an undergraduate course. Hence, Table 5 summarizes respondents' answers to five statements relevant to this scenario.

Notably, almost all the surveyed learners agree or strongly agree that they perceive the app's feedback as correct (93%). Additionally, a large majority would (strongly) recommend the app to their fellow learners (84%). Of the 43 surveyed learners, 77% state that the app adds or

Journal Pre-proofs
strongly adds value to the traditional teaching approach—i.e., discussing the exam solution in class.

Slightly more than half of all surveyed learners (58%) agree or strongly agree that they fully understood the feedback (comprehensibility). Hence, these results suggest that there might be room for improvement in providing more comprehensible textual feedback. Nevertheless, when interpreting this result, we have to consider that learners had little time in the class to digest the feedback, which might result in underestimating the feedback's comprehensibility.

Table 5 INSERT APPROXIMATELY HERE

6.4.4. Qualitative Feedback Assessment

Beyond asking learners to answer standardized questions, we also prompted them to answer qualitatively by providing optional fields for "positive comments," "negative comments," and "general comments." 49% (21 out of 43) left a positive, 41% (18 out of 43) left a negative, and 23% (10 out of 43) left a general comment.

Among the positive comments, several learners explicitly mention that they like the app's constructive qualitative feedback, which suggests ways to improve an incorrect solution. Additionally, several learners mentioned the high level of detail and quality of feedback. Furthermore, several learners stated that the app was intuitive, easy to use, functional, and well-designed.

On the negative side, some learners stated that the app sometimes lacked feedback on individual aspects of an exercise and that some feedback was incorrect. These qualitative comments align with the results of Study B, where we let experts assess the quality of the feedback. This analysis shows that incorrect feedback occurs but is comparatively infrequent. On the technical side, a few respondents complained about the waiting time until the feedback showed up (it takes around one minute to send multiple requests and receive the corresponding answers from the OpenAI's APIs). Additionally, respondents felt that transferring each sub-exercise manually from their coding environment was cumbersome. Some additional suggestions for improvements to the app's display of the resulting feedback are no longer relevant because the current version incorporated this improvement.

Finally, the general comments mostly expressed support for the app. They highlighted that learners would like to use it for their studies (e.g., "astonishing," "please continue developing the app because there is much potential," "nice idea, great implementation," "very recommendable also for other subjects because it promotes self-study").

7. Discussion

7.1. Summary

We described and evaluated an app for automated feedback that uses Generative AI. In three studies, we tested the app's quantitative and qualitative performance and asked learners how usable it is and whether it adds value to the traditional educational approach.

Study A evaluated the quantitative performance by comparing the number of achieved points from the app with human expert evaluations across 4,349 sub-exercise submissions to marketing analytics exams. In four app settings, varying in context and chosen Generative AI model, we show that performance is increased by supplying more context (here: the correct

Journal Pre-proofs
solution) and using more advanced Generative AI models (here: GPT-4). The best-performing setting (GPT-4 with solutions) evaluates almost unbiasedly at a mean error of -2.5 points (out of 90 achievable points, i.e., -2.8%) and a mean absolute error of 5.6 points (i.e., 6.2%).

Furthermore, Study A's comparison of seven Generative AI models shows that there is a cost–performance tradeoff. Typically, better performance is available at a higher cost. Educators need to assess this cost–performance tradeoff for their specific educational setting by considering the desired feedback accuracy and ensuring that the cost of generating this feedback remains within their budget.

Overall, we find that providing automated feedback through the Generative AI app is considerably faster (typically less than one minute per exam) and approximately one (GPT-4) to three (Gemini 1.0 Pro) orders of magnitude less expensive than human evaluators.

While Study A assessed the app's quantitative performance, Study B presented expert assessments of the qualitative (i.e., textual) feedback. The results show that when the app uses GPT-4 with correct exercise solutions, the feedback is most often factually correct, sufficiently addresses all relevant parts of the exercise, and is highly appropriate in helping learners to understand their mistakes and showing them how to improve.

Study C then tested the app in the field and surveyed 43 undergraduate learners submitting their solutions to the app after a professor-led discussion of the mock exam (i.e., the "traditional approach" to providing feedback on the mock exam). High standardized usability scores (e.g., System Usability Score = 96) and overall positive responses show that learners find the app easy to use and expect to benefit from it in their learning journey.

7.2. Conclusions

Our studies' results lead us to conclusions regarding the performance of Generative AI for generating feedback, its costs, and its multilingual support.

First, our result shows that the most advanced model (GPT-4) with the greatest amount of context produces the most accurate feedback. The relatively high accuracy and the major improvement over the previous model (GPT-3.5) lead us to conclude that future GPT versions will further improve our app's performance, although more incrementally, because the GPT-4 feedback is already highly accurate.

Second, Study A's analysis of seven Generative AI models' cost–performance relationship indicates which models educators should consider when providing feedback through the application. In Study A's setting, educators should either use GPT-4, Mistral Large, Claude 3 Opus, or Gemini 1.0 Pro because there exists no alternative model that has either i) better performance at lower or the same cost or ii) at least the same performance at lower cost. Conversely, educators should avoid GPT-3.5, Claude 3 Sonnet, and Gemini 1.5 in this setting because an alternative model exists with either i) or ii).

Third, since Generative AI models are multilingual, our app is multilingual by default and can process exercises and solutions in most major languages and respond in that given language. Hence, our app works internationally, even if the instructional language differs from English. We tested the app in a German-language class with learners submitting German solutions. Because OpenAI (2023a) suggests that GPT-4's performance in English is marginally better than in German, we expect that using the app in English will yield at least similar, or marginally better, performance.

Journal Pre-proofs

Finally, because generating automated feedback through Generative AI is at least one and up to three orders of magnitude cheaper than human expert feedback, using the app in an educational setting might free up educators' resources and save costs. However, especially when using GPT-4, variable costs are non-trivial (US$0.79 for a 90-minute exam, and learners might submit multiple solutions) and increase in direct proportion to the number of learners using the app. Hence, the app is, on the one hand, easily scalable—generating additional feedback imposes no additional *labor* costs on the educator. On the other hand, each feedback request comes with additional costs, albeit much lower than human expert labor costs. Nevertheless, the many alternative models already available suggest intense competition among Generative AI models, which will likely decrease costs and improve performance further.

7.3. Implications

Our results and conclusions give rise to several implications for (marketing analytics) educators, learners, and firms. Most directly, educators can use our app to provide additional and personalized feedback to learners, even in settings that suffer from limited educator resources. Learners, correspondingly, can receive more personalized feedback on their work, which might improve their learning outcomes.

The novel contribution of our article is to show that Generative AI can provide accurate feedback to learners without subject-specific training of the feedback system. This characteristic is a significant departure from previous automated feedback systems, which mostly depended on a subject-specific training process that required extensive data on past exercises, solutions, and human grading. Hence, the availability of an accurate feedback app such as the one we have described and evaluated here could materially increase the frequency of learners receiving feedback on their work. Because the app does not require subject-specific training and instead works with the educator providing only the exercise and the correct solution, the app might be well-suited and ready-to-use in other educational settings across many levels (e.g., schools, universities, professional training) and subjects (beyond marketing analytics).

Nevertheless, the Generative AI feedback is not always perfectly accurate. Analyzing these deviations suggests in which settings and under which conditions educators can confidently use the app. Educators should be transparent with learners that the app and its underlying technology can make mistakes—similar to humans—and can provide suboptimal or incorrect feedback. Hence, the received feedback serves as an indication but not a final judgment of the submission's merit. This precaution also implies that the app is most suitable for providing *formative feedback*, i.e., feedback during the learning process that helps learners to better understand a given subject.

Simultaneously, the results imply that educators, so far, cannot fully rely on our app to always provide accurate feedback. Hence, the app should not be the sole method of providing *summative feedback* to learners when feedback errors have important consequences—for example, when grading an exam.

While useful for educators and learners, our results also suggest implications for other settings in which humans typically receive feedback to learn a new skill. For example, firms continuously need to train new and existing employees in firm-specific tasks, software, or regulatory requirements. This process typically includes feedback from colleagues, which might be expensive in terms of labor costs. To do so, firms or external learning platforms such as Coursera, DataCamp, and Udemy can offer learning exercises and automated feedback when training learners on a particular task or software.

Journal Pre-proofs

From a learner's perspective, our results imply that Generative AI can be a valuable addition to the learning process and accelerate a learner's understanding of a given subject through immediate and reasonably accurate feedback. Especially when human expert feedback is unavailable or too slow, our app's immediate feedback can eliminate roadblocks, such as a learner missing a key detail for a correct solution.

Through the link studylabs.ai/en/invite/ijrm, the app is available for educators and learners. While this stand-alone app can be integrated into any course (simply by linking to the website), educators or technology providers in higher education can follow our described approach to build a similar app according to custom requirements. Thereby, universities or learning software providers can, for example, integrate such an app natively into their learning management system.

7.4. Limitations and Further Research

While this manuscript evaluated four settings with different app design choices, there are other ways in which the app's performance could still be increased. For example, we use a single, only slightly optimized prompt to provide the Generative AI model with instructions. While we define its role, task, and desired output format, further research could systematically evaluate the impact of different prompts on the app's performance.

Additionally, we use off-the-shelf Generative AI models without tuning the underlying model. We expect such fine-tuned models to achieve higher performance, albeit at the potential cost of being less flexible for other subjects. For example, fine-tuning a model for marketing analytics exercises through reinforcement learning (i.e., iteratively letting it evaluate exams and providing feedback on how good those evaluations were) could increase the performance for other marketing analytics exercises but potentially make the app less useful for evaluating other subjects' exercises.

While this article sheds light on how well Generative AI can provide feedback to learners, we do not assess ethical considerations when using such technology to provide automated feedback to learners. Thus, we urge future researchers to assess the ethical implications of feedback generation through Generative AI. For example, there might be scenarios in which the Generative AI feedback misleads learners by stating that a correct answer is incorrect or vice versa. This erroneous feedback could then mislead the learner and provide false certainty that they did or did not sufficiently understand the course materials, which could adversely impact the learner's exam performance.

By evaluating the app's performance, we showed how useful the automated feedback app is and under which circumstances. Testing the app on marketing analytics exercises shows that the app can handle many different input types—from simple single-choice questions to very complex coding exercises analyzing data and correctly interpreting the output economically. Importantly, we did not change the underlying Generative AI model in any way to accommodate marketing analytics exercises beyond providing the exercise and solutions. By design, our app is subject-, exercise type-, and language-agnostic and ready to use for any other exercise outside of marketing analytics. Because we evaluated the app's performance within a specific subject (marketing analytics), we encourage further research comparing the app's performance in other subjects.

Another relevant area for further research could examine the consistency of the Generative AI feedback. Our comparison of the quantitative feedback generated by querying the same model two times shows that the feedback correlates highly but not perfectly. Also, the model's performance compared to human grading varies. However, the grading of the same human

Journal Pre-proofs
can also differ over time—as can the grading of different humans. Hence, further research could compare Generative AI systems' consistency to human graders' consistency.

While this manuscript emphasizes the educational context, the approach can be transferred to other domains. Humans can benefit from feedback before deciding on many relevant choices—e.g., personal life decisions (investments, education, career), managerial decisions (choosing between complex alternatives), or consumption decisions (choosing products under constraints). For each of these examples, humans could improve based on feedback to their proposed decision and underlying reasoning. Through Generative AI models' "general knowledge" of the world, feedback on such decisions could provide a valuable second opinion before implementing them.

Journal Pre-proofs

References

Albers, Sönke (2012), "Optimizable and Implementable Aggregate Response Modeling for Marketing Decision Support," *International Journal of Research in Marketing*, 29 (2), 111–22.

Anthropic (2024), "The Claude 3 Model Family: Opus, Sonnet, Haiku," *Technical Report*, (accessed April 15, 2024), https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Bangor, Aaron, Philip T. Kortum, and James T. Miller (2008), "An Empirical Evaluation of the System Usability Scale," *International Journal of Human–Computer Interaction*, 24 (6), 574–94.

Barnett, G. Octo, Richard Winickoff, Joseph L. Dorsey, Mary M. Morgan, and Robert S. Lurie (1978), "Quality Assurance through Automated Monitoring and Concurrent Feedback Using a Computer-Based Medical Information System," *Medical Care*, 16 (11), 962–70.

Botelho, Anthony, Sami Baral, John A. Erickson, Priyanka Benachamardi, and Neil T. Heffernan (2023), "Leveraging Natural Language Processing to Support Automated Assessment and Feedback for Student Open Responses in Mathematics," *Journal of Computer Assisted Learning*, 39 (3), 823–40.

Brand, James, Ayelet Israeli, and Donald Ngwe (2023), "Using GPT for Market Research," *Working Paper*, (accessed April 3, 2024), https://papers.ssrn.com/abstract=4395751.

Brooke, John (1996), "SUS: a 'Quick and Dirty' Usability Scale," in *Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester and Ian L. McClelland (editors) "Usability Evaluation in Industry,"* London: Taylor & Francis, 189–94.

Brown, Gavin T. L. and Hasan H. A. Abdulnabi (2017), "Evaluating the Quality of Higher Education Instructor-Constructed Multiple-Choice Tests: Impact on Student Grades," *Frontiers in Education*, 2 (24), 1–12.

Chickering, Arthur W. and Zelda F. Gamson (1987), "Seven Principles for Good Practice in Undergraduate Education," *AAHE Bulletin*, 3, 3–7.

Czaplewski, Andrew J. (2009), "Computer-Assisted Grading Rubrics: Automating the Process of Providing Comments and Student Feedback," *Marketing Education Review*, 19 (1), 29–36.

Davis, Fred D. (1989), "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, 13 (3), 319–40.

Deeva, Galina, Daria Bogdanova, Estefanía Serral, Monique Snoeck, and Jochen De Weerdt (2021), "A Review of Automated Feedback Systems for Learners: Classification Framework, Challenges and Opportunities," *Computers & Education*, 162, 104094.

Dixson, Dante D. and Frank C. Worrell (2016), "Formative and Summative Assessment in the Classroom," *Theory Into Practice*, 55 (2), 153–59.

Journal Pre-proofs

Journal Pre-proofs

Eisenbeiss, Maik and Alexander Bleier (2020), "VaycayNation: Driving Website Traffic through Second-Screen Analytics," *Case 9B20A027, Ivey Publishing*.

Germann, Frank, Gary L. Lilien, and Arvind Rangaswamy (2013), "Performance Implications of Deploying Marketing Analytics," *International Journal of Research in Marketing*, 30 (2), 114–28.

Gibbs, Graham and Claire Simpson (2005), "Conditions Under Which Assessment Supports Students' Learning," *Learning and Teaching in Higher Education*, 1, 3–31.

Goli, Ali and Amandeep Singh (2024), "Frontiers: Can Large Language Models Capture Human Preferences?" *Marketing Science*, Forthcoming.

Google Gemini Team (2024a), "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context," *Technical Report*, https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

Google Gemini Team (2024b), "Gemini: A Family of Highly Capable Multimodal Models," *Technical Report*, (accessed April 15, 2024), https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

Guha, Abhijit, Dhruv Grewal, and Stephen Atlas (2024), "Generative AI and Marketing Education: What the Future Holds," *Journal of Marketing Education*, 46 (1), 6–17.

Heron, Gavin (2011), "Examining Principles of Formative and Summative Feedback," *The British Journal of Social Work*, 41 (2), 276–95.

Huang, Ming-Hui and Roland T. Rust (2023), "The Caring Machine: Feeling AI for Customer Care," *Journal of Marketing*, forthcoming.

Jansen, Tijmen, Mark Heitmann, Martin Reisenbichler, and David A. Schweidel (2023), "Automated Alignment: Guiding Visual Generative AI for Brand Building and Customer Engagement," *Working Paper*, (accessed April 15, 2024), https://papers.ssrn.com/abstract=4656622.

Koltovskaia, Svetlana (2020), "Student Engagement with Automated Written Corrective Feedback (AWCF) Provided by Grammarly: A Multiple Case Study," *Assessing Writing*, 44, 100450.

Kumar, Harsh, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman (2023), "Math Education with Large Language Models: Peril or Promise?," *Working Paper*, (accessed December 1, 2023), https://papers.ssrn.com/abstract=4641653.

Lewis, James R. (2018), "The System Usability Scale: Past, Present, and Future," *International Journal of Human–Computer Interaction*, 34 (7), 577–90.

Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary (2024), "Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis," *Marketing Science*, 43 (2), 254–66.

Liu, Xia and Alvin C. Burns (2018), "Designing a Marketing Analytics Course for the Digital Age," *Marketing Education Review*, 28 (1), 28–40.

Journal Pre-proofs

McAfee, Andrew and Erik Brynjolfsson (2012), "Big Data: The Management Revolution," *Harvard Business Review*, 90 (10), 59–68.

Mistral AI (2024), "Au Large: Mistral Large, Our New Flagship Model," (accessed April 15, 2024), https://mistral.ai/news/mistral-large/.

National Center for Education Statistics (2023), "Digest of Education Statistics," (accessed April 2, 2024), https://nces.ed.gov/programs/digest/d23/tables/dt23_303.10.asp.

Ooi, Keng-Boon, Garry Wei-Han Tan, Mostafa Al-Emran, Mohammed A. Al-Sharafi, Alexandru Capatina, Amrita Chakraborty, Yogesh K. Dwivedi, Tzu-Ling Huang, Arpan Kumar Kar, Voon-Hsien Lee, Xiu-Ming Loh, Adrian Micu, Patrick Mikalef, Emmanuel Mogaji, Neeraj Pandey, Ramakrishnan Raman, Nripendra P. Rana, Prianka Sarker, Anshuman Sharma, Ching-I Teng, Samuel Fosso Wamba, and Lai-Wan Wong (2023), "The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions," *Journal of Computer Information Systems*, forthcoming.

OpenAI (2023a), "GPT-4 Technical Report," (accessed December 14, 2023), http://arxiv.org/abs/2303.08774.

OpenAI (2023b), "OpenAI Platform," (accessed September 6, 2023), https://platform.openai.com.

Peres, Renana, Martin Schreier, David Schweidel, and Alina Sorescu (2023), "On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice," *International Journal of Research in Marketing*, 40 (2), 269–75.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018), "Improving Language Understanding by Generative Pre-Training," *Technical Paper*, (accessed April 3, 2024), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019), "Language Models are Unsupervised Multitask Learners," *Technical Paper*, (accessed April 3, 2024), https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Reisenbichler, Martin, Thomas Reutterer, David A. Schweidel, and Daniel Dan (2022), "Frontiers: Supporting Content Marketing with Natural Language Generation," *Marketing Science*, 41 (3), 441–52.

Ringel, Daniel (2023), "Creating Synthetic Experts with Generative Artificial Intelligence," *Working Paper*, (accessed April 15, 2024), https://papers.ssrn.com/abstract=4542949.

Sauro, Jeff and James R. Lewis (2016), *Quantifying the User Experience: Practical Statistics for User Research*, Amsterdam et al.: Morgan Kaufmann.

Singh, Rishabh, Sumit Gulwani, and Armando Solar-Lezama (2013), "Automated Feedback Generation for Introductory Programming Assignments," in *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '13, New York, NY, USA: Association for Computing Machinery, 15–26.

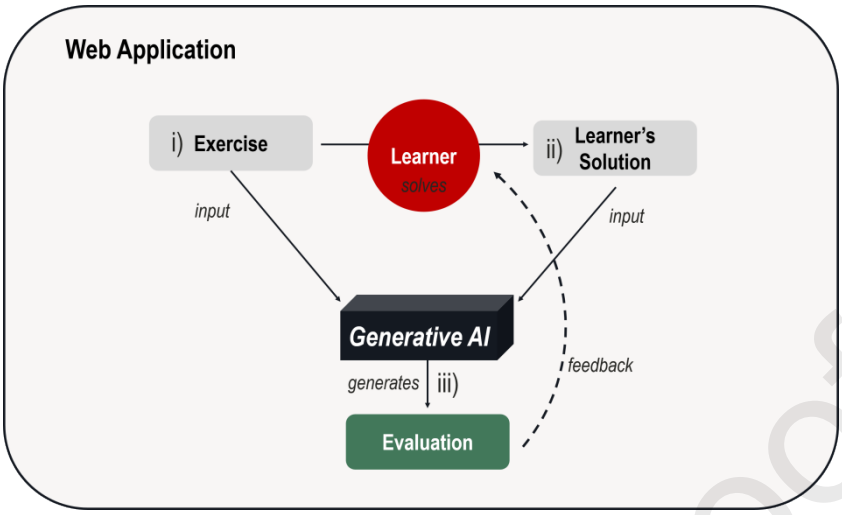
Journal Pre-proofs

Skiera, Bernd and Lukas Jürgensmeier (2024), "Teaching Marketing Analytics: A Pricing Case Study for Quantitative and Substantive Marketing Skills," *Journal of Marketing Analytics*, forthcoming.

Zhang, Zhe (Victor) and Ken Hyland (2018), "Student Engagement with Teacher and Automated Feedback on L2 Writing," *Assessing Writing*, Special Issue: The Comparability of Paper-Based and Computer-Based Writing: Process and Performance, 36 (April), 90–102.

Journal Pre-proofs

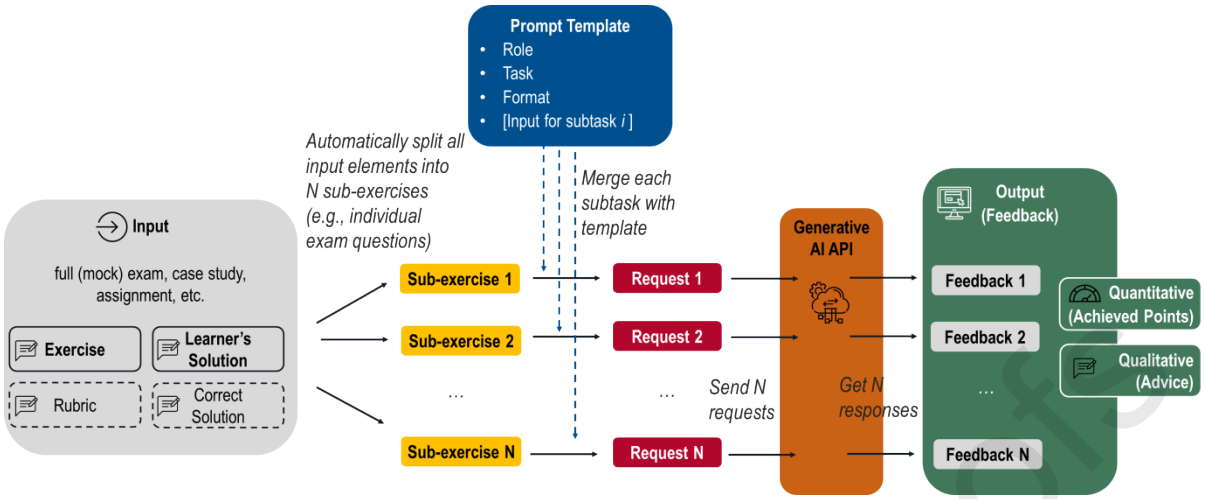
Figure 1: Illustrative App Process Chart to Provide Feedback on Learner's Exercise Solutions



Journal Pre-proofs

Journal Pre-proofs

Figure 2: Illustration of the App's Back End



Journal Pre-proofs

Figure 3: Distribution of Achieved Points by Exam, According to Human Expert

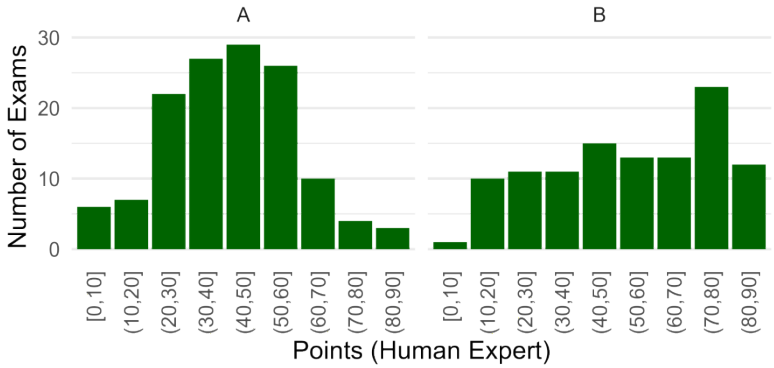
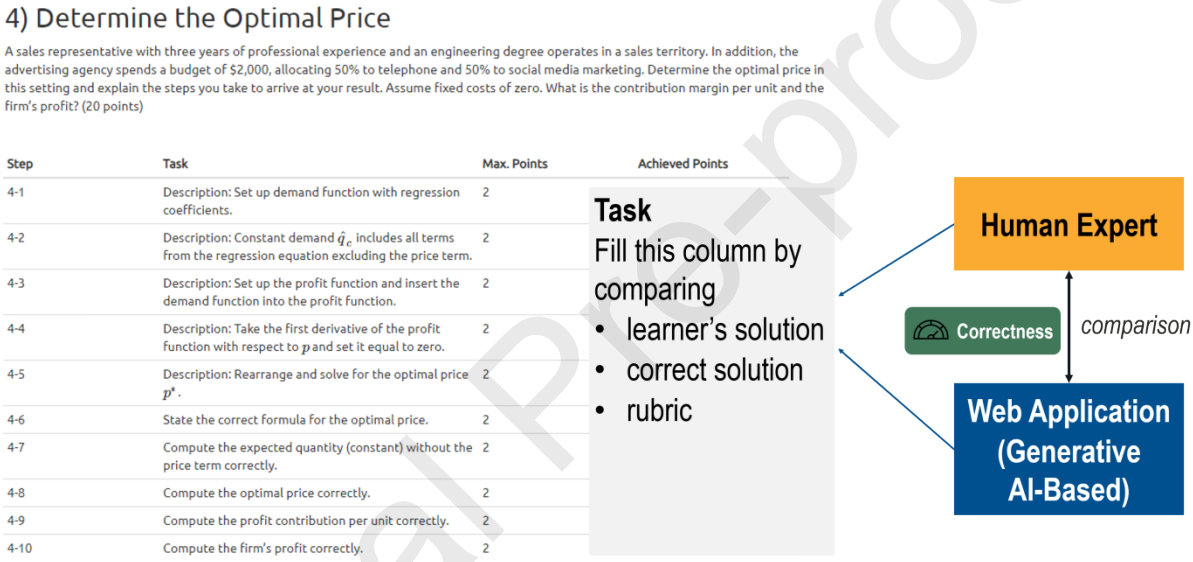


Figure 4: Exemplary Exercise to be Evaluated by Human Experts and the Web App According to the Rubric



4) Determine the Optimal Price

A sales representative with three years of professional experience and an engineering degree operates in a sales territory. In addition, the advertising agency spends a budget of $2,000, allocating 50% to telephone and 50% to social media marketing. Determine the optimal price in this setting and explain the steps you take to arrive at your result. Assume fixed costs of zero. What is the contribution margin per unit and the firm's profit? (20 points)

Step	Task	Max. Points	Achieved Points
4-1	Description: Set up demand function with regression coefficients.	2	
4-2	Description: Constant demand $\hat{q}_c$ includes all terms from the regression equation excluding the price term.	2	
4-3	Description: Set up the profit function and insert the demand function into the profit function.	2	
4-4	Description: Take the first derivative of the profit function with respect to p and set it equal to zero.	2	
4-5	Description: Rearrange and solve for the optimal price p^*.	2	
4-6	State the correct formula for the optimal price.	2	
4-7	Compute the expected quantity (constant) without the price term correctly.	2	
4-8	Compute the optimal price correctly.	2	
4-9	Compute the profit contribution per unit correctly.	2	
4-10	Compute the firm's profit correctly.	2	

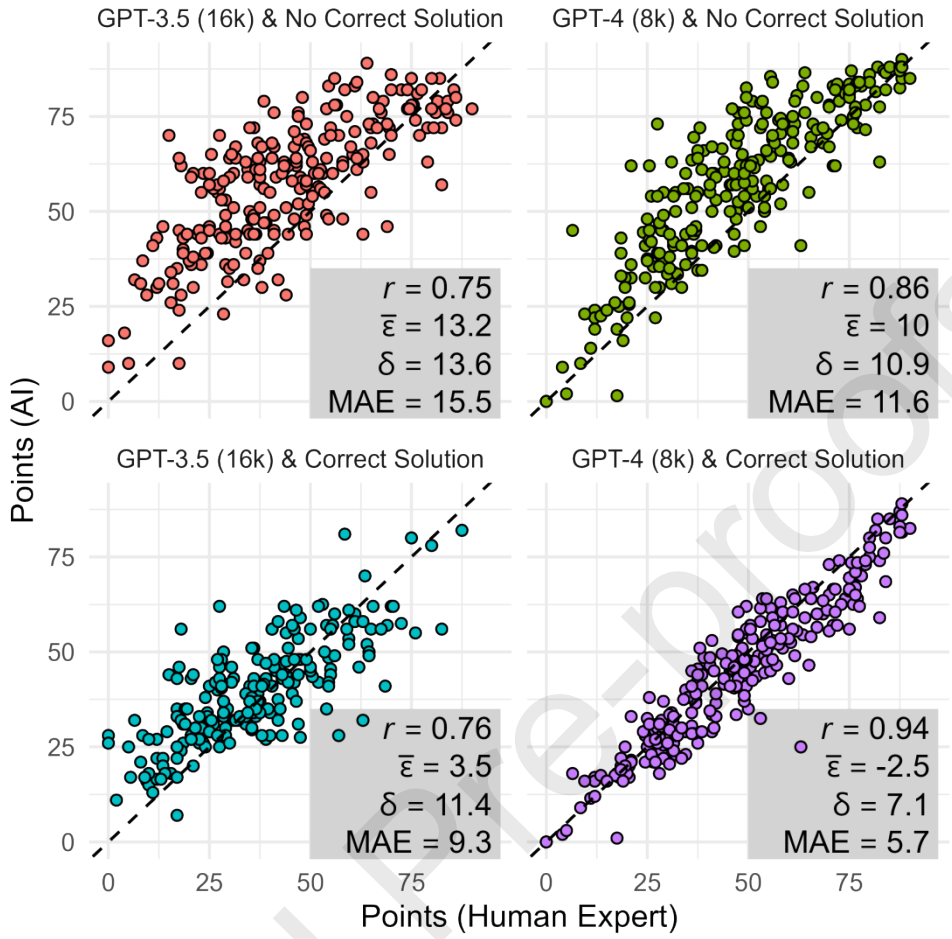
Note: In a previous sub-exercise, learners set up a demand function and received information about the variable costs.

Journal Pre-proofs

Journal Pre-proofs

Journal Pre-proofs

Figure 5: Scatterplot and Performance of Exam Evaluation by AI vs. Human Expert Depending on Model and Degree of Input

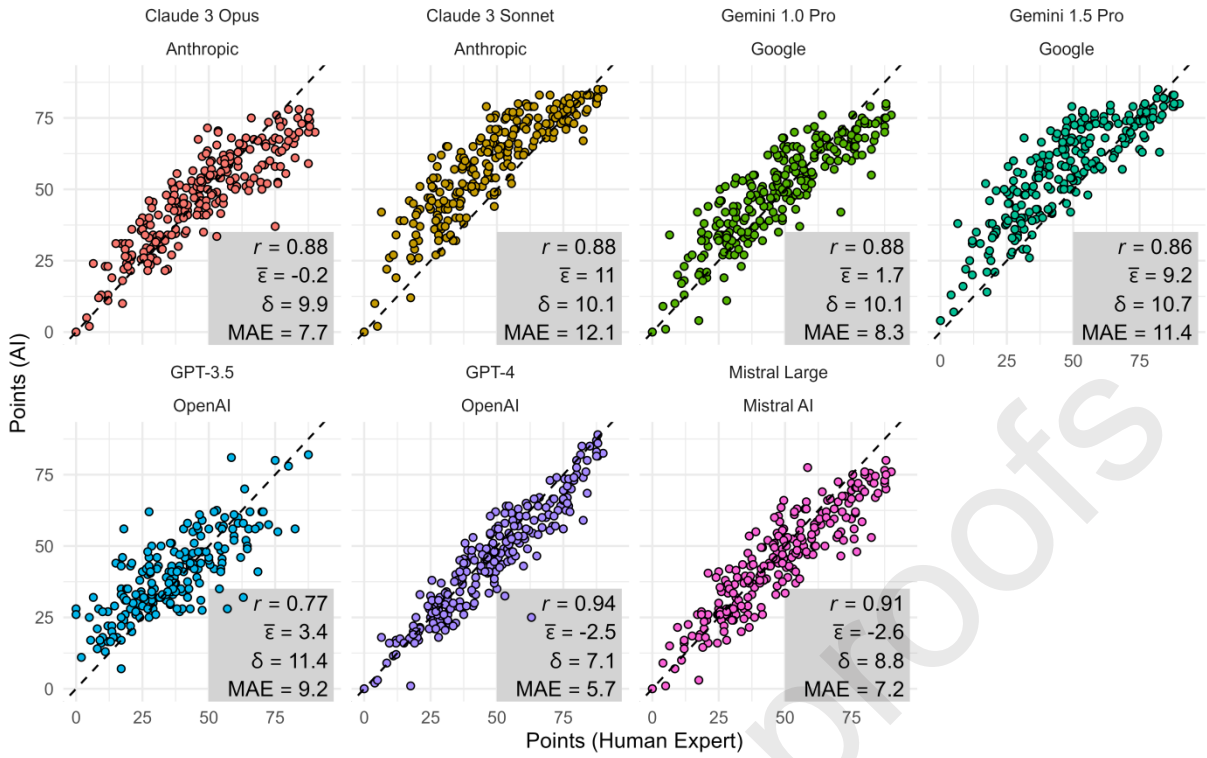


Notes: Correlation r, mean error $\bar{\varepsilon}$, standard deviation of error δ, mean absolute error MAE.

Figure 6: Scatterplot and Performance of Exam Evaluation by AI vs. Human Expert Depending on Alternative Model

Journal Pre-proofs

Journal Pre-proofs

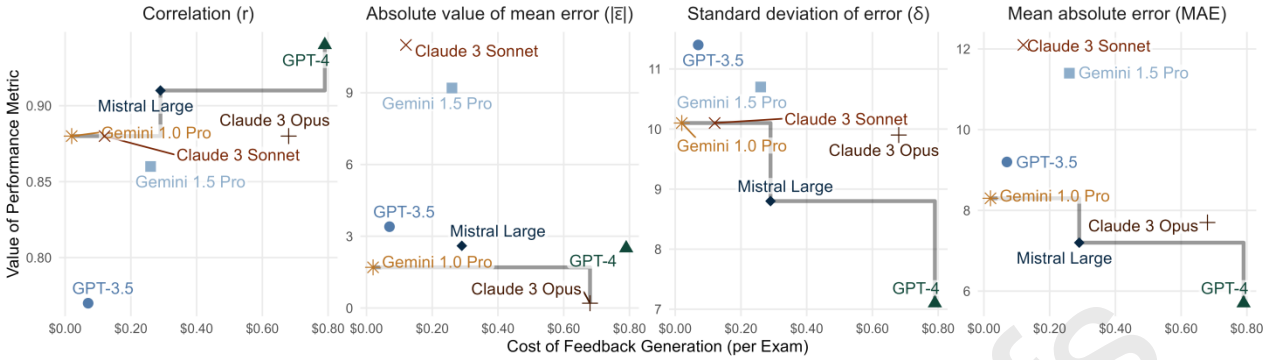


Notes: All model evaluations with the correct solution. Correlation r, mean error $\bar{\epsilon}$, standard deviation of error δ, mean absolute error MAE.

Journal Pre-proofs

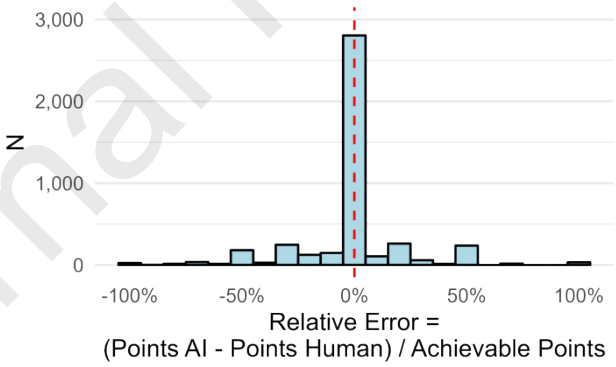
Journal Pre-proofs

Figure 7: Scatterplot of Generative AI Models' Cost vs. Performance and Efficiency Frontier



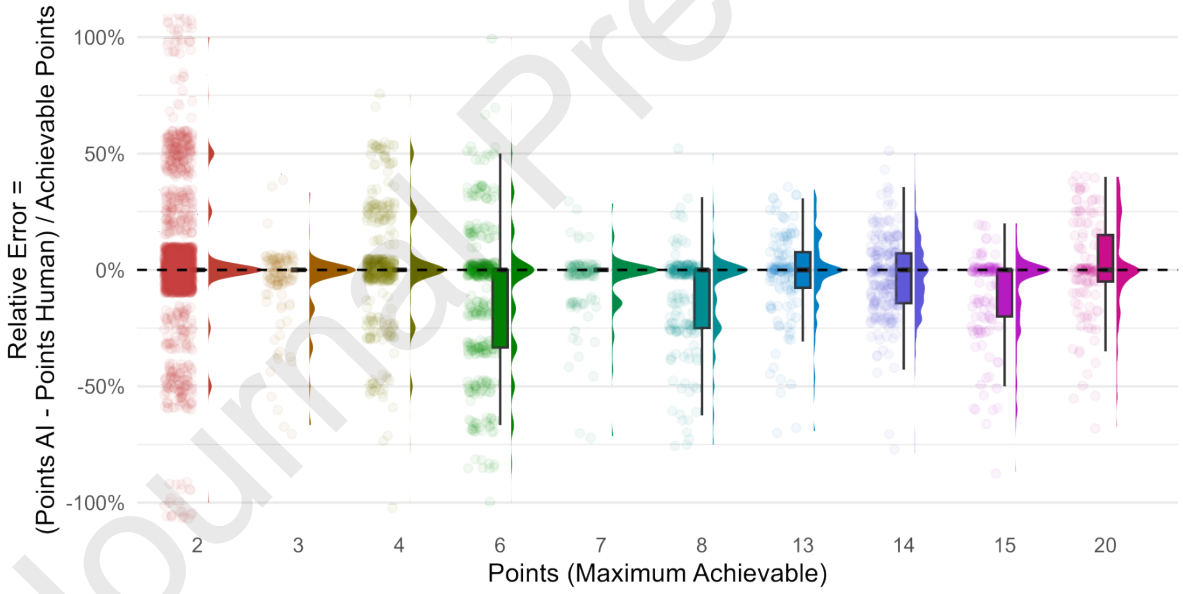
Notes: The line corresponds to the efficiency frontier, i.e., all models deviating from the line are dominated by at least one model on the line. Jointly evaluating all four performance metrics and relating them to costs, GPT-4, Mistral Large, Claude 3 Opus, and Gemini 1.0 Pro are efficient, i.e., no alternative model has either i) better performance at the same or lower cost or ii) the same or better performance at a lower cost. For the panel showing the correlation, all models below the efficiency frontier are dominated by another model, i.e., an alternative model exhibits either i) or ii). For all three remaining panels, all models above the frontier are dominated by another model. Additionally, models lying on the efficiency frontier, but not on its corners, are inefficient. We looked at the absolute value of the mean error because this metric's optimal value is zero.

Figure 8: Histogram of the Relative Error in the Best-Performing Setting (GPT-4 with Solutions) on Exercise Level



Journal Pre-proofs

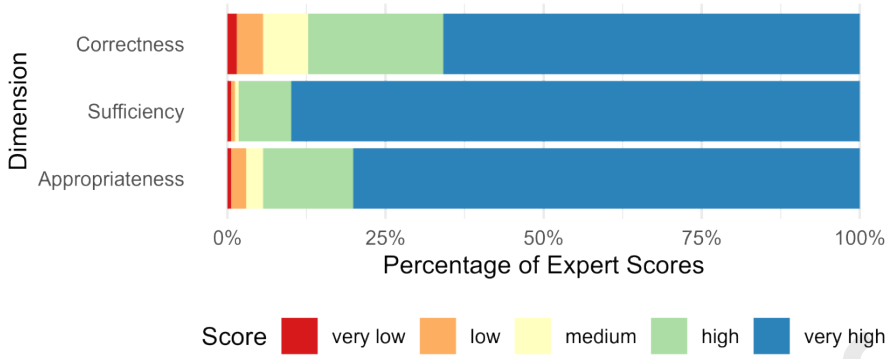
Figure 9: Distribution of Relative Error of Best-Performing Setting (GPT-4 with Solutions) Depending on Maximum Achievable Points



Journal Pre-proofs

Journal Pre-proofs

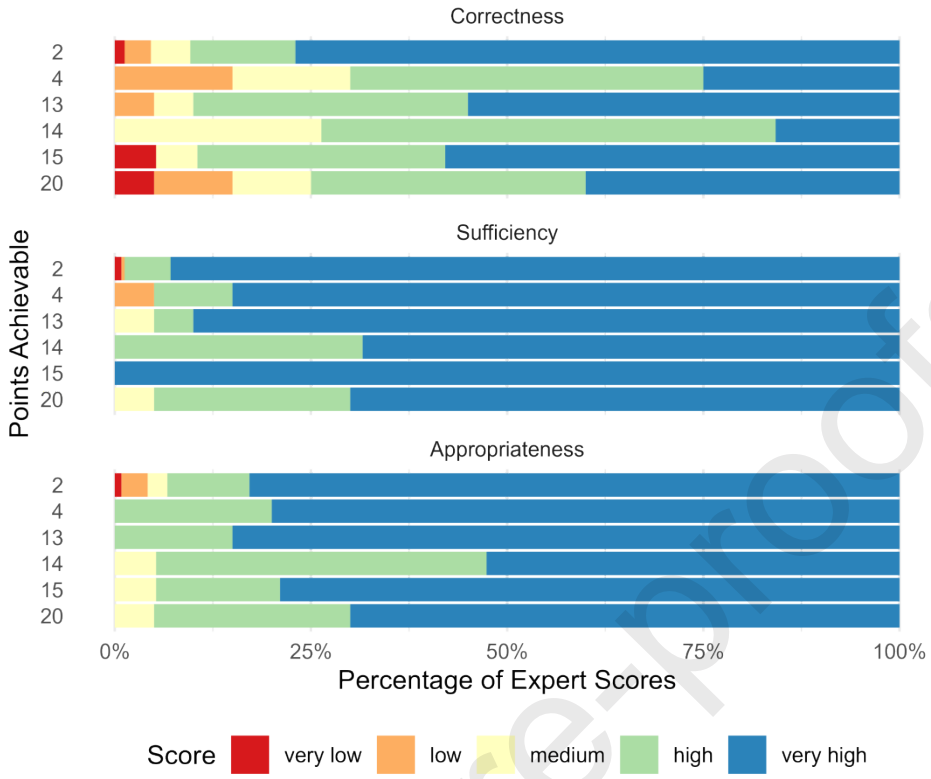
Figure 10: Expert Evaluation of Qualitative Generative AI Feedback



Note: N = 340 solutions to sub-exercises.

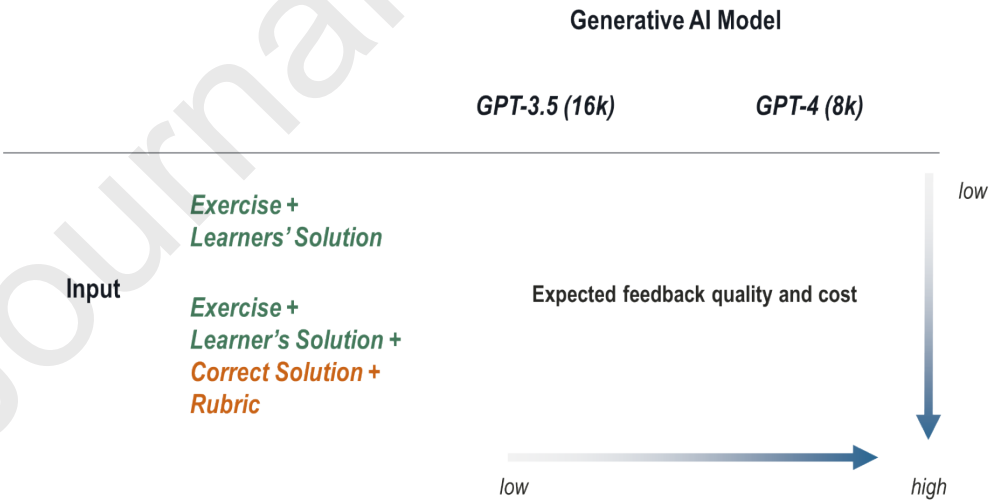
Journal Pre-proofs

Figure 11: Expert Evaluation of Qualitative Generative AI Feedback by Achievable Points of Sub-Exercise



Note: N = 340 solutions to sub-exercises.

Table 1: Design Choices for the Web App and Expected Feedback Quality and Cost



Journal Pre-proofs

Journal Pre-proofs

Table 2: Overview and Descriptive Statistics of the Two Exams

Metric	Exam A	Exam B
Exam Characteristics		
Number of Exercises	2	2
Number of Sub-Exercises	14 + 3 = 17	13 + 6 = 19
Exam Mode	Computer-based	Computer-based
Subject	Marketing Analytics	Marketing Analytics
Exam Language	German	German
Programming Language	R	R
Allotted Time to Complete the Exam	90 minutes	90 minutes
Maximum Achievable Points	90	90
Learner Submissions		
Number of Learner Submissions	134	109
Number of Solutions to Sub-Exercises	2,278	2,071
Human Evaluation (Achieved Points)		
Mean (Share of Achievable Points)	41.7 (46.3%)	53.9 (59.9%)
Standard Deviation	17.3	22.8
Minimum	0	8.5
Maximum	87.5	90

Journal Pre-proofs

Journal Pre-proofs

Journal Pre-proofs

Journal Pre-proofs

Table 3: Feedback Quality Dimensions Assessed by Expert Evaluators

Feedback Quality Dimension	Definition	Considerations for Evaluators
Correctness	The accuracy of the feedback in identifying the precise strengths and weaknesses of the learner's work.	Does the feedback • correctly acknowledge what the learner did well? • accurately identify areas that require improvement? • align with the rubric or expected outcomes?
Sufficiency	The comprehensiveness of the feedback in explaining the basis for the evaluation, especially the scores or grades.	• provide enough detail so that the learner understands why they received their score? • cover all the essential elements of the learner's response? Is the rationale behind the points awarded or deducted clear and justified?
Appropriateness	The suitability of the feedback in showing learners how to improve their performance in similar future tasks.	Is the feedback • constructive, offering specific guidance or suggestions for improvement? • presented in a way that is sensitive to the learner's level of understanding and potentially motivational?

Journal Pre-proofs

Journal Pre-proofs

Table 4: Measurements of the App's Perceived Usefulness, Perceived Ease-of-Use, and Behavioral Intent from the Technology Acceptance Model

Construct	Mean	Standard Deviation	Share of "agree" (= 4) and "strongly agree" (= 5)	Cronbach's α
Perceived Usefulness	4.13	.74	77%	.89
Perceived Ease-of-Use	4.47	.55	95%	.66
Behavioral Intent	4.35	.81	88%	(single item)

Note: Items measured on a five-point Likert scale from "strongly disagree" (= 1) to "strongly agree" (= 5), N = 43 survey participants.

Journal Pre-proofs

Table 5: Learners' Evaluation of the App's Feedback

Evaluation Metric	Mean	Standard Deviation	Share of "agree" (= 4) and "strongly agree" (= 5)
Correctness: *I perceived the app's feedback as correct.*	4.05	.84	93%
Completeness: *The feedback addressed all relevant parts of my solution.*	3.84	.84	70%
Comprehensibility: *I fully understood the app's feedback.*	3.84	1.04	58%
Helpfulness: *I found the app a valuable tool to prepare for the exam.*	4.21	.86	72%
Added Value: *I found the app a useful addition to discussing the exam in class.*	4.56	.80	77%
Recommendation: *I would recommend using the app to my fellow learners.*	4.42	.91	84%

Note: Items measured on a five-point Likert scale from "strongly disagree" (= 1) to "strongly agree" (= 5), N = 43 survey participants.

Journal Pre-proofs

Journal Pre-proofs

Journal Pre-proofs