Human and LLM-Based Resume Matching: An Observational Study

Swanand Vaishampayan, Virginia Tech Chris Brown, Virginia Tech

Introduction

Resume matching assesses the extent of overlap between the content of candidate resume and the job description. Modern hiring is increasingly automating tasks related to resume matching such as parsing and rating resumes using NLP based deep learning models to save time and increase efficiency. Given increasing interest in using large language models (LLMs) for this purpose, we explore the applicability of GPT-4, the most recent advancement in NLP, for resume matching in an observational study. We compare zero-shot GPT-4 and human resume ratings for 736 resumes submitted to job openings from diverse fields—analyzing the resumes using real-world resume matching criteria. We concentrate on understanding "how" ratings are formulated for these constructs to investigate differences between GPT-4 and human ratings. Additionally, we analyze the effect of prompt engineering techniques, such as Chain of Thoughts (CoT), on the GPT-4 ratings and compare differences in GPT-4 and human ratings across racial and gender groups.

Study Design & Data Analysis

The 736 resumes each received two ratings: a) Human ratings from 4 raters; and b) an LLM-generated rating on constructs of Work experience, Skills, Certifications and Education. Through A pilot interview study with 3 experienced recruiters, we finalized the following rating scale:

- 1: Vastly does not meet minimum requirements
- 2: Does not meet minimum requirements
- 3: Meets minimum requirements
- 4: Exceeds minimum requirements
- 5: Vastly exceeds minimum requirements

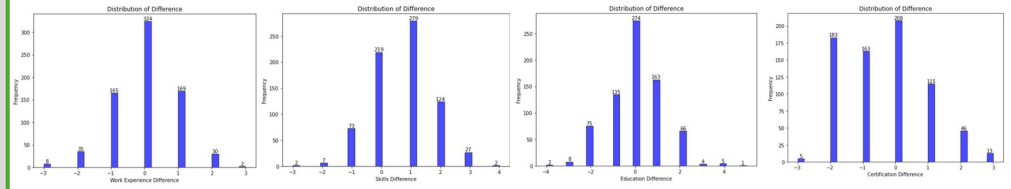
Following methods were used for analysis:

- Human-GPT Agreement: Fleiss Kapaa, Pearson's correlation
- Human-GPT Reasoning Differences: Open coding approach
- Human-GPT Group Differences: Standardized Cohen's d

Research Questions and Findings

RQ1: How do zero-shot GPT-4 and human resume ratings compare across constructs such as work experience, skills, educational qualifications and certification(s), given the job description?

- GPT and Human resume ratings differ in terms of work experience, skills, education and certifications.
- GPT ratings are more lenient across skills but are more stringent for certifications when compared to human ratings



RQ2: What is the effect of prompt engineering techniques such as Chain of Thought (CoT) on zero-shot LLM ratings?

- GPT-4 rating performance improved using all three prompt engineering techniques (Task based, Task based CoT, and Task based CoT with example)
- Certifications saw the biggest improvement in agreement with humans while work experience had the least improvement for advanced prompt engineering techniques compared with zero-shot GPT-4

Category	Zero-Shot GPT-4	Task Based	Task Based CoT	Task Based CoT with Example	Total samples
Work exp.	65	70	60	68	198
Skills	42	47	45	57	198
Education	62	102	110	88	198
Certifications	34	108	119	122	198

Samples with Perfect Match Between GPT and Human Ratings

Category	Zero-shot GPT-4	Task Based	Task Based COT	Task Based COT with Example
Work Experience	0.1104	0.2643	0.2562	0.3625
Skills	0.0697	0.2174	0.2367	0.3208
Education	0.2317	0.4792	0.4961	0.6109
Certification	0.1497	0.3677	0.4310	0.6588

Correlation for Prompt Engineering Techniques Compared to Human Raters

RQ3: What are the group differences in scores generated by GPT-4 and humans across race/ethnicity and gender demographics?

For Intra group human differences:

- Statistically significant differences for work experience across Asian and White subgroups with close to large magnitude of observed effect size
- Statistically significant differences for work experience across African American and White subgroups with medium magnitude of observed effect size

For Intra Group GPT differences:

- Statistically significant differences for work experience across African American and White subgroup with close to medium magnitude of observed effect size

For inter-group GPT-Human rating differences:

- Human ratings for work experience and certifications across Asian and White subgroups differ significantly more than GPT
- GPT ratings for certifications across Male and Female subgroups differ significantly more than human ratings

Demographic	Work Experience		Skills	
	GPT-4	Human	GPT-4	Human
Asian-White	0.3284 [0.1155, 0.5413]	0.7651 [0.5475, 0.9827]	0.1406 [-0.0714, 0.3526]	0.1264 [-0.0856, 0.3384]
African-American-White	0.4892 [0.2511, 0.7273]	0.5415 [0.3028, 0.7802]	0.3719 [0.1347, 0.6091]	0.3258 [0.0889, 0.5627]
Hispanic/Latino-White	0.3184 [0.1107, 0.5261]	0.2433 [0.0361, 0.4505]	0.0625 [-0.1442, 0.2692]	0.0809 [-0.1258, 0.2876]
Multiracial-White	0.4489 [0.1879, 0.7099]	0.4301 [0.1693, 0.6909]	0.4147 [0.154, 0.6754]	0.1002 [-0.159, 0.3594]
Male-Female	0.3627 [0.2164, 0.509]	0.1435 [-0.0018, 0.2888]	0.1332 [-0.0121, 0.2785]	0.0022 [-0.1429, 0.1473]

Demographic	Education		Certification	
	GPT-4	Human	GPT-4	Human
Asian-White	0.3282 [0.1153, 0.5411]	0.3472 [0.1342, 0.5602]	0.1279 [-0.0841, 0.3399]	0.5635 [0.3485, 0.7785]
African-American-White	0.0794 [-0.1565, 0.3153]	0.3734 [0.1362, 0.6106]	0.2249 [-0.0114, 0.4612]	0.1544 [-0.0817, 0.3905]
Hispanic/Latino-White	0.2799 [0.0725, 0.4873]	0.0818 [-0.1249, 0.2885]	0.1235 [-0.0833, 0.3303]	0.2786 [0.0712, 0.486]
Multiracial-White	0.0648 [-0.1944, 0.324]	0.2873 [0.0274, 0.5472]	0.0428 [-0.2163, 0.3019]	0.0375 [-0.2216, 0.2966]
Male-Female	0.0313 [-0.1138, 0.1764]	0.1382 [-0.0071, 0.2835]	0.3848 [0.2384, 0.5312]	0.0284 [-0.1167, 0.1735]

Comparison of GPT-4 and Human Ratings by Demographic Groups Across Work Experience, Skills, Education and Certification. Values in Brackets are the respective Confidence Intervals

Acknowledgements:
Special thanks to Dr. Louis Hickman and Dr. Brent Stevenor for their help and collaboration.