AI Essay Evaluator: Reviewing College Apps with LLMs

Nick Walker
Department of Computer Science
Stanford University
nhwalk13@stanford.edu

Ori Spector
Department of Computer Science
Stanford University
orispec@stanford.edu

Abstract

This paper presents a comprehensive study on the development and evaluation of an AI-driven tool designed to assist in the grading and constructive feedback of Common App personal statement essays. Utilizing Large Language Models (LLMs), this innovative tool aims to democratize access to high-quality essay guidance, previously limited by financial and geographical constraints. Central to our research is the creation of a sophisticated AI assistant, grounded in a robust database comprising essays accepted to top colleges, augmented with expert assessments and salient writing qualities. This tool not only highlights the strengths and weaknesses in student essays but also provides actionable suggestions for improvement, focusing on guiding students rather than rewriting their work.

Our experimental phase can be explained in terms of 2 separate stages, both of which involved our system outputting essay feedback and grades: 1) An exploration of 78 separate model configurations, focusing on the accuracy of the assigned grade. 2) An exploration of 5 model configurations, focusing on human feedback.

In stage 1, we ran 78 experiments with each possible combination of 1 of 13 prompts, 1 of 3 temperatures, and 1 of 2 models. In this stage, we focused our assessment on grading accuracy, comparing system-determined grades with expert-given grades. At the conclusion of this stage, we chose 5 models that demonstrated qualities that we believed would result in a diverse user testing phase.

In stage 2, we tested the 5 models with 3 Stanford students. Each student input their Common App personal statement essay and received feedback on it. Participants rated the feedback from 1-5 for each model before optionally leaving written feedback. Our findings revealed that models with the greatest grading accuracy were not necessarily those which receive the best user feedback. This highlights the need for a balanced approach in AI-driven educational tools, which must be accurate in their assessments yet also align with user expectations and experiences.

Our research underscores the transformative potential of AI in the educational domain, particularly in enhancing the accessibility and quality of essay writing guidance. However, it also brings to light the complex interplay between technical accuracy and user satisfaction, a balance that is crucial for the success of such AI tools in educational contexts.

For more specifics please see our GitHub Repository.

1. Introduction

Navigating the complexities of college application essays poses a significant challenge for many students, particularly with the high costs and limited access to expert guidance. Addressing this critical issue, our project introduces an advanced, budget-friendly AI solution that leverages Large Language Models (LLMs). This AI assistant, powered by a comprehensive database of essays accepted to top colleges, salient writing qualities, and expert evaluations, is adept at grading and providing constructive feedback on Common App personal statement essays. It highlights strengths, pinpoints weaknesses, and suggests improvements, focusing on guidance rather than rewriting. Across 78 configuration experiments and 3 user feedback experiments, we demonstrate the early potential of our system as an adept essay grader and feedback provider, achieving as high as a 63% WOGA and 4/5 user rating. This technology not only makes expert essay advice more accessible but also empowers students with insights to refine their writing skills. It's an innovative leap towards democratizing educational resources, offering a practical tool for students seeking to improve their essay writing without the traditional financial burden.

2. Related Works

2.1. Automated Essay Scoring (AES)

Previous research in automatic essay grading has been performed with various machine learning techniques. For instance, Masaki Uto's *A review of deep-neural automated*

essay scoring models [6] provides an overview of his review on deep-neural automated essay scoring (DNN-AES) models. AES, which automates essay scoring to replace traditional human grading, has evolved with the advent of deep neural network models that eliminate the need for manual feature engineering. Despite the development of various DNN-AES models with distinct characteristics, there has been no detailed, comprehensive review of these models. Uto's review addresses this gap by offering an in-depth survey of DNN-AES models, discussing their core concepts and detailed architectures. He categorizes the AES task into four types and organizes the existing DNN-AES models within this framework, providing a structured and thorough examination of the field. Uto's work aided us in building our intuition of essay grading and inspired our belief in AES feasibility, contributing to our desire to improve upon existing models with LLMs.

2.2. Essay Quality Signals as Weak Supervision for Source-based Essay Scoring

The paper *Essay Quality Signals as Weak Supervision for Source-based Essay Scoring* by Zhang et al. (2021) [7] addresses the inefficiency of human essay grading and proposes Automated Essay Scoring (AES) as a solution. Recognizing the challenge of obtaining a human-graded essay corpus for training AES systems, the authors explore an alternative approach using less robust but easily available indicators like word count and topic distribution similarity. Through experiments with two essay scoring corpora, the study finds that these weak supervision methods are ineffective for training neural source-based AES models but useful for extracting Topical Components (TCs) necessary for supervised feature-based AES models. Crucially, the study demonstrates that feature-based AES models perform similarly whether TCs are constructed manually or automatically, suggesting a viable route for developing efficient AES systems without relying on extensive human-graded essay collections. This paper's findings convinced us that salient qualities would not be the only feature we used for our system's context and evaluation. We instead provide salient essay qualities to our system and the LLM while also allowing the LLM to reason outside of them.

2.3. A User-Centered Design Approach to Evaluating the Usability of Automated Essay Scoring Systems

In *A User-Centered Design Approach to Evaluating the Usability of Automated Essay Scoring Systems* [4], a paper, by Erin Hall from Virginia Tech, the focus is on evaluating the usability of Automated Essay Scoring (AES) systems through a user-centered design approach, given the advancements in AI technologies like Large Language Models (LLMs). The study addresses a research gap in AI explainability and algorithm transparency and their impact on AES platform usability. A qualitative study involving surveys, interviews, and a focus group was conducted to understand the experiences of students and graders using an AI-based essay writing and grading platform. Key findings highlight the importance of feedback clarity, its impact and actionability, user understanding, trust in AI, major user concerns, system strengths, the user interface, and improvement areas. These considerations are crucial for developing effective, transparent, and explainable essay feedback and grading tools, enhancing usability in computer science education. We strive to implement them in our own work by employing LLMs to explain strengths and weaknesses in essays, but need to improve upon feedback clarity and actionability in future work by allowing users to interact with feedback and ensuring that our system's output cites specific excerpts from a user's essay to show specifically what is good versus what is bad, enabling easy actionability.

3. Dataset and Knowledge Base

3.1. Data

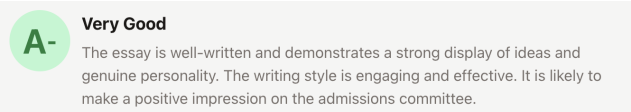
To power our AI-driven college essay evaluation tool, we manually curated a comprehensive dataset from two key sources, ensuring a diverse and insightful collection of essay examples along with expert assessments.

Our primary dataset was sourced from Essays That Worked [3], a rich repository of hundreds of successful Common Application essays. These essays are not only exemplary in content but also come with feedback, letter grades assigned by experts (based on various salient qualities that define successful essays), and explanation of the given grade. Additionally, the dataset provides valuable context by including the list of colleges where the applicants were accepted, offering a real-world benchmark of essay effectiveness.

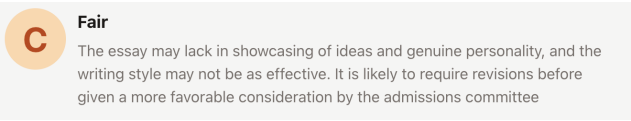
Complementing our primary source, we utilized Admit Report [5], which offered another layer of depth to our dataset. Here, we again found graded essays complete with expert comments and detailed quality assessments. Unlike Essays That Worked, Admit Report inclued examples of essays receiving a "D" grade and an "F" grade, enriching our database with diverse grades.

This dual-source approach allowed us to build a dataset that is not only varied in terms of content and style but also rich in expert analysis and grading metrics. By feeding this data into our LLM, we ensured that our AI tool is grounded in real-world examples and expert knowledge, enabling it to provide accurate, contextual, and constructive feedback to users. Sample grades and their definitions from Essays that Worked can be viewed in Figure 1 and some of the evaluated qualities can be found in Figure 2. While it is not pictured in these figures, expert-provided feedback on

specific excerpts was also available on Essays That Worked and was used in some of our prompts.



(a) Good Grade Sample



(b) Bad Grade Sample

Figure 1. Sample Data

(a) Negative Quality (b) Positive Quality

Figure 2. Sample Data

3.2. Knowledge Base

In our project, a crucial component of our data management and experimentation was the use of *Supabase*, a cloud-based Postgres database management system. Supabase played a pivotal role in various aspects of our project, primarily in data storage and retrieval, which were essential for the efficient functioning of our system.

Firstly, Supabase stored 48 expert-graded essays that we used in this system. Each record contained an essay id, essay transcript, essay prompt, and expert-given grade. A separate Supabase table was used to store 10 of these 48 records that were held-out and used in few-shot prompts. The other 38 records formed our test set and were processed by our system during experiment runs.

Secondly, we used Supabase to store the results of our 78 experiments across distinct configurations. For each experiment run, 38 records were produced for each of the essays that were evaluated. Each of these records contained the experiment id, essay id, full GPT response, extracted (via regular expression) grade, and expert-given grade (for easy comparison). These records were helpful in debugging the system. Additionally, each experiment produced a single record containing the experiment id, model, temperature, prompt, (precise) accuracy, and within one grade accuracy.

Beyond this, we used Supabase to store the 5 model configurations used in our user testing, allowing for quick retrieval by our web application during user testing based on the selected configuration. Each of these 5 configuration records contained the full prompt, temperature, and model.

Finally, Supabase was used to collect user feedback during the user testing stage. Our web application sent user responses to Supabase, where we collected ratings and written feedback pertaining to our different model configurations.

Ultimately, our use of Supabase allowed us to manage our essay knowledge base as well as our experiment results by handling scalability and data durability for us.

Outside of Supabase, we kept the full text of each of our prompts in a file called ”prompts.py” which allowed for quick retrieval during initial model configuration experiments. These prompts also contain the information that we grounded our system in. Prompts that have ”FewShot” in their name used few-shot learning and were provided with examples of grading, grading rationale, and essay feedback from the 10 held-out essay records.

4. Experimental Results

A significant portion of our project involved experimenting with various configurations of Large Language Models (LLMs) to optimize essay grading accuracy and feedback quality. In order to test out our different configurations we needed to create distinct parameters to test and build an evaluation system.

4.1. Hyperparameters

We focused on three key parameters: 1) Temperature, 2) OpenAI Chat Completion Model, and 3) Prompt Text and Structure. Specifically, for temperature we tried three different values (0.1, 0.5. 0.9) to see how our grading model worked with various levels of randomness. Additionally, we experimented with two of OpenAI’s models (gpt-3.5-turbo-16k, gpt-4-1106-preview) to see how they varied in assigning grades to essays and salient qualities (across different prompt structures). Lastly, we curated 13 different prompts to test on. This included prompts that performed essay grading and provided feedback without performing any intermediate work first, prompts that had the LLM first score salient qualities from 0 to 10 or 0 to -10 depending on if the quality was positive or negative, prompts that had the LLM first give letter grades to salient properties, and other prompts that leveraged few-shot learning by providing essay examples with their overall grades, feedback on excerpts, and general feedback. To find more information on our prompts, please refer to the *Appendix* section or the ”prompts.py” file in our project repository, which contains all the prompts we used. Since we tested every combination of our hyperparameters, we had 78 different model configurations to evaluate.

4.2. Metrics

In order to measure which configuration is most successful, we use the metrics shown by equation 1 and 2 below. For each configuration, we calculated accuracy and within one grade accuracy (WOGA). From our dataset we had essays that were scored by experts. We used those grades as

the foundation to evaluate our LLM-generated essay grades. If the LLM produced the exact grade the expert did for the same essay, then it would be considered a +1 in the accuracy numerator and a +0 otherwise (i.e. "A-" = "A-" and yields a +1 and "A-" != "A" and yields a +0). However, essay grading can be subjective, so we provided our system with leeway. Specifically, we also created a metric that captured the fraction of essays that were graded by the LLM within one grade of the expert. For example, if the expert grade was A- and our LLM produced a B+, A-, or A grade, it would be considered as correct within one grade and contribute +1 to the numerator and +0 otherwise. We refer to this as "within one grade accuracy," or WOGA.

$$\text{Accuracy} = \frac{\text{Number of Exact Matches}}{\text{Total Number Essays Graded}} \quad (1)$$

$$\text{WOGA} = \frac{\text{Number of Matches Within One Grade}}{\text{Total Number Essays Graded}} \quad (2)$$

4.3. Experiments & Methods

As mentioned before, we split our experimentation into two stages: 1) An exploration of 78 separate model configurations, focusing on the accuracy and WOGA of the assigned grade. 2) An exploration of 5 model configurations, focusing on human feedback.

Before discussing results, and so it is easier to understand why we chose the 5 models we did for user testing, we'd first like to explain some of our prompting in greater depth. There were 5 base prompts:

- 1) Baseline: This was a baseline prompt that simply instructed the model to grade the essay.
- 2) PosNegSalientProps1OverallGrade: This was a prompt that provided 20 salient qualities in college essays (both positive and negative qualities) and asked the LLM to give the essay an overall grade. These properties were derived from providing a list of 66 expert-determined salient qualities in college essays as well as our graded essay dataset to ChatGPT-4 and asking it to determine the 20 most salient qualities, or those qualities that most influence the grade of the essay.
- 3)PosNegSalientPropsIntermediateGrades1OverallGrade: This was a prompt that provided the same 20 salient qualities but also asked the LLM to give intermediate letter grades to the 20 salient grades and then an overall grade informed by the intermediate grades.
- 4) PosSalientPropsIntermediateGrades1OverallGrade: This was a prompt that provided the 20 salient qualities, but reworded any negative salient qualities to be there positive inverse (i.e., "Surface-Level Takeaways" would be converted to "Deep Takeaways"). This change was in response to the LLM sometimes being confused about how to assign a letter grade to negative qualities. The LLM then was told to give a letter grade to each intermediate quality to inform an overall letter grade.
- 5)PosNegScoresSalientPropsIntermediateScores1OverallGrade: This was a prompt that again used the 20 salient qualities (worded positively and negatively) but assigned positive qualities a score from 0 to 10 and negative qualities a score from 0 to -10 rather than letter grades. It then used these scores to inform the final grade it gave.

The latter 4 of these base prompts then had few-shot and "short" few-shot versions of them made. The few-shot versions included expert-assessed examples of "A+," "A," "A-," "B+," "B," "B-," "C+," "C," "D," and "F" essays (1 of each), their grade, the rationale behind the grade, feedback on specific excerpts in the essays, and feedback on what could have gone better and what went well. On the other hand, the "short" few-shot versions provided examples of "A+," "B," "C," "D," and "F" essays (1 of each), their grade, the rationale behind the grade, and feedback on what could have gone better and what went well.

Varying these prompts, temperatures, and models led to a rich set of experiments.

4.4. Results

4.4.1 Stage 1 (Configuration) Results

After running the 78 configuration experiments, we chose 5 models of interest for user testing. We computed accuracy and WOGA across the 78 and selected these 5 configurations for the following reasons:

Configuration 1 (Greatest WOGA): This configuration, achieved the highest WOGA (0.631579), demonstrating a robust capability to approximate the grades assigned by human experts. It used a temperature of 0.9, GPT-4, and the PosNegScoresSalientPropsIntermediateScores1OverallGrade prompt. While not always exact in its grading, it showed a commendable general understanding of essay quality, which is pivotal for a tool aimed at providing broad yet accurate feedback. It was selected for user testing since it achieved the greatest WOGA accuracy.

Configuration 2 (Greatest Accuracy): Exhibiting the highest accuracy (0.263158) and a high WOGA (0.605263), this model demonstrated the ability to emulate experts precisely and approximately. It stood out as the most reliable configuration for consistent and extremely accurate grading, underscoring the efficacy of our AI tool in aligning closely with expert evaluations. It used a temperature of

0.9, GPT-3.5-Turbo, and the PosSalientPropsIntermediateGrades1OverallGrade prompt. It was selected for user testing since it was the model that attained the greatest (precise) accuracy.

Configuration 3 (Greatest Accuracy and WOGA Among Baseline Prompt Configurations): This configuration resulted in an accuracy of 0.184211 and WOGA of 0.526316. This established a baseline for us to improve over, which we did. We selected it for user testing to see if GPT-3.5-Turbo could satisfy users out of the box. It used a temperature of 0.5, GPT-3.5-Turbo, and the baseline prompt.

Configuration 4 (Greatest WOGA Among "Short" Few-Shot Prompt Configurations) This configuration resulted in an accuracy of 0.131579 and a WOGA of 0.605263. It used a temperature of 0.5, GPT-4, and the ShortFewShotPosNegScoresSalientPropsIntermediateScores1OverallGrade prompt. It was selected for user testing since it was the best performing "short" few-shot prompt configuration (by WOGA), and we wanted to test how a prompt using "short" few-shot prompting would perform with users.

5 (Greatest WOGA Among Few-Shot Prompt Configurations): This configuration achieved an accuracy of 0.131579 and a WOGA of 0.552632. It used a temperature of 0.9, GPT-4, and the FewShotPosNegScoresSalientPropsIntermediateScores1OverallGrade prompt. It was selected for user testing because it was the best performing few-shot prompt configuration (by WOGA) and we wanted to test how a prompt using few-shot prompting would perform with users.

Interestingly enough, it did not seem that GPT-4 greatly outperformed GPT-3.5-Turbo in this context. However, GPT-4 did excel in scenarios in which short or long few-shot prompting was used, perhaps indicating that it is more well-suited for few-shot learning than GPT-3.5-Turbo is within the essay grading space.

Table 1. Model Performance Comparison

Model #	Temperature	AI Model	Accuracy	WOGA
1: Best WOGA	0.9	gpt-4-1106-preview	0.184211	**0.631579**
2: Best Accuracy	0.9	gpt-3.5-turbo-16k	**0.263158**	0.605263
3: Best Baseline	0.5	gpt-3.5-turbo-16k	0.184211	0.526316
4: Best Few-Shot (Short)	0.5	gpt-4-1106-preview	0.131579	0.605263
5: Best Few-Shot (Long)	0.9	gpt-4-1106-preview	0.131579	0.552632

4.4.2 User Test Results

In addition to the technical evaluation, user feedback played a critical role in assessing the practical effectiveness of our configurations. For each of the 5 chosen configurations, 3 users gave us feedback in the form of a rating between 0-5. Models 2 and 5 received the highest average user ratings of 4.00, indicating a strong alignment with user expectations and perceived utility of the feedback. These findings and the fact that having the greatest accuracy and/or WOGA does not necessarily correspond with being the most user-satisfying model, highlight the importance of user-centric design in AI systems, where user satisfaction is at least as crucial as technical accuracy.

Reflection on Results

The combined analysis of technical performance and user feedback underscores a key insight: the necessity for a balanced approach in AI-driven educational tools. While technical accuracy is fundamental, the subjective experience of the users - their satisfaction with and understanding of the feedback - is equally vital. Our results suggest that future improvements should focus on integrating the technical proficiency of configurations like Configuration 2 with the user-approved approach of Configuration 5. This integrated strategy aims to develop a tool that is not only technically sound but also resonates well with its end-users, enhancing both the efficacy and the usability of the AI assistant in real-world applications.

Another future step could be simply taking Configuration 1 or 2 and changing the prompting to ensure that specific excerpts from the graded essay are cited. Configuration 1 was our best WOGA configuration, showing that it has a strong understanding of what makes an essay high quality or not. However, at least 1 user noted that they didn't like it because it wasn't specific enough in its critiques. If we improve on this dimension and call it out in our prompting more strongly, we may be able to achieve greater performance in user testing while also being able to deploy our most (WOGA) accurate model configuration.

Table 2. User Ratings by Model

Model	Avg. Rating
Model 1: Desc	3.67
Model 2: Desc	**4**
Model 3: Desc	3.67
Model 4: Desc	**3.33**
Model 5: Desc	**4**

5. Discussion & Future Works

5.1. Limitations & Areas for Improvement

5.1.1. Increased User Testing

To refine accuracy and usability, we emphasize the importance of increased user testing. This involves a systematic approach to gather feedback from a broad user base, thus ensuring that the system's responses are not only accurate but also user-friendly. This iterative process is crucial for identifying and addressing specific issues that may not be apparent during initial development phases.

5.1.2. Increasing Concrete Examples in Responses

Based on user feedback, incorporating more concrete examples in responses is pivotal. This approach not only aids in clarifying complex concepts but also enhances the overall user experience. It serves as an educational tool, helping users to better understand the context and application of the information provided.

5.1.3. Varying Context Lengths

Adjusting the lengths of context provided to the Language Learning Models (LLMs) could potentially enhance their understanding and the quality of feedback. This involves experimenting with short, medium, and long context inputs to determine the optimal length that balances comprehensive understanding without overwhelming the model with irrelevant information.

5.1.4. Expanded Data and Metrics

To evaluate model configurations more effectively, we propose the incorporation of a broader range of essays and diverse metrics. This expansion will allow for a more comprehensive analysis of the models' performance across various topics and writing styles, leading to more nuanced and tailored feedback.

5.2. Risks and Considerations

5.2.1. Balancing Constructive Criticism and Positive Reinforcement

It is essential to balance constructive criticism with positive reinforcement to maintain user engagement. While critical feedback is necessary for improvement, excessive criticism can be demotivating. Positive reinforcement encourages continued learning and exploration.

5.2.2. Avoiding Promotion of Formulaic Writing

There is a risk of promoting formulaic writing, which can stifle creativity. The system should encourage diverse writing styles and ideas, ensuring that users feel free to express themselves creatively while still receiving constructive feedback.

5.2.3. Communicating Model Limitations

Finally, communicating the limitations and inherent subjectivity of the models to the users is crucial. Users should be aware that the feedback provided is based on algorithmic interpretations and may not always align with human judgment. This transparency is key to setting realistic expectations and fostering trust in the system.

5.3. Future Work

5.3.1. Real-Student Testing

In future developments, a significant focus should be on testing the tool with actual high school seniors who are in the process of applying to college. This real-world testing environment is crucial for several reasons. First, it allows for the evaluation of tool performance in the specific context it is intended to be used. Second, it provides direct feedback on user satisfaction from the target demographic. Third, and most importantly, it offers an opportunity to assess the impact of the tool on the quality of the essays produced by these students. By engaging with students who are actively using the tool for a real and consequential task, we can gain invaluable insights into the tool's efficacy and areas for improvement.

5.3.2. Interactive Messaging

Another key area for future work is the implementation of interactive messaging features. This would transform the tool from a passive feedback mechanism to an active, conversational interface. By allowing students to interact with the tool - asking questions, seeking clarifications, and receiving responses in real time - the tool can emulate a tutor-like experience. This approach not only makes the feedback process more engaging but also allows for more personalized and context-specific assistance. The interactivity can help address specific student concerns and queries, making the feedback process more tailored and effective.

5.3.3. Enhanced Prompting Techniques

The exploration of enhanced prompting techniques represents a crucial area for development. The goal here is to refine the strategies used to prompt AI feedback, with a focus on improving the quality and accuracy of the feedback, particularly in the nuanced context of essay evaluation. This involves experimenting with different types of prompts that can guide the AI in providing more precise, detailed, and contextually relevant feedback. By fine-tuning these prompting techniques, the tool can be better equipped to handle the complexities and subtleties inherent in personal essay writing, thereby offering more valuable and actionable insights to the students. This approach also opens up possibilities for the AI to better understand and respond to the unique styles and voices of individual writers, further enhancing the personalization of feedback.

6. Conclusion

This study explored the development of an AI-driven tool utilizing Large Language Models (LLMs) to assist in evaluating and providing feedback on Common App personal statement essays. Our goal was to make quality essay guidance more accessible, leveraging a database of successful college essays and expert evaluations. The AI tool was designed to supplement human judgment by identifying key elements in student essays and suggesting areas for improvement.

Our research involved analyzing various LLM configurations to assess their grading accuracy and user satisfaction. Through experiments with five models, we found that different configurations have specific strengths: Model 1 excelled in Within One Grade Accuracy, and Model 2 effectively balanced technical accuracy with user satisfaction. These findings highlight the challenges in creating AI tools that are both accurate and user-friendly.

The study also emphasized the need for balance in AI feedback, avoiding formulaic writing and transparently communicating the limitations of AI models to users. Future work includes extensive testing with real students, implementing interactive features for a more engaging experience, and refining prompting techniques for better feedback quality.

In conclusion, this paper demonstrates the potential of AI in educational assistance, particularly in enhancing essay writing skills. The key takeaway is the importance of balancing technical precision with user-centric design. Future efforts will focus on integrating these aspects to develop an even more effective and engaging educational tool.

References

[1] Openai api, 2023. 7

[2] Supabase api, 2023. 7

[3] Ryan Chiang. Essays that worked, 2023. 2

[4] Erin Elizabeth Hall. A user-centered design approach to evaluating the usability of automated essay scoring systems, 2023. 2

[5] Alex McNeil. Admit report, 2023. 2

[6] Masaki Uto. A review of deep-neural automated essay scoring models, 2021. 2

[7] Haoran Zhang and Diane Litman. Essay quality signals as weak supervision for source-based essay scoring, 2021. 2

7. Apendix

Thanks to OpenAI [1] and Supabase [2] for making our project possible.

Sample Prompt

1) Please evaluate each of the following qualities within the following essay and assign a grade from A+ to F.

2) After doing this, give the essay an overall grade based on your evaluation and weighting of each quality. All qualities don't need to be weighted equally. Use your best judgment and still grade the essay holistically to give a suitable grade.

3) Explain why you gave this overall grade in terms of the qualities you graded.

4) Write at the end of the response "Grade: grade", where grade is the overall grade you gave the essay.

5) After all of this, write up a feedback report for the college student whose essay you just reviewed. Explain the key points to them and cite specific passages in their essay to demonstrate your point, if possible. Make sure you tell them the overall grade they received, things they did well, and things they could improve on. Again, cite specific passages in their essay to demonstrate your point, if possible. Separate this section off at the end under the message "OVERALL FEEDBACK"

***List of Qualites (Postive): ***

Being Genuine
Having Unique Ideas
Reflects Deeply And Meaningfully
Shows Gratitude And Maturity
Connects to Activities & Achievements
Shows Change In Perspective
Shows Humility
Strong Narrative
Be Authentic
Showcase Your Values and Identity
Non-Surface-Level Ideas
Non-Vague-And-Impersonal Writing
Non-Unnecessary Storytelling
Clear Main Idea
Shows, Not Tells
Not Overly Critical/Negative
Does Not Lack Reflection
Not Overly Self-Confident
Does Not Lack Unique And Deep Ideas
Strong Intro/Conclusion

You must follow this example format for each response:
==========
List of Qualities:
Being Genuine: A+

Having Unique Ideas: B-

Reflects Deeply And Meaningfully: A

...

Overall Grade: A+
Explanation: Being genuine was the most important trait for this essay and made it stand out, helping counteract the B- grade on having unique ideas...

OVERALL FEEDBACK
You receive an A grade because...
Things you did well:
Things you could improve:
==========

*** Essay Context:

Sample Prompt
Grade this college essay and put your answer as Grade: grade so I can parse it.
After all of this, write up a feedback report for the college student whose essay you just reviewed. Explain the key points to them and cite specific passages in their essay to demonstrate your point, if possible. Make sure you tell them the overall grade they received, things they did well, and things they could improve on. Again, cite specific passages in their essay to demonstrate your point, if possible. Separate this section off at the end under the message "OVERALL FEEDBACK"

Example Response:
==========
You received: B+

OVERALL FEEDBACK
You receive a B+ grade because...
Things you did well:
Things you could improve:
==========

Here's the essay:

Sample Prompt
1) Please evaluate each of the following positive qualities within the essay and assign a score from 0 to 10 for each. Assign a negative score from 0 to -10 for each negative quality, going lower for negative qualities that show up more. If the negative quality does not show up at all, score it a 0 (no deduction).

2) After evaluating each quality, give the essay an overall grade based on your evaluation and weighting of each quality. Not all qualities need to be weighted equally. Use your best judgment and grade the essay holistically to give a suitable overall grade.

3) Explain why you gave this overall grade in terms of the individual qualities you graded.

4) At the end of your response, write "Grade: grade", where grade is the overall grade you gave the essay.

5) After all of this, write up a feedback report for the college student whose essay you just reviewed. Explain the key points to them and cite specific passages in their essay to demonstrate your point, if possible. Make sure you tell them the overall grade they received, things they did well, and things they could improve on. Again, cite specific passages in their essay to demonstrate your point, if possible. Separate this section off at the end under the message "OVERALL FEEDBACK"

List of Qualities (Positive/Negative):
Being Genuine: Positive Having Unique Ideas: Positive Reflects Deeply And Meaningfully: Positive Shows Gratitude And Maturity: Positive Connects to Activities & Achievements: Positive Shows Change In Perspective: Positive Shows Humility: Positive Strong Narrative: Positive Be Authentic: Positive Showcase Your Values and Identity: Positive Surface-Level Ideas: Negative Vague And Impersonal Writing: Negative Unnecessary Storytelling: Negative Unclear Main Idea: Negative Tells, Not Shows: Negative Overly Critical/Negative: Negative Lacks Reflection: Negative Overly Self-Confident: Negative Lacks Unique And Deep Ideas: Negative Weak Intro/Conclusion: Negative

Graded Essay Examples:
1) **Grade: A+**

Essay:
People love to ask why. Why do you wear a turban? Why do you have long hair? Why are you playing a guitar with only 3 strings and watching TV at 3 A.M.—where did you get that cat? Why won't you go back to your country, you terrorist? My answer is...uncomfortable. Many truths of the world are uncomfortable. A couple of examples are that an equal number of pets are euthanized as are adopted each year and that cats roam the streets at night because they are actually looking for owners with better food. One of those statements is a horrible truth and the other is a thought I had in the shower. Either way, the point still stands. Uncomfortable truths are just that, uncomfortable. The answer to 'Why won't you go back to your country, you terrorist?' is the most uncomfortable answer I can give, barring the current status of aboriginal street cats.

Sikhs like myself have borne the brunt of the backlash through our forced subjection to hate crimes, bullying, and job discrimination. In [Date], a misguided gunman took the lives of six Sikhs who were praying peacefully in their house of worship in Oak Creek, Wisconsin. Their families, through their tears, asked the nation, like I continue to ask myself, "Why?"

The uncomfortable truth is that as a society, we have not found a solution to the growing trend of extremism and hate crimes—we failed at the whole "freedom of religion" clause in the Bill of Rights. The media tells us that these crimes are carried out by individuals that are ignorant and motivated by hate. I

would personally call them losers, but that would solve none of underlying system problems that have grown from anti-immigration rhetoric. When my cousin joined the US Army, he was told that he'd have to cut his beard and hair. Every time I tell that part of the story I can't help but guffaw at how ridiculous it sounds. My then eleven-year-old angst came to a climactic fruition hearing those words—it was a call to action.

I helped to gather signatures for a petition to Robert Gates, then Secretary of Defense, pleading with him to allow Sikhs to serve without having to cut our hair. We garnered over 15,000 signatories, receiving generous media attention. We called and convinced our local congressional offices to support this issue. I created a Facebook page to help spread awareness, and helped to organize fundraisers to help fight this ban on our articles of faith. Our message is simple. Through service, we can push back against both hate and intolerance. But, if the largest employer in the U.S. does not allow us to serve with our articles of faith, then we will continue to be victimized as outsiders, contrary to the founding principles of our nation.

I'm proud to say my cousin deployed to Afghanistan as the first Sikh to be granted a religious waiver in nearly a generation. He saved countless lives as a doctor on the front lines of war and was awarded the Bronze Star Medal for his efforts. But, (there's always a but) Sikhs today still face a presumptive ban. Despite being in perfect—for the army's sake—physical condition, I cannot join the US Army because of my hair.

So now the uncomfortable story that was the uncomfortable answer to an uncomfortable question comes to an uncomfortable ending. And, like all great uncomfortable answers, I never really answered the main question. I don't have the answers to why people do the hateful things they do. But by wearing my turban proudly every morning, by answering questions when they come up, by being willing to talk about everything that is wrong, I become a personification of what is right. My solution to the systemic problem starts with me.

Feedback on Specific Parts:
Part: "The answer to 'Why won't you go back to your country, you terrorist?' is the most uncomfortable answer I can give, barring the current status of aboriginal street cats."

Feedback: Great. Combines meaningful idea with lighthearted remark. When dealing with serious topics, it is easy for your essay to come off as a "downer." To combat this, consider adding elements of personality that show lightheartedness, while still being serious when dealing with your main ideas.

Overall Feedback:
The essay is well-written, thoughtful, and showcases an impressive display of ideas and genuine personality. The writing style is creative and engaging. It stands out from other essays and is likely to make a strong impression on the admissions committee.

Central Theme: Having a unifying idea is key to successful personal statements. Find your deepest idea or realization and focus your essay around that. Connects to Accomplishments: Find a way to showcase your achievements while connecting to broader, more universal ideas. Strong Conclusion and Hook: Connecting your ending to your beginning is a powerful way to bring your essay full circle. A great conclusion expands on your ideas introduced earlier, while leaving some room for more to be said.

2) **Grade: A Essay:** Watching my coach demonstrate the drill, it seemed so simple. But when I tried to do the Carioca drill (it sounds like "karaoke", but doesn't involve wailing into a microphone - it's more like shuffling sideways while doing the Irish jig), everything fell apart.

...
Prompt Pattern Repeats
...

Overall Feedback:
There's no easy way to say it, but this essay just doesn't meet the mark. That's why it gets an F. It reads like a free write rather than an essay because it is stream-of-consciousness and doesn't really make a clear point. I learn that the writer loves chemistry, but the overall message is not clear.

What does this essay do well?

Ideas: All hope is not lost! Once we dig into what each sentence of the essay is saying, there are some good ideas that the writer can turn into a more cohesive topic.

What could you improve on?

Organization: I hesitate to make any extreme claims about college essays, but I feel pretty confident

in saying that the vast majority of college essays should always be more than one paragraph. You need paragraphs to break up your thoughts into digestible chunks. Each paragraph should contain a single point you're trying to convey to the reader. This writer should break all these ideas up into several paragraphs.

Theme: We see that the topic of the essay is chemistry, that chemistry is interesting because it's the foundation of everything, and that chemistry can help people. But we don't really get any deeper meaning from the writer. They haven't made an attempt to be vulnerable or to show us something significant about themself.

Length: The essay is almost a hundred words over the word count. The writer needs to pare things down as they organize and clarify their ideas.

You must follow this example format for each response:
========== List of Qualities: Being Genuine: 8 Having Unique Ideas: 9 Reflects Deeply And Meaningfully: 4 . . . Lacks Unique and Deep Ideas: -5 (Include all qualities)
Overall Grade: A Explanation: Being genuine was the most important trait for this essay (which attained a score of 10 on it) and made it stand out, helping counteract the -5 grade on having unique ideas...
OVERALL FEEDBACK
You receive an A grade because. . . Things you did well: Things you could improve:
==========
*** Essay Context:

User Flow & UI

Uploading Essay + Model Selection

Sample qualities graded numerically

List of Qualities:
Being Genuine: 8
Having Unique Ideas: 7
Reflects Deeply And Meaningfully: 7
Shows Gratitude And Maturity: 8
Connects to Activities & Achievements: 9
Shows Change In Perspective: 6
Shows Humility: 6
Strong Narrative: 8
Be Authentic: 7
Showcase Your Values and Identity: 8

Sample Feedback from Model 2

OVERALL FEEDBACK
You receive an A- grade because your essay successfully conveys a strong and genuine interest in AI, particularly regarding its technical complexities and ethical implications. Your narrative effectively connects your past experiences to your future aspirations and demonstrates how Stanford's Computer Science Coterm program is a natural fit for your goals.

Things you did well:
- You adeptly showcase your academic and research background and how it aligns with your passions.
- The dual focus on the technical aspects of AI and the ethical considerations provides depth to your essay, illustrating your thoughtful and holistic approach to the subject.
- You show a strong linkage to past achievements and future goals, which enforces the idea that you are a candidate who plans strategically and with clear intention.

Things you could improve:
- Dive deeper into specific experiences that shaped your ethical considerations in AI to offer more personal insight.
- Use more "showing" rather than "telling" in your writing to make your reflections on AI and its impact on society more tangible and vivid.
- Be cautious of sounding overly assured when referencing your capabilities and future plans—maintaining a balance between confidence and humility is key to a well-rounded application essay.

Rate the Essay Feedback

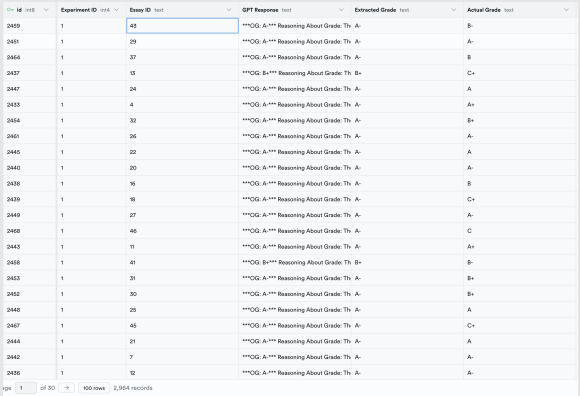
☆ ☆ ☆ ☆ ☆

Write Feedback

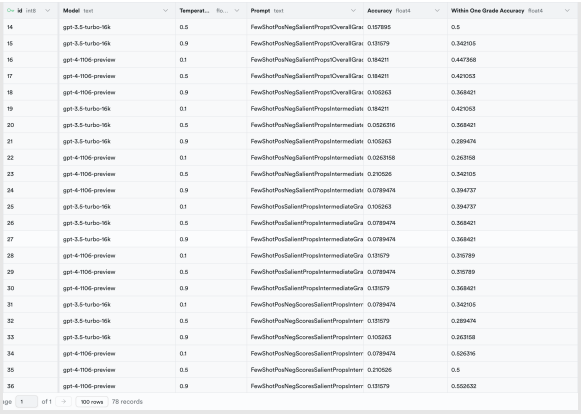
Write your feedback here.

Submit

Supabase Experiment-Content Table



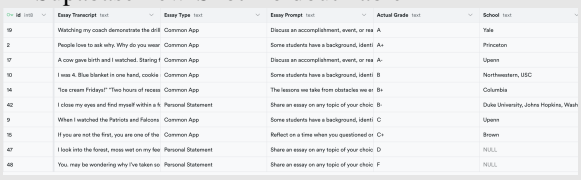
Supabase Experiment-Data Table



id int8	Model text	Temperat... flo...	Prompt text	Accuracy float4	Within One Grade Accuracy float4
14	gpt-3.5-turbo-16k	0.5	FewShotPosNegSalientPropsIOverallGra	0.157895	0.5
15	gpt-3.5-turbo-16k	0.9	FewShotPosNegSalientPropsIOverallGra	0.131579	0.342105
16	gpt-4-1106-preview	0.1	FewShotPosNegSalientPropsIOverallGra	0.184211	0.447368
17	gpt-4-1106-preview	0.5	FewShotPosNegSalientPropsIOverallGra	0.184211	0.421053
18	gpt-4-1106-preview	0.9	FewShotPosNegSalientPropsIOverallGra	0.105263	0.368421
19	gpt-3.5-turbo-16k	0.1	FewShotPosNegSalientPropsIntermediate	0.184211	0.421053
20	gpt-3.5-turbo-16k	0.5	FewShotPosNegSalientPropsIntermediate	0.0526316	0.368421
21	gpt-3.5-turbo-16k	0.9	FewShotPosNegSalientPropsIntermediate	0.105263	0.289474
22	gpt-4-1106-preview	0.1	FewShotPosNegSalientPropsIntermediate	0.0263158	0.263158
23	gpt-4-1106-preview	0.5	FewShotPosNegSalientPropsIntermediate	0.210526	0.342105
24	gpt-4-1106-preview	0.9	FewShotPosNegSalientPropsIntermediate	0.0789474	0.394737
25	gpt-3.5-turbo-16k	0.1	FewShotPosSalientPropsIntermediateGra	0.105263	0.394737
26	gpt-3.5-turbo-16k	0.5	FewShotPosSalientPropsIntermediateGra	0.0789474	0.368421
27	gpt-3.5-turbo-16k	0.9	FewShotPosSalientPropsIntermediateGra	0.0789474	0.368421
28	gpt-4-1106-preview	0.1	FewShotPosSalientPropsIntermediateGra	0.131579	0.315789
29	gpt-4-1106-preview	0.5	FewShotPosSalientPropsIntermediateGra	0.0789474	0.315789
30	gpt-4-1106-preview	0.9	FewShotPosSalientPropsIntermediateGra	0.131579	0.368421
31	gpt-3.5-turbo-16k	0.1	FewShotPosNegScoresSalientPropsInterr	0.0789474	0.342105
32	gpt-3.5-turbo-16k	0.5	FewShotPosNegScoresSalientPropsInterr	0.131579	0.289474
33	gpt-3.5-turbo-16k	0.9	FewShotPosNegScoresSalientPropsInterr	0.105263	0.263158
34	gpt-4-1106-preview	0.1	FewShotPosNegScoresSalientPropsInterr	0.0789474	0.526316
35	gpt-4-1106-preview	0.5	FewShotPosNegScoresSalientPropsInterr	0.210526	0.5
36	gpt-4-1106-preview	0.9	FewShotPosNegScoresSalientPropsInterr	0.131579	0.552632

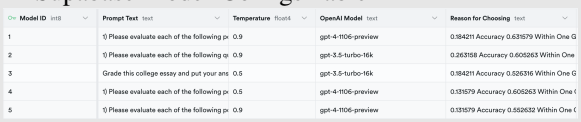
age 1 of 1 → 100 rows 78 records

Supabase Few-Shot Holdout Table



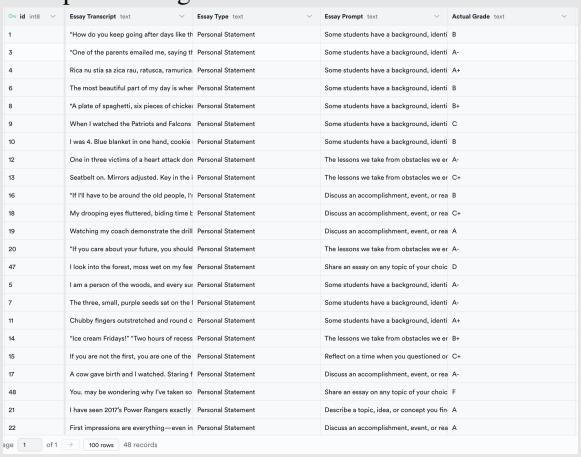
id int8	Essay Transcript text	Essay Type text	Essay Prompt text	Actual Grade text	School text
19	Watching my coach demonstrate the drill	Common App	Discuss an accomplishment, event, or rea	A	Yale
2	People love to ask why. Why do you wear	Common App	Some students have a background, identi	A+	Princeton
17	A cow gave birth and I watched. Staring f	Common App	Discuss an accomplishment, event, or rea	A-	Upenn
10	I was 4. Blue blanket in one hand, cookie	Common App	Some students have a background, identi	B	Northwestern, USC
14	"Ice cream Fridays!" "Two hours of recess	Common App	The lessons we take from obstacles we er	B+	Columbia
42	I close my eyes and find myself within a fi	Personal Statement	Share an essay on any topic of your choic	B-	Duke University, Johns Hopkins, Wash
9	When I watched the Patriots and Falcons	Common App	Some students have a background, identi	C	Upenn
15	If you are not the first, you are one of the	Common App	Reflect on a time when you questioned or	C+	Brown
47	I look into the forest, moss wet on my fee	Personal Statement	Share an essay on any topic of your choic	D	NULL
48	You. may be wondering why I've taken so	Personal Statement	Share an essay on any topic of your choic	F	NULL

Supabase Model-Configs Table



Model ID int8	Prompt Text text	Temperature float4	OpenAI Model text	Reason for Choosing text
1	1) Please evaluate each of the following pa	0.9	gpt-4-1106-preview	0.184211 Accuracy 0.631579 Within One G
2	1) Please evaluate each of the following qu	0.9	gpt-3.5-turbo-16k	0.263158 Accuracy 0.605263 Within One
3	Grade this college essay and put your ans	0.5	gpt-3.5-turbo-16k	0.184211 Accuracy 0.526316 Within One G
4	1) Please evaluate each of the following pa	0.5	gpt-4-1106-preview	0.131579 Accuracy 0.605263 Within One (
5	1) Please evaluate each of the following pa	0.9	gpt-4-1106-preview	0.131579 Accuracy 0.552632 Within One (

Supabase OrigData Table



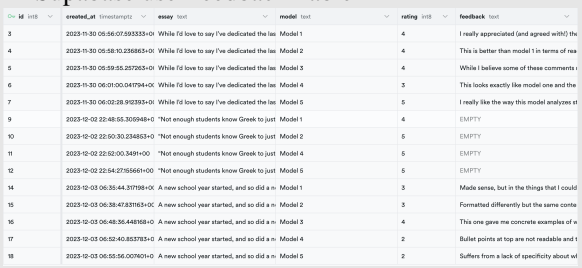
id int8	Essay Transcript text	Essay Type text	Essay Prompt text	Actual Grade text
1	"How do you keep going after days like th	Personal Statement	Some students have a background, identi	B
3	"One of the parents emailed me, saying tl	Personal Statement	Some students have a background, identi	A-
4	Rica nu stia sa zica rau, ratusca, ramurica	Personal Statement	Some students have a background, identi	A+
6	The most beautiful part of my day is whe	Personal Statement	Some students have a background, identi	B
8	"A plate of spaghetti, six pieces of chicke	Personal Statement	Some students have a background, identi	B+
9	When I watched the Patriots and Falcons	Personal Statement	Some students have a background, identi	C
10	I was 4. Blue blanket in one hand, cookie	Personal Statement	Some students have a background, identi	B
12	One in three victims of a heart attack don	Personal Statement	The lessons we take from obstacles we er	A-
13	Seatbelt on. Mirrors adjusted. Key in the i	Personal Statement	The lessons we take from obstacles we er	C+
16	"If I'll have to be around the old people, I'l	Personal Statement	Discuss an accomplishment, event, or rea	B
18	My drooping eyes fluttered, biding time b	Personal Statement	Discuss an accomplishment, event, or rea	C+
19	Watching my coach demonstrate the drill	Personal Statement	Discuss an accomplishment, event, or rea	A
20	"If you care about your future, you should	Personal Statement	The lessons we take from obstacles we er	A-
47	I look into the forest, moss wet on my fee	Personal Statement	Share an essay on any topic of your choic	D
5	I am a person of the woods, and every sur	Personal Statement	Some students have a background, identi	A-
7	The three, small, purple seeds sat on the l	Personal Statement	Some students have a background, identi	A-
11	Chubby fingers outstretched and round c	Personal Statement	Some students have a background, identi	A+
14	"Ice cream Fridays!" "Two hours of recess	Personal Statement	The lessons we take from obstacles we er	B+
15	If you are not the first, you are one of the	Personal Statement	Reflect on a time when you questioned or	C+
17	A cow gave birth and I watched. Staring f	Personal Statement	Discuss an accomplishment, event, or rea	A-
48	You. may be wondering why I've taken so	Personal Statement	Share an essay on any topic of your choic	F
21	I have seen 2017's Power Rangers exactly	Personal Statement	Describe a topic, idea, or concept you fin	A
22	First impressions are everything—even in	Personal Statement	Discuss an accomplishment, event, or rea	A

age 1 of 1 → 100 rows 48 records

Supabase Results Table

model text	Avg. Rating float8
Model 1	3.67
Model 2	4
Model 3	3.67
Model 4	3.33
Model 5	4

Supabase userFeedback Table



id int8	created_at timestamptz	essay text	model text	rating int8	feedback text
3	2023-11-30 05:56:07.593333+0(	While I'd love to say I've dedicated the las	Model 1	4	I really appreciated (and agreed with!) the
4	2023-11-30 05:58:10.236863+0(	While I'd love to say I've dedicated the las	Model 2	4	This is better than model 1 in terms of rea
5	2023-11-30 05:59:55.257263+0(	While I'd love to say I've dedicated the las	Model 3	4	While I believe some of these comments
6	2023-11-30 06:01:00.041794+0(	While I'd love to say I've dedicated the las	Model 4	3	This looks exactly like model one and the
7	2023-11-30 06:02:28.912393+0(	While I'd love to say I've dedicated the las	Model 5	5	I really like the way this model analyzes st
9	2023-12-02 22:48:55.305948+0	"Not enough students know Greek to just	Model 1	4	EMPTY
10	2023-12-02 22:50:30.234853+0	"Not enough students know Greek to just	Model 2	5	EMPTY
11	2023-12-02 22:52:00.3491+00	"Not enough students know Greek to just	Model 4	5	EMPTY
12	2023-12-02 22:54:27.155661+00	"Not enough students know Greek to just	Model 5	5	EMPTY
14	2023-12-03 06:35:44.317198+0(	A new school year started, and so did a ne	Model 1	3	Made sense, but in the things that I could
15	2023-12-03 06:38:47.831163+0(	A new school year started, and so did a ne	Model 2	3	Formatted differently but the same conte
16	2023-12-03 06:48:36.448168+0	A new school year started, and so did a ne	Model 3	4	This one gave me concrete examples of w
17	2023-12-03 06:52:40.853783+0	A new school year started, and so did a ne	Model 4	2	Bullet points at top are not readable and t
18	2023-12-03 06:55:56.007601+0	A new school year started, and so did a ne	Model 5	2	Suffers from a lack of specificity about wl