Automating the Curation and Summarization of Rheumatology Fellowship Applications using Generative Artificial Intelligence (AI) Agents and Transformer Models

Andrew R. Trotter[2], John D. Osborne (PhD)[1], Amanda S. Alexander (MD)[2], Angelo L. Gaffo (MD)[2], Jeffrey R. Curtis (MD, MS, MPH)[2]

[1] Department of Biomedical Informatics and Data Science, the University of Alabama at Birmingham, AL 35205, USA

[2] Department of Immunology and Rheumatology, the University of Alabama at Birmingham, AL 35205, USA

Introduction

- Each year, rheumatology fellowship application reviewers estimate the manual processing and curation of fellow applicant data takes **~40-50 hours each application cycle**. The growing accessibility of building custom agentic frameworks and fine-tuning artificial intelligence (AI) models allows for opportunities to automate these tasks in ways previously very difficult(1). We introduce a pipeline applying advances in machine learning to automate many tasks in a manner reproducible by reviewers on safe, local computing environments. This software will be released on GitHub with a suite of tools allowing for other fellowship programs to utilize the same general pipeline, while also being able to implement their own data and preferred feature types.

Methods

- We test the performance capabilities of several machine learning model archetypes for extracting and analyzing key terms and phrases, for predicting whether a rheumatology fellow applicant received an interview, and for generating an accurate narrative synthesis (one-to-two-page summary) for applicants. We note significantly better performance on a fine-tuned Clinical-Longformer(2) compared to a zero-shot GPT 4o-mini (deployed on Azure using a HIPAA-compliant server), also getting similar performance to 4o-mini with a logistic regression model purely using USMLE (MD physicians) & COMLEX (Doctor of Osteopathic Medicine [DO] physicians) exam scores. Additionally, we highlight factors important to UAB reviewers, testing the significance of key features like ties to the southeastern geographic region, USMLE score percentiles, medical school locations, and residency program locations, and identify significant variation in feature data across years. These variations across many years make it difficult to train accurate models.
- We utilize UAB rheumatology fellow applicant data provided for download from the Electronic Residency Application System (ERAS) portal from **2017-2024**; we use spreadsheets with **2,675 columns** containing structured and semi-structured features for all applicants, and **~40-page PDFs per-applicant** which include additional unstructured content like letters of recommendation (LoRs), school/program evaluations, and personal statements. Our pipeline automatically curates these sources into compact, easy-to-read text files for each applicant. Of the **1308 total applicants** in the corpus, **184 received interviews** and **1124 did not**. Roughly ~23-24 applicants received an interview each year, but there was no set minimum required.

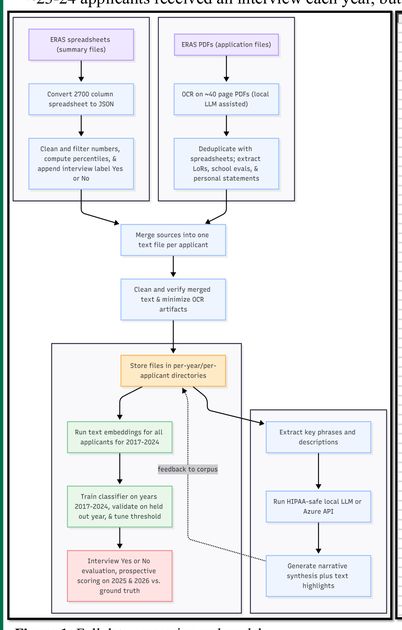


Figure 1: Full data processing and model evaluation pipeline steps.

Feature (normalized)	Count
ACLS	1308
LoR Received	1308
Malpractice Cases Pending	1308
Felony Conviction	1308
Medical Licensure Problem	1308
PALS	1308
Board Certification	1308
Personal Statement Received	1308
Applicant Name	1308
Medical Education or Training Interrupted	1308
Misdemeanor Conviction	1308
COMLEX-USA Transcript Received	1308
Interviewed	1308
Composite Score	1308
Count of Non Peer Reviewed Online Publication	1308
Count of Oral Presentation	1308
Count of Other Articles	1308
Count of Peer Reviewed Book Chapter	1308
Count of Peer Reviewed Journal Articles/Abstracts	1308
Count of Peer Reviewed Journal Articles/Abstracts(Other than Published)	1308
Count of Peer Reviewed Online Publication	1308
Count of Poster Presentation	1308
Count of Scientific Monograph	1308
Language Fluency	1308
Missing Documents	1307
LoR Type	1306
LoR Author	1306
LoR Latest Date Received	1306
LoR Received_1	1305
LoR Latest Date Received_1	1305
LoR Type_1	1305
LoR Author_1	1305
Personal Statement Latest Date Received	1303
LoR Type_2	1301
LoR Latest Date Received_2	1301
LoR Received_2	1301
LoR Author_2	1301
Professional Experience Organization	1300
Professional Experience Organization_1	1282
ACLS Expiration Date	1254
Professional Experience Organization_2	1248
USMLE Step 1 Score	1218
USMLE Step 2 CK Score	1212
Professional Experience Organization_3	1209
Professional Experience Organization_4	1160
USMLE Step 3 Score	1123
Professional Experience Organization_5	1114
USMLE Step 2 CS Score	1087
Most Recent Medical School	1080
Professional Experience Description	1064
Professional Experience Description_1	1048
Awards & Recognitions - Others	1047
Professional Experience Organization_6	1046
Professional Experience Description_2	1019
LoR Latest Date Received_3	1017
LoR Received_3	1017
LoR Type_3	1017
LoR Author_3	1017
Expected Visa Status	1010
Professional Experience Description_3	987
Medical License State	968
Medical License Type	968

Figure 2: Examples of extracted and normalized features from ERAS spreadsheets.

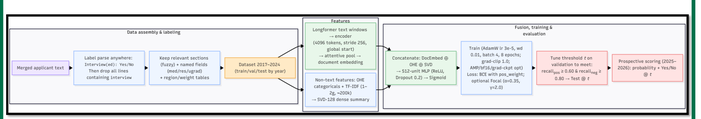
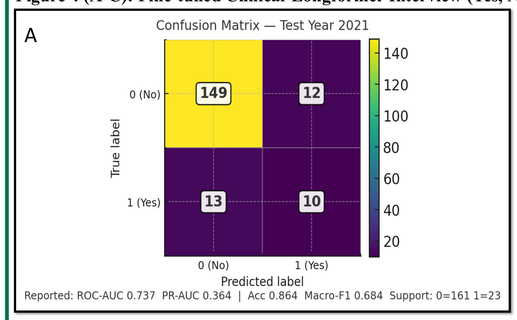
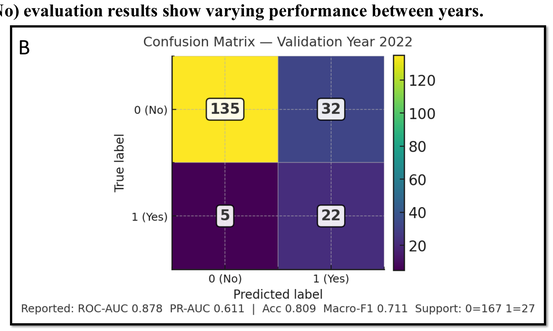


Figure 3: Model processing, training, and tuning pipeline. Longformer base and Clinical-Longformer(2) were trained and tested with this pipeline, the latter of which is displayed in the results section.

Results

Figure 4 (A-C): Fine-tuned Clinical-Longformer Interview (Yes, No) evaluation results show varying performance between years.





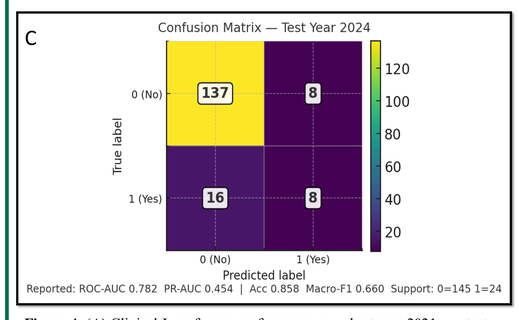
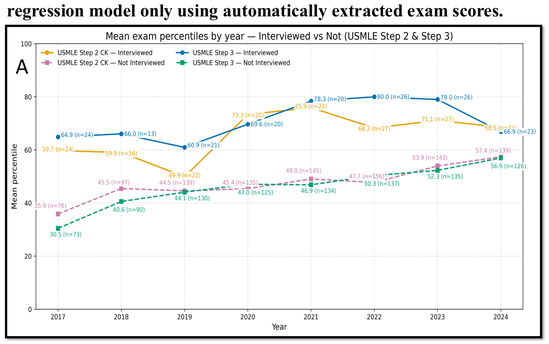
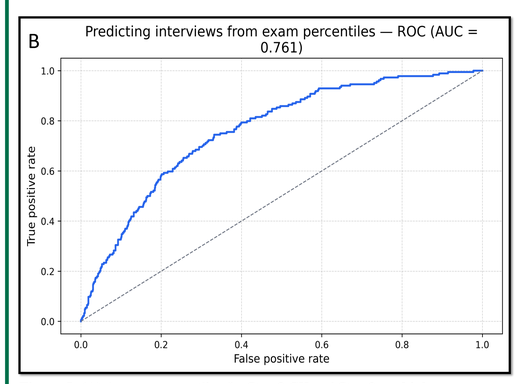


Figure 4: **(A)** Clinical-Longformer performance on cohort year 2021 as a test dataset. **(B)** performance on year 2022 as a validation dataset. **(C)** performance on year 2024 as a test dataset; **results clearly vary across years.**

Figure 5 (A-C): Evaluating Interview (Yes, No) label with a logistic regression model only using automatically extracted exam scores.





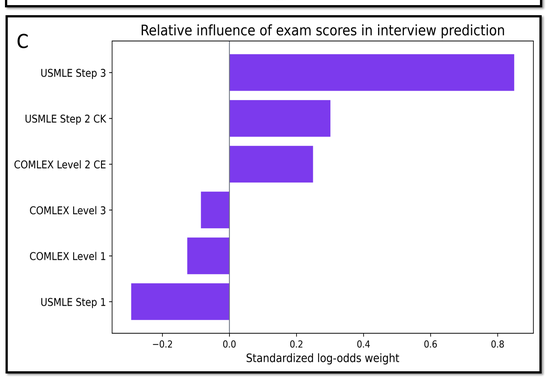
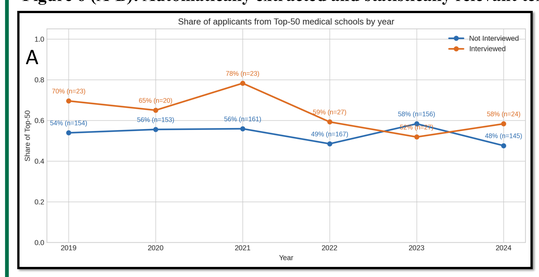


Figure 5: (A) Average percentiles for Step 2 CK and Step 3 are **higher every year** for interviewed applicants. **(B)** Performance of logistic regression model using **all USMLE and COMLEX percentiles**. **(C)** The weighted influence of exams and their scores on model performance. **COMLEX results may be misleading due to small sample size.**

Figure 6 (A-B): Automatically extracted and statistically relevant text-based features for model extraction and analysis.



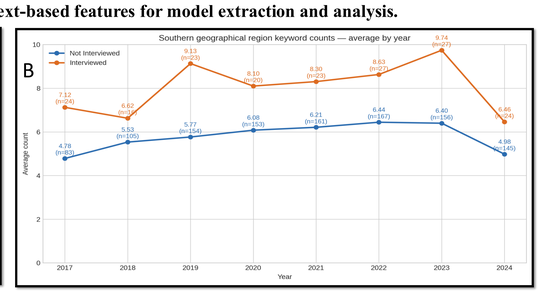


Figure 6: **(A)** Percentage of applicants who attended a top-50 ranked medical school. **(B)** Yearly average of southern geographical region terms from semi-structured ERAS data.

Results

Figure 7: GPT applicant scoring metric histogram.

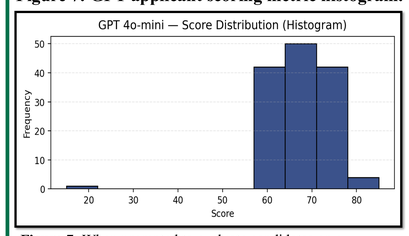


Figure 7: When prompted to evaluate candidates on a scale from 1-100, GPT 4o-mini almost always labels a candidate as **above average.**

Figure 8: Zero-shot GPT prediction results.

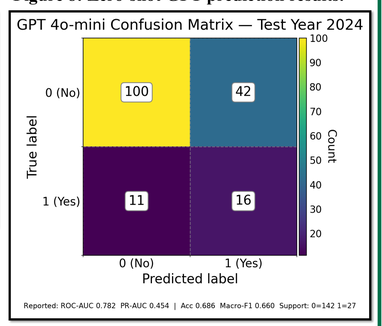


Figure 8: GPT 4o-mini (Azure endpoint API) performance on cohort year 2024 as a test dataset. Performance is significantly **worse than fine-tuned models.**

- Year effects were significant for **top-50 med-school status** (χ^2 p=1.0e-36, 2.1e-09) and **southern keyword means** (Kruskal–Wallis p=3.1e-10, 0.0449). By interview outcome, **southern keyword means** differed (Welch t p=1.2e-13), but **top-50 med-school counts** did not (two-proportion Z-test p=0.2009).

Discussion

- For rheumatology fellowship applications, our automated machine learning pipeline processes 2,675 structured Electronic Residency Application System (ERAS) features and ~40-page PDFs per applicant, extracts the information most relevant to program directors and faculty, outputs concise and easy-to-read text files per applicant, and produces accurate narrative summaries. This pipeline significantly reduces the time reviewers need to spend processing and managing applicant data.
- The interview-recommendation pipeline architecture and model fine-tuning still needs major development before deployment. Though, even if one could create a model with near-perfect performance, fully automating decisions to interview applicants raises many ethical concerns, especially given year-to-year data variability.
- Fine-tuned models, machine learning, and generative AI can still meaningfully assist with processing, staging, and deployment of application data, even without relying on predicted outcomes.
- The worse-case scenario is for a model to produce false-negatives, which would discriminate against applicants who would otherwise have received interviews. Therefore, it seems it would be better if a model were to have higher false positives if it could be highly accurate with true negatives and scarcely predict false negatives.
- Zero-shot API models are both ineffective and costly for this use. GPT-4o-mini/5o-mini produce reasonable summaries, but 4o-mini was sycophantic when scoring applicants. It seems better to fine-tune locally on LoRs, personal statements, etc., to extract key phrases and types of information (e.g., "good teamwork," rheumatology research interests and experiences) and generate summaries without API calls. It is necessary to protect applicant data and prevent leakage during training and evaluation, so it is best to use local models or LLMs behind an institutional firewall or other protected servers.

Conclusion

- Modern machine learning pipelines can offer significant benefits to rheumatology fellowship application reviewers (e.g. faculty at academic medical centers) by eliminating the need for tedious, manual review of many very long and dense documents, with our automated pipeline achieving substantial time savings by condensing these complicated files into concise text files for each applicant.
- Time savings notwithstanding, training highly accurate models for interview prediction is difficult, particularly for correctly predicting who was offered an interview. Human-in-the-loop insights remain crucial for continuing to develop these fine-tuned language models which need many well-curated feature types to properly contextualize what it means to be a "good applicant."
- The applicant data curation and summarization pipeline is almost ready for public release, but model fine-tuning requires further development to accurately suggest candidates for interview prioritization.

References

Drum, Benjamin MD, PhD1; Shi, Jianlin MD, PhD2; Peterson, Bennet3; Lamb, Sara MD4; Hurdle, John F. MD, PhD5; Gradick, Casey MD, MPH6. Using Natural Language Processing and Machine Learning to Identify Internal Medicine–Pediatrics Residency Values in Applications. Academic Medicine 98(11):p 1278-1282, November 2023. | DOI: 10.1097/ACM.0000000000005352 (1)

Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc. 2023 Jan 18;30(2):340-347. doi: 10.1093/jamia/ocac225. PMID: 36451266; PMCID: PMC9846675. (2)

ACKNOWLEDGEMENTS

We acknowledge the support of NIAMS P30AR072583 “Building and InnovatinG: Digital heAlth Technology and Analytics (BIGDATA)”

uab.edu