Modeling Elite College Admissions with Domain-Adapted Large Language Models

Aiden Jia
Montclair Kimberley Academy
New Jersey, USA
aiden.jia712@gmail.com

Haolin Jiang
Electrical and Computer Engineering
Rutgers University
New Jersey, USA
hj462@scarletmail.rutgers.edu

Hang Liu
Electrical and Computer Engineering
Rutgers University
New Jersey, USA
hl1097@scarletmail.rutgers.edu

***Abstract*—Large language models (LLMs) have demonstrated remarkable capabilities in reasoning and language understanding, yet their application to complex human systems remains underexplored. In this study, we leverage recent advancements in open-source LLMs to model and analyze the U.S. selective college admissions process. We first curate and release a cleaned, anonymized dataset of 2,825 historical college applications and admissions outcomes from 2022-2025. Using this dataset, we fine-tune the Mistral-Small-24B-Instruct-2501 model through domain-adaptive pretraining, producing a specialized LLM capable of identifying latent patterns in application success. Our model achieves 71.7% accuracy in predicting acceptance to Top 20 (T20) colleges, outperforming baseline general-purpose LLMs. Leveraging chain-of-thought prompting and statistical validation via Pointwise Mutual Information, we extract actionable insights, highlighting demonstrated leadership, initiative, and community-mindedness as key factors in successful applications. These results provide both empirical understanding of elite college admissions and practical guidance for prospective applicants, demonstrating the potential of domain-adapted LLMs to analyze complex, context-rich human decision-making processes. Implications include improved transparency in holistic admissions and scalable tools for student guidance.**

***Index Terms*—Large language models, domain-adaptive pretraining, QLoRA, chain-of-thought, college admissions, educational data mining**

I. Introduction

Large language models (LLMs) have become increasingly powerful tools for structured reasoning and domain-specific analysis [1]–[5]. Initially trailing behind their proprietary counterparts, open-source LLMs such as Mistral [6], LLaMA 3 [3], and DeepSeek [4], [5] now deliver state-of-the-art performance across a wide range of reasoning tasks, including chain-of-thought inference, document summarization, and multi-step problem-solving. This growing capability has unlocked new opportunities to apply LLMs in domains that require deep contextual reasoning over rich, unstructured data.

One such domain is U.S. college admissions, a complex and historically opaque process shaped by institutional priorities, holistic evaluation, and context-specific heuristics [7]. Top-ranked colleges, particularly those in the "T20" group, assess applicants not only by academic metrics (e.g., GPA, SAT scores) but also by qualitative elements such as essays, recommendation letters, extracurriculars, and perceived character [8]. Over the past two decades, admission rates at these institutions have declined sharply. For instance, Harvard's acceptance rate fell from 10.9% in 2000 to 3.9% in 2024; Duke dropped from 25.4% to 5.1% [9]. As selectivity intensifies, students increasingly invest in crafting and optimizing every component of their applications. This is often done with the help of private college counselors, who may charge thousands of dollars [10]. These services typically rely on experience and historical patterns to advise students, mimicking a data-driven evaluative process grounded in past outcomes [11].

While LLMs have shown remarkable success in various evaluation and reasoning tasks, their application to the college admissions domain remains limited. Existing research on admissions has predominantly focused on system-level concerns such as racial and socioeconomic disparities [12] or legacy preferences [13], while paying limited attention to the structure and content of applications themselves. A few studies have shifted this focus, including Ben-Michael et al. [14], which analyzes the role of recommendation letters, and *The Early Admissions Game* [15], which explores the benefits of applying early. However, these efforts are conducted in an earlier admissions context and examine only narrow components of the application, which may limit their applicability to today's more competitive landscape.

Such a gap raises a natural question: if both counselors and institutions rely on historical data, could an LLM trained on past applications and outcomes approximate this evaluative process while also generating actionable insights into what drives admissions decisions? To explore this, we develop a domain-adapted LLM fine-tuned on real application records from the 2022-2025 cycles. Our model is designed to reason over application content and outcomes, uncover latent patterns in successful profiles, and offer interpretable, data-driven guidance for future applicants navigating an increasingly competitive admissions landscape.

In this work, we make three primary contributions that advance understanding of the college admissions process and demonstrate the potential of domain-adapted LLMs:

1) Dataset: We curate and release a cleaned, anonymized corpus of recent college applications and admissions outcomes, standardized for use in NLP training.

2) Domain-Adapted LLM: We fine-tune an open-source large language model on this corpus, producing a domain-

specialized model that captures latent patterns in the college admissions process.

3) Empirical Insights Into Admissions: Using both model outputs and statistical analysis of the raw dataset, we identify key factors that influence elite admissions outcomes and translate these findings into data-backed, actionable insights for applicants.

II. Dataset

Our first core contribution is the creation of a text corpus of 2,825 historical college applications and corresponding admissions results used for the fine-tuning of large language models (LLMs). We use a multi-step process to ensure that each entry in the corpus is valid, properly structured, and suitable for model tuning.

A. Data Collection

To build our dataset, we needed historical applications along with results; however, direct access to this data proved to be infeasible due to legal privacy constraints. Government legislation, such as the Family Educational Rights and Privacy Act (FERPA), prohibits the disclosure of personally identifiable data (PID) by schools, colleges, or counselors. As a result, we turned to self-reported sources, focusing primarily on Reddit communities related to college admissions. Among these, the subreddit r/CollegeResults offered the most consistent and accessible format for extracting structured application data.

1) Collecting Data from the Reddit API: We collected posts using the official Reddit API through the Python Reddit API Wrapper (PRAW). However, the Reddit API only allows access to the newest 1,000 posts, which, at the time of collecting, extended back to just February 1, 2025.

2) Collecting Data from Pushshift Data Dumps: To obtain posts beyond this limit, we utilized the Pushshift Reddit data dumps [16], a collection of all posts and comments from the top 40,000 subreddits. Included in these dumps was content from the r/CollegeResults subreddit dating back to the 2015 admissions cycle. We restricted the included posts to only those in the last three college admission cycles (2022-2025) to ensure the data we used was relevant to today's admission climate.

B. Data Curation

To curate (clean and standardize) the collected posts to prepare for model training, we needed to ensure that each one was relevant, included all required details (demographics, GPA, test scores, extracurriculars, etc.), and contained the student's admissions results.

1) Data Validation: As the data from r/CollegeResults is all self-reported, we utilize the "score" of each post (all upvotes minus downvotes) as well as community self-moderation to ensure the validity of the data to the best of our ability. This approach provided us with about 3,000 case posts of individual students' college applications, including their acceptance and rejection outcomes.

2) Filtering Data: To ensure that our dataset was both relevant and high-quality, we first applied a keyword filter to the titles of each post, removing entries that were clearly irrelevant, satirical, or off-topic. Next, to standardize the dataset and guarantee that each post contained all required information, we employed a two-step query to GPT-4o-mini [17]. In the first step, the model summarized the content of each post into a consistent, structured format, capturing elements such as academic performance, standardized test scores, extracurricular involvement, awards, and demographic details. In the second step, GPT-4o-mini verified the presence of all key application components, flagging posts that were incomplete or missing essential information. This approach allowed us to systematically curate a dataset that was both comprehensive and consistent, while minimizing the need for manual review.

3) Data Standardization: We also utilized the standardized post content from 4o-mini as the content to appear in our final corpus. We converted the content to plain text and removed all placeholder content (N/A, none, not provided, etc.) that was left behind to ensure that the LLM would not inadvertently highlight these during training. Lastly, we removed any emojis or Reddit artifacts from the data to ensure that all of our data was in plain text.

C. Dataset Construction

With cleaned and standardized data, we then compiled all of our individual example cases (plain text of each individual application followed by the corresponding admissions results) into a tokenizable, unannotated corpus suitable for training use. Each example case was delimited with "===". Lastly, we scanned the dataset by hand to ensure there were no errors in formatting or content. This provided us with a training-ready corpus of 2,825 case examples, each being a real college application and corresponding admissions results.

All data was collected in compliance with the Reddit API Terms of Service. No personally identifying information, such as usernames, real names, or other unique identifiers, was retained in the final dataset.[1]

III. Domain-Adaptive Fine-Tuning

Our second core contribution is the development of an LLM specialized for the college admissions domain. In this section (Part A, contribution #2), we describe how we adapted the Mistral-Small-24B-Instruct-2501 model [18] to this domain using continued pre-training on our curated text corpus. In the following section (Part B, contribution #2), we evaluate this domain-adapted model against existing LLMs to benchmark its performance and demonstrate the effectiveness of our fine-tuning approach.

A. Fine-Tuning Approach

As outlined above, we employed continued pre-training (also known as domain-adaptive pretraining or DAPT), a technique shown to enhance model performance on domain-specific tasks by exposing the model to unlabeled in-domain

[1]The anonymized dataset used in this study, including curated application features and admissions outcomes will be released after paper acceptance.

text [19]. Similar approaches have proven effective in specialized contexts, such as biomedical NLP [20], demonstrating the capabilities of adapting general-purpose models to specific domains. In our case, this approach, combined with our unannotated corpus, enables the model to internalize latent patterns across the data, such as recurring traits among successful applicants to highly selective institutions, through the inherent co-occurrence of specific application details and admissions results.

B. Model Selection

We select Mistral AI's Mistral-Small-24B-Instruct-2501 as the base model for our pre-training due to its balance of scale, reasoning ability, and accessibility [18]. At 24 billion parameters, this model is large enough to capture complex semantic relationships and latent patterns across long, list-like application data, while still being computationally feasible for continued pre-training using techniques such as QLoRA. Additionally, the model's instruction-tuned variant offers improved compatibility for structured prompting and interpretability, which is particularly important for our latter steps of prompting for evaluation and insight extraction.

C. Training Method

To carry out continued pre-training on Mistral-Small-24B-Instruct-2501, we utilized Quantized Low-Rank Adapters (QLoRA) to enable fine-tuning of the model within the constraints of a single NVIDIA A100 GPU with 40 GB of VRAM. QLoRA builds upon Low-Rank Adaptation (LoRA), which reduces the number of trainable parameters by introducing low-rank updates to a frozen pre-trained model [21]. LoRA has been shown to retain model performance while significantly lowering computational costs.

QLoRA further improves efficiency by incorporating 4-bit quantization of the base model weights during training, enabling fine-tuning of LLMs on limited hardware without sacrificing performance [22]. Additionally, LoRA has been shown to maintain model performance even in low-resource settings, where the availability of relevant training data is limited [21]. With our data containing fewer than 3,000 examples, these factors made performing continued pre-training through QLoRA the optimal choice.

To further optimize our model tuning with limited hardware, we fine-tuned using a per-device batch size of 1 with gradient accumulation over 8 steps, giving us an effective batch size of 8 sequences per GPU. We used a learning rate of 2×10^{-4} and trained for 3 epochs with a maximum sequence length of 2,048 tokens. Mixed precision training (fp16) was enabled to reduce memory usage and improve throughput.

IV. Model Evaluation and Results

To assess the effectiveness of our domain-adapted model (Part B, contribution #2), we design an evaluation task based on a real-world prediction scenario: determining whether a given applicant would be accepted to a Top 20 (T20) U.S. college based on their application profile. We then compare our model's results to those of state-of-the-art general-purpose systems such as GPT-4.1 [23] and Gemini 2.5 Pro [24]. We also compare our model to more accessible proprietary LLMs such as GPT-4.1-mini [25] and Gemini 2.5 Flash [26], which are widely used due to their lower latency and reduced cost, but typically trade off some performance.

A. Evaluation Method

For evaluation, we curated a test set of 60 new applicant profiles from r/CollegeResults posts published after our training corpus had been created. This ensures that our model has not inadvertently seen any of the test cases during training. The test set consisted of nearly the same information as the training data, containing all key application details (academics, standardized testing, extracurriculars, demographics, awards, etc.), except that the admissions results were omitted.

Our primary prediction task was binary classification: given an application profile, the model predicts whether the applicant was accepted to a T20 university. We compared our model's predictions against those of several baselines: large general-purpose LLMs (GPT-4.1 and Gemini 2.5 Pro), more lightweight proprietary models (GPT-4.1-mini and Gemini 2.5 Flash), and the unmodified Mistral-Small-24B-Instruct-2501 base model. All models were prompted with the same structured input format, with temperature set to 0, and instructed to output a binary "Yes" or "No" label, enabling fair and quantitative performance comparisons. The prompt structure is available in Appendix A.

Model accuracy was calculated as the proportion of correct predictions on the test set, with additional metrics such as precision, recall, and F1 score used to explore differences in model behavior.

B. Evaluation Results

Fig. 1 shows the accuracy of each model by the percentage of correct predictions. The fine-tuned Mistral-Small-24B-Instruct-2501 had the highest accuracy, correctly predicting 71.7% (43/60) of the test cases, while the base Mistral model and Gemini 2.5 Flash performed the worst, correctly predicting only 58.3% (35/60) of the test cases. On the other hand, Gemini 2.5 Pro performed reasonably well (68.3%; 41/60), coming close to the fine-tuned Mistral model.

Table I shows the precision, recall, and F1 score of each model to provide more insight into the evaluation. Precision measures how often predicted acceptances were correct, recall measures how many true acceptances were identified, and F1 balances the two. These metrics help evaluate not just overall accuracy, but how well each model captures the nuanced patterns of successful applicants in the college admissions domain.

The fine-tuned Mistral model achieved the most balanced performance, demonstrated by its F1 score of 0.712. In contrast, the base Mistral model achieved higher recall (0.852) but lower precision (0.523), suggesting it tends to overpredict acceptances. Interestingly, while GPT-4.1 and 4.1-mini achieved the same accuracy (65%), 4.1-mini was quite balanced in its

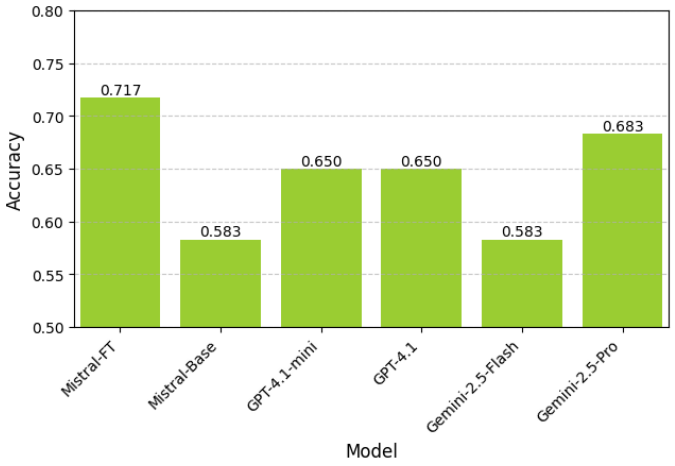

Fig. 1. Accuracy of each model on the evaluation dataset (percent of correct predictions).

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Precision	Recall	F1 Score
Mistral-FT	0.656	0.778	0.712
Mistral-Base	0.523	0.852	0.648
GPT-4.1-mini	0.607	0.630	0.618
GPT-4.1	0.875	0.259	0.400
Gemini-2.5-Flash	0.528	0.704	0.603
Gemini-2.5-Pro	0.654	0.630	0.642

predictions, whereas GPT-4.1 exhibited high precision (0.875) but very low recall (0.259), indicating a conservative pattern that misses many true acceptances.

By benchmarking against both the largest available general-purpose models and smaller, faster proprietary models, we show that our continued pre-training approach yields a model whose performance reflects both high predictive accuracy and strong alignment with domain-specific patterns. These results validate that our model is not only well-trained but also adapted to the nuanced decision-making context of elite college admissions.

V. EMPIRICAL INSIGHTS INTO ADMISSIONS

Our third core contribution is the extraction of meaningful insights from the college admissions process using our domain-adapted LLM and supporting statistical analysis. In this section (Part A, contribution #3), we utilize our fine-tuned model to identify what application features are most strongly correlated with admissions outcomes at elite U.S. colleges. We then cross-validate these findings against the underlying dataset and compare them with prior studies to ensure that our observations are grounded in empirical evidence. The next section (Part B, contribution #3) will build on these insights to provide actionable, data-driven guidance for prospective applicants.

A. Insight Extraction Method

To extract insights from the model after training, we configured the model with a temperature of 0.1 to ensure output stability and reliability, while still allowing for the generation of alternative reasoning paths. This balance would be restricted at a temperature setting of 0. To assess the model's confidence in its responses, each prompt was asked five times. Prompts for which the model produced consistent answers (four or more agreeing responses out of five) were considered to reflect high confidence and were recorded as valid insights.

For prompting, we employed chain-of-thought (CoT) prompting, a technique shown to significantly improve the reasoning capabilities of LLMs on complex tasks. CoT encourages models to generate intermediate reasoning steps rather than jumping directly to a final output, which not only enhances accuracy but also provides transparency into the model's decision-making process [27]. For our use case, analyzing how specific features of college applications relate to admissions outcomes, CoT is even more practical, as prior work has demonstrated that CoT prompting is particularly beneficial in domains where multi-step inference, latent pattern recognition, or qualitative judgment is required [28]. The prompt starter used is given in Appendix B.

B. Core Admission Factors

As our end goal is to provide actionable insights for students, we begin by prompting the LLM with a series of questions to identify the most important factors in college admissions where applicants also have the most control. The analysis highlighted three main areas: extracurricular involvement, personal qualities, and relevant skills, along with important details in each.

1) Common Extracurricular Activities: The LLM consistently identified several extracurricular domains that showcased strong correlations to acceptances at T20 colleges (Table II).

Successful applicants to top-tier colleges often demonstrated not just involvement, but depth and leadership in their extracurricular activities. These activities involved experiences that reflected intellectual curiosity, leadership, and initiative,

TABLE II
EXAMPLES OF COMMON ACTIVITIES BY CATEGORY

Category	Common Examples
Academic Competitions & Research	Science/Math Olympiads, computer science competitions, mentored research, published papers, or conference presentations
Leadership Roles	President of clubs/organizations, sports team captain
Community Service	Long-term volunteering, leadership in service organizations, founding service projects
Athletics	Varsity sports participation, team captain, awards
Performing & Visual Arts	Arts club leadership, recognition in competitions, solo performances
Internships & Work Experience	Internships aligned with academic interests, jobs showing responsibility and initiative
Cultural Activities	Language clubs, cultural event leadership, teaching/tutoring in another language
Entrepreneurship & Innovation	Launching a business or nonprofit, participation in pitch or startup competitions
Summer Programs	Selective summer schools (e.g., Ivy League pre-college), subject-specific camps
Awards & Honors	National and international level recognition, academic or artistic honors

particularly in areas that connect to specific academic or personal interests.

2) Core Character Traits: The LLM showed that successful applicants consistently demonstrated several key personal qualities, indirectly communicated through their listed activities and accomplishments (Table III).

Leadership continued as a strong trend among top applicants. The LLM also noted that these applicants demonstrated the ability to "take action" and make meaningful contributions to their areas of choice. The ability to adapt and overcome, especially in a collaborative manner, was a trait frequently demonstrated through activities as well. Finally, following academic interests, either through outside activities or coursework, was a recurring characteristic.

3) Hard and Soft Skills Correlated with Admission: The LLM identified several concrete skills that were disproportionately demonstrated in accepted applications (Table IV).

Expectedly, the most important "hard skills" were based on demonstrated academic strength and the ability to perform high-level research. Importantly, leadership was once again present as one of the top "soft skills" among successful T20 applicants, along with collaboration, resilience, and demonstrated community-mindedness.

Interestingly, qualities such as leadership, initiative, and community-mindedness appeared frequently in the LLM's responses. Importantly, these factors are much more accessible for students and easier to showcase compared to things like awards or research. To better understand their relative impact and how students might strategically highlight them, we probed the model further to explore nuances in how these traits influence admissions outcomes.

TABLE III
PERSONAL TRAITS AND EVIDENCE IN APPLICATIONS

Trait	Description / Evidence in Applications
Leadership	Held key roles in organizations, managed teams or initiatives
Initiative	Founded new clubs, launched projects, independently pursued opportunities
Commitment	Long-term involvement, visible contributions
Collaboration	Active in team-based efforts such as sports, performing arts, or group community service
Resilience	Overcame challenges in rigorous activities or personal adversity
Community Mindedness	Demonstrated social awareness and impact through volunteering or advocacy
Creativity & Innovation	Engaged in the arts, STEM creation, or unique project-based learning
Academic Curiosity	Conducted research, joined academic teams, challenging coursework beyond requirements

TABLE IV
SKILLS AND INDICATORS IN COLLEGE APPLICATIONS

Skill Type	Skill	Indicators / Evidence
Hard Skills	Academic Rigor	High GPA in AP/IB/Honors courses
	Standardized Testing	SAT/ACT scores above 90th percentile, 4s and 5s on AP exams
	Extracurricular Achievements	Awards, competition placements, and leadership in academic or artistic activities
	Research Experience	Mentored research, published work, STEM or social science investigations
Soft Skills	Leadership	Led clubs, teams, or initiatives
	Community Service	Ongoing volunteer work with measurable or personal impact
	Collaboration	Contributions in team-based environments
	Resilience & Adaptability	Evidence of growth through challenge, ability to navigate change
	Passion & Authenticity	Deep engagement in specific interests, consistent narrative across application

C. Exploring Nuances

Building on the accessible qualities highlighted by the LLM (leadership, initiative, community-mindedness), we next examined the finer-grained impact of these traits on admissions outcomes. By probing the model with more detailed prompts, we explored how specific personal qualities appear in successful applications and how they compare to traditional academic metrics.

1) Leadership: Leadership emerged as the most frequently cited trait among successful applicants, with most holding two to four roles ranging from team captaincies to community initiatives. While less influential than GPA, which appeared to function as a baseline qualifier, the LLM stated that leadership carried weight comparable to strong essays or recommendations. Notably, non-recruited athletes without leadership roles were less likely to gain admission compared to non-athletes, whereas team captains experienced higher acceptance rates, underscoring the role of leadership in strengthening competitiveness.

2) Initiative: Demonstrated initiative, which we define as the ability to start and carry out meaningful projects, was another key theme in successful applications to T20 universities. The LLM reported that accepted applicants to T20 universities typically started between two and three initiatives. These are projects, organizations, or events that the student starts and leads themselves, and include things such as community events, entrepreneurial ventures, technology projects, interest-based clubs, and more. Initiative was found to have a slightly more substantial impact than that of a strong personal statement or supplemental essay, and comparable in weight to that of demonstrated leadership.

3) Community-Mindedness: A demonstrated sense of community-mindedness was also frequently observed among successful applicants to top-tier universities. According to the LLM, successful applicants had between two and three different volunteer positions. However, among the three most commonly exhibited personal qualities (leadership, initiative, and community-mindedness), it appeared to have the weakest influence on admissions outcomes. According to the model, community-mindedness carried a lower impact on admissions decisions than academic performance (e.g., GPA) and a lower impact than strong writing components such as personal statements or supplemental essays.

4) Significance of Ideal Personal Qualities: Importantly, the combination of leadership, initiative, and community-

mindedness (subsequently referred to as "ideal personal qualities") was found to carry a notably significant weight in the admissions process when demonstrated in college applications. While not as influential as a high GPA, the LLM reported that the presence of these personal qualities had an impact stronger than that of good writing or recommendations, and nearly equivalent to that of a strong standardized test score (i.e., SAT/ACT).

D. Statistical Validation

While the LLM provides qualitative insight into the traits and behaviors correlated with successful T20 college applications, we assess the robustness of these findings through quantitative analysis. To do so, we cross-validated the LLM's observations against the training dataset by transforming each application into a categorized binary format, encoding the presence or absence of key features such as leadership roles, volunteering, initiated projects, GPA, and standardized test scores.

Using this structured dataset, we apply Pointwise Mutual Information (PMI) to quantify the strength and direction of association between individual application features x and acceptance outcomes. Formally, PMI between a feature x and outcome y is defined as:

$$\text{PMI}(x, y) = \log \frac{P(x, y)}{P(x) \cdot P(y)}$$

Where $P(x, y)$ is the joint probability of both the feature x and the outcome y occurring together, and $P(x)$ and $P(y)$ are the marginal probabilities. A positive PMI indicates a feature is more frequently associated with acceptance than expected by chance, while a negative PMI indicates under-occurrence. Laplace smoothing ($+0.01$) was applied to reduce instability for rarer features.

Our PMI evaluation confirmed that features the LLM identified, such as leadership roles, volunteer positions, and started initiatives (indicators of "ideal personal qualities"), were positively correlated with acceptance to at least one T20 college (Fig. 2). Conversely, lacking these features was negatively correlated with acceptance at a T20. Unsurprisingly, strong GPA and standardized test scores were positively correlated with T20 acceptance as well (Fig. 3).

The PMI analysis also revealed more nuanced patterns (Fig. 3). General participation in sports (without demonstrated leadership) correlated negatively with T20 acceptance. Interestingly, holding only a single leadership position showed a negative correlation with T20 admission, emphasizing the importance of demonstrating consistent leadership ability across several disciplines.

It is important to note that PMI measures co-occurrence strength rather than causal influence, and its values are sensitive to feature frequency. High-frequency features yield large PMI scores due to consistent co-occurrence with admissions, whereas rarer but potentially more impactful features may yield smaller PMI values simply due to limited sample size. While our tests did not show any extreme outliers, this nuance should still be taken into account when analyzing results. Fig. 4 displays the frequency distribution of all notable features.

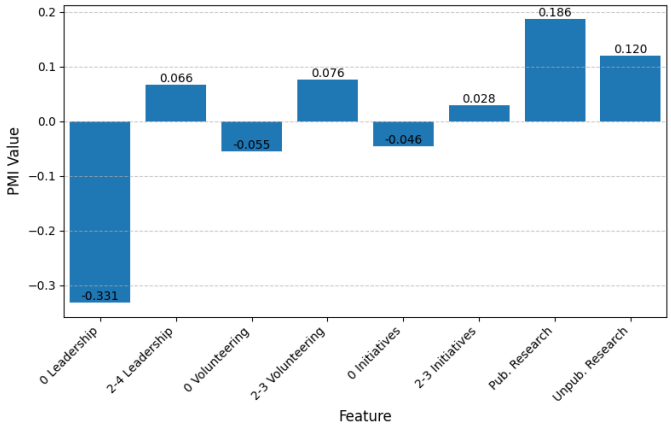

Fig. 2. PMI values of LLM-identified features with acceptance to one or more T20 universities.

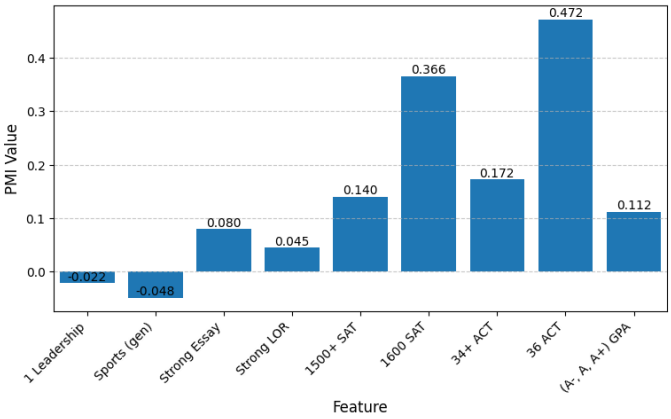

Fig. 3. PMI values of other notable features with acceptance to one or more T20 universities.

This analysis serves to validate the LLM's findings and confirm that traits emphasized by the LLM also show measurable associations with T20 admissions in the dataset. The

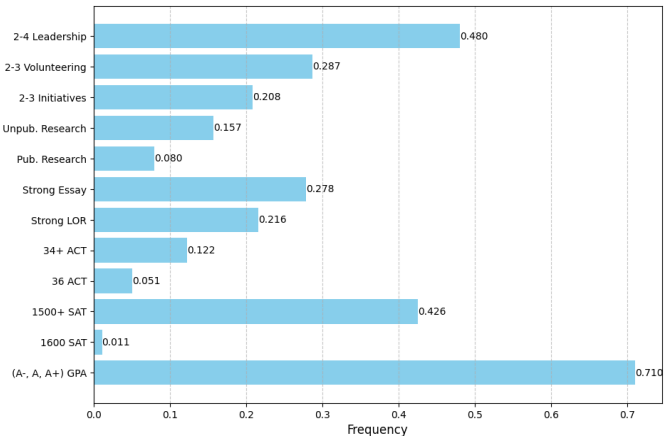

Fig. 4. Frequency of significant features in the dataset.

alignment between qualitative observations and quantitative trends supports the reliability of the patterns observed by the LLM.

E. Alignment with Prior Works

As noted in the introduction, the college application process is relatively underexplored in academic research. To contextualize our results, we compare our findings with a 2023 survey by the National Association for College Admission Counseling (NACAC), which reported on the factors admissions officers consider most important in evaluating applicants. While not a statistical evaluation, the NACAC *Factors in the Admissions Decision* survey [29] offered meaningful insight from admissions officers who evaluate applications firsthand. Our findings showed partial alignment with the survey, which notably ranked "positive character attributes" as the third most important factor, behind only GPA and the rigor of high school coursework. This result was somewhat unexpected and served to further validate the LLM's conclusion that demonstrating "ideal personal qualities", such as leadership, initiative, and community-mindedness, can meaningfully strengthen an application.

VI. Actionable Insights for Applicants

The preceding analysis established consistent patterns across both LLM-generated insights and quantitative validation, identifying traits such as academic achievement, leadership, initiative, and sustained engagement as strongly associated with admissions outcomes. In this section (Part B, contribution #3), we build on these findings by formalizing them into actionable guidance for applicants.

A. Evidence-Based Insights

While academic metrics such as GPA remain foundational, other non-academic factors prove to contribute significantly to the strength of an application. Importantly, these factors are largely within students' control, as they are shaped by the choices applicants make in how they spend their time, pursue their interests, and present their experiences throughout the application.

Of these factors, demonstrated leadership, initiative, and community-mindedness (through volunteering) consistently emerged as the three most influential. Successful T20 applicants frequently held between two and four leadership roles and had undertaken two to three self-initiated projects during high school. These projects ranged from founding clubs and nonprofits to launching independent research or creative endeavors. Successful applicants also typically held between two to three volunteering roles as well, demonstrating a sense of care for the community.

Notably, the presence of such activities and the effective communication of the underlying traits they reflect, which we term "ideal personal qualities," proved as influential in admissions outcomes as more traditionally emphasized application components such as essays, letters of recommendation, and standardized test scores. This finding is supported not only by our own data but also by external benchmarks such as the NACAC Factors in the Admission Decision survey [29], underscoring the value for students in ensuring that their applications clearly demonstrate these personal qualities.

In sum, students aiming for admission to competitive institutions should understand that while academic excellence is necessary, it is often not sufficient on its own. Demonstrating the ability to take initiative, lead with purpose, and engage meaningfully with one's community will allow students to take ownership of how they are perceived and improve their chances of admission to the United States' most selective universities.

VII. Discussion

While we are confident our work has provided a capable domain-adapted LLM and meaningful insights into the college admissions process, there are some limitations to this study that should be addressed, along with some areas for future expansion.

A. Data Limitations

Due to FERPA regulations and broader privacy concerns, we were unable to access a large enough volume of verified historical application data from institutional sources, limiting our ability to analyze a more comprehensive and diverse applicant pool. As a result, we relied on self-reported data from the r/CollegeResults subreddit, which introduces several limitations.

First, self-reported data depends on the honesty and accuracy of individual users, with no mechanism for external verification, although we do use our own process to verify the data. Second, the subreddit's user base is not representative of the overall college applicant population. It is heavily skewed toward students who are more competitive and deeply invested in the admissions process. This bias is evident in our dataset, where 54% of students were accepted to at least one T20 college, far above the national average. It also leads to the over-representation of certain features; for example, nearly 50% of students in our sample reported conducting some type of academic research, which does not reflect broader applicant trends.

Finally, our sample size of approximately 1,000 entries per admissions cycle limited the statistical power needed to detect more subtle patterns, particularly for rare features or feature combinations. The limited availability of data also constrained the size of our evaluation set, which contained only 60 test cases.

B. Limitations on Statistical Verification

While the statistical analysis provided us valuable insights into the relationships between applicant features and T20 admissions outcomes and allowed us to verify findings from the LLM, several limitations should also be acknowledged here.

First, the analysis required the conversion of qualitative, often subjective, textual data into categorized binary features

(e.g., strong essay, published research, 2-4 leadership roles). This process, while necessary for quantification, introduces simplifications that may ignore important nuances. In the real world, college applications are evaluated based on the context of the particular applicant, all of which is ignored when encoding applications into binary values.

Second, as discussed earlier, PMI values are sensitive to low-frequency features, especially when a feature or its co-occurrence with admission outcomes is rare. In these cases, PMI can be artificially inflated or deflated, making it difficult to interpret small-sample results as reliable indicators of association strength. Also, PMI assumes independence between features, although real-world application characteristics often co-occur, limiting the method's ability to capture combined effects, such as the "ideal personal qualities" identified by the LLM.

C. Model Limitations

Lastly, our domain-adapted LLM is also subject to inherent limitations. Training on a relatively small and skewed dataset may lead to overfitting or the amplification of spurious correlations. While chain-of-thought prompting improves transparency, the model's outputs remain not fully interpretable, and some identified correlations may reflect dataset bias rather than causal relationships. Moreover, reliance on self-reported, competitive-focused data risks biasing the model toward traits emphasized by highly engaged students, which could misrepresent broader admissions patterns.

D. Directions for Future Work

Future studies could address these limitations by incorporating verified institutional data using privacy-preserving methods, expanding datasets to better capture underrepresented populations, and including rarer applicant features. Additional work could explore richer representations of qualitative content, such as essays or recommendation letters, and utilize more sophisticated statistical approaches to capture feature interactions.

VIII. Conclusion

Drawing on historical college application and admissions data from the three most recent college admission cycles (2022-2025), this study makes three key contributions to the understanding of selective college admissions. First, we curated and released a cleaned, anonymized corpus of real student applications and outcomes, standardized for use in natural language processing. Second, we fine-tuned an open-source large language model, Mistral-Small-24B-Instruct-2501, on this dataset to produce a domain-adapted model capable of capturing latent patterns in the college admissions process. Third, by leveraging both the outputs of our domain-adapted model and statistical validation techniques like Pointwise Mutual Information (PMI), we identified actionable insights into the factors that most consistently influence admissions outcomes, such as demonstrated leadership, initiative, and community-mindedness.

Together, these contributions enable a multifaceted analysis of the college admissions process and highlight not only which student-controlled features carry the most significance, but also how applicants can strategically present their strengths to competitive institutions. Future work could extend these contributions by incorporating verified institutional datasets, simulating diverse applicant scenarios, and exploring richer qualitative representations, enabling more personalized, scalable, and data-driven guidance as the college admissions landscape continues to evolve.

References

[1] OpenAI et al., "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, Mar. 2024. DOI: 10.48550/arXiv.2303.08774.

[2] Gemini Team Google, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, May. 2025. DOI: 10.48550/arXiv.2312.11805.

[3] Aaron Grattafiori et al., "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, Jul. 2024. DOI: 10.48550/arXiv.2407.21783.

[4] DeepSeek-AI et al., "DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, Jan. 2025. DOI: 10.48550/arXiv.2501.12948.

[5] T. Gao, J. Jin, Z. T. Ke, and G. Moryoussef, "A comparison of DeepSeek and other LLMs," *arXiv preprint arXiv:2502.03688*, Feb. 2025. DOI: 10.48550/arXiv.2502.03688.

[6] A. Q. Jiang et al., "Mistral 7B," *arXiv preprint arXiv:2310.06825*, Oct. 2023. DOI: 10.48550/arXiv.2310.06825.

[7] O. L. Liu, "Holistic admissions in higher education: challenges and promise," *Journal of Postsecondary Student Success*, vol. 1, no. 4, pp. 1–19, Jul. 2022. DOI: 10.33009/fsop_jpss131099.

[8] C. Claybourn, "How colleges choose which students to admit," *U.S. News & World Report*, Aug. 16, 2022. [Online]. Available: https://www.usnews.com/education/best-colleges/articles/how-colleges-choose-which-students-to-admit

[9] R. Rubin, "Ivy League acceptance rates [updated for 2024 admissions]," *Spark Admissions*, Dec. 23, 2024. [Online]. Available: https://www.sparkadmissions.com/blog/ivy-league-acceptance-rates/

[10] C. Avery, "The effects of college counseling on high-achieving, low-income students," NBER Working Paper no. 16359, National Bureau of Economic Research, Cambridge, MA, Sept. 2010.

[11] S. B. Kyte, C. Atkins, E. Collins, and R. Deil-Amen, "Understanding the impact of data-driven tools on advising practice and student support," *Journal of Postsecondary Student Success*, vol. 2, no. 4, pp. 1–23, 2023. DOI: 10.33009/fsop_jpss132841.

[12] S. Kim and M. N. Bastedo, "Who gets their first choice? Race and class differences in college admissions outcomes," *AERA Open*, vol. 10, 2024. DOI: 10.1177/23328584241298951.

[13] E. J. Castilla and E. J. Poskanzer, "Through the front door: why do organizations (still) prefer legacy applicants?," *American Sociological Review*, vol. 87, no. 5, pp. 782–826, Oct. 2022. DOI: 10.1177/00031224221122889.

[14] E. Ben-Michael, A. Feller, and J. Rothstein, "Varying impacts of letters of recommendation on college admissions: approximate balancing weights for subgroup effects in observational studies," *arXiv preprint arXiv:2008.04394*, Feb. 2021. DOI: 10.48550/arXiv.2008.04394.

[15] C. Avery, A. Fairbanks, and R. Zeckhauser, *The Early Admissions Game: Joining the Elite*. Cambridge, MA: Harvard Univ. Press, 2004.

[16] Watchful1, "PushshiftDumps," GitHub, Mar. 27, 2025. [Online]. Available: https://github.com/Watchful1/PushshiftDumps

[17] OpenAI, GPT-4o-mini, OpenAI API, 2024. [Online]. Available: https://platform.openai.com

[18] Mistral AI Team, Mistral-Small-24B-Instruct-2501, Hugging Face, 2025. [Online]. Available: https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

[19] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, May 2020. DOI: 10.48550/arXiv.2004.10964.

[20] Z. Guo and Y. Hua, "Continuous training and fine-tuning for domain-specific language models in medical question answering," *arXiv preprint arXiv:2311.00204*, Nov. 2023. DOI: 10.48550/arXiv.2311.00204.
[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, Oct. 2021. DOI: 10.48550/arXiv.2106.09685.
[22] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: efficient finetuning of quantized LLMs," *arXiv preprint arXiv:2305.14314*, May 2023. DOI: 10.48550/arXiv.2305.14314.
[23] OpenAI, GPT-4.1, OpenAI API, 2025. [Online]. Available: https://platform.openai.com/
[24] Google DeepMind, Gemini 2.5 Pro, Google AI for Developers, 2025. [Online]. Available: https://ai.google.dev/
[25] OpenAI, GPT-4.1-mini, OpenAI API, 2025. [Online]. Available: https://platform.openai.com/
[26] Google DeepMind, Gemini 2.5 Flash, Google AI for Developers, 2025. [Online]. Available: https://ai.google.dev/
[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, Jan. 2023. DOI: 10.48550/arXiv.2201.11903.
[28] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, Jan. 2023. DOI: 10.48550/arXiv.2205.11916.
[29] NACAC, "Factors in the admission decision," National Association for College Admission Counseling, Aug. 23, 2023. [Online]. Available: https://www.nacacnet.org/factors-in-the-admission-decision/

APPENDIX A
PROMPT USED FOR MODEL EVALUATION

Given the following college application, respond with 'yes' if the student would be accepted to a Top 20 U.S. college, and 'no' if not. If you are unsure, respond with 'no'.

APPENDIX B
PROMPT STARTER USED FOR INSIGHT EXTRACTION

You are an AI trained on thousands of individual historical college applications paired with their admissions outcomes. Using your specialized knowledge of patterns in successful applications to the top 20 colleges, think step-by-step to answer the following question: Based on the extracurricular activities and roles available in each individual application...