An Investigation into Model Performance for Identification of Human vs. AI text

A report to the Quantum Insight AI research team.

Jenny, Lauren, Amalia

2024-10-11

1 Statement of the Problem

Quantum Insight is aiming to distinguish human-authored text from AI-generated content using a parallel corpus of human and machine-generated text. The initial models trained with Biber's linguistic features performed well in identifying AI-generated text on the training set, achieving high in-sample accuracy. However, when applied to new data from a different corpus, model accuracy decreased significantly. Our goal is to determine the factors contributing to this performance drop and recommend ways to improve model accuracy across different corpora.

2 Summary of Findings

We initially determined a subset of the original models we would investigate. These models were chosen because they were representative of their LLM type and had the smallest and largest differences in overall accuracy and loss of accuracy. We first used the varImp function to identify the most important features for classification in a subset of the original models. We then used correlation matrices to find commonalities between the relationships of features belonging to human-written and AI-generated texts.

A problem arises when a comparison is drawn for the correlation between the top Biber features used for classification by each of the analyzed models in Tables 1, 2, and 3 and other features. Many of the classification features were similarly positively and negatively correlated with similar features across all relevant data sets. This issue can be highlighted by observing the ChatGPT 4o model, where nominalization, present participle, and mean word length become important. As expressed in Figure 1, nominalization is similarly positively correlated with phrasal coordination, and mean word length is similarly positively and negatively correlated with adj_attr and other_nouns, respectively. This lack of significant difference could have led to confusion during classification.

Furthermore, for the important features in which differences can be observed (e.g. present participle), similar relationships cannot be distinguished within the original data. We discov-

ered similar problems with the Meta Llama 70B and Meta Llama 70B Instruct models with an additional notable problem being the past tense feature. This feature was categorized as important for the model according to Table 2, but was not present in the out-of-sample data. This discrepancy is due to the prompts for the out of sample data requesting that all responses be worded in "active voice", which resulted in a loss of categorizing capabilities for the Meta Llama 70B Instruct model.

Following this line of thought, we examined the text to determine whether the prompts could have contributed to the similar correlations across samples. A potential contributing factor was that the prompts for the initial data sets were to "extend" a human-generated text. This specification is highly likely to have resulted in limited differences between the machine and human data sets, which can be observed in Figures 2 and 3 and Tables 4, 5, and 6. There are a limited amount of discrepancies between the LLMs and the humans, though the greatest differences can be observed in the ChatGPT 4o models (Figure 2). Overall, the models trained to look for very specific patterns rather than something broad and widely applicable.

3 Recommendations

Future research could involve collecting new training and testing datasets similar to the out of sample dataset where human and machine text were generated through open-ended prompts. This could result in more variance between the human and machine texts. We could also investigate different models for classification. Currently the models are random forests which could be more susceptible to over-fitting with large and over-specified data. Training the models using lasso regression or other machine learning models might improve performance. Additionally, this data was generated by a previous version of ChatGPT, so it could be that ChatGPT performs well because it has a very specific style of writing. Comparing performance on machine generated text from other LLMs could generate more insights into the problems seen in our models.

4 Appendix

Table 1: Top 5 Important Features for ChatGPT

Feature	Overall
f_25_present_participle	13.346290
f_44_mean_word_length	8.745116
f_14_nominalizations	8.075695
f_29_that_subj	7.525795
f_17_agentless_passives	6.452369

Table 2: Top 5 Important Features for Meta Llama 70B Instruct

Feature	Overall
f_24_infinitives	6.490659
f_42_adverbs	5.804206
f_01_past_tense	5.438775
f_65_clausal_coordination	5.168773
f_25_present_participle	4.933457

Table 3: Top 5 Important Features for Meta Llama 70B

Feature	Overall
f_39_prepositions	5.201562
f_38_other_adv_sub	4.953442
f_49_emphatics	4.806458
f_51_demonstratives	4.241690
f_57_verb_suasive	4.101157

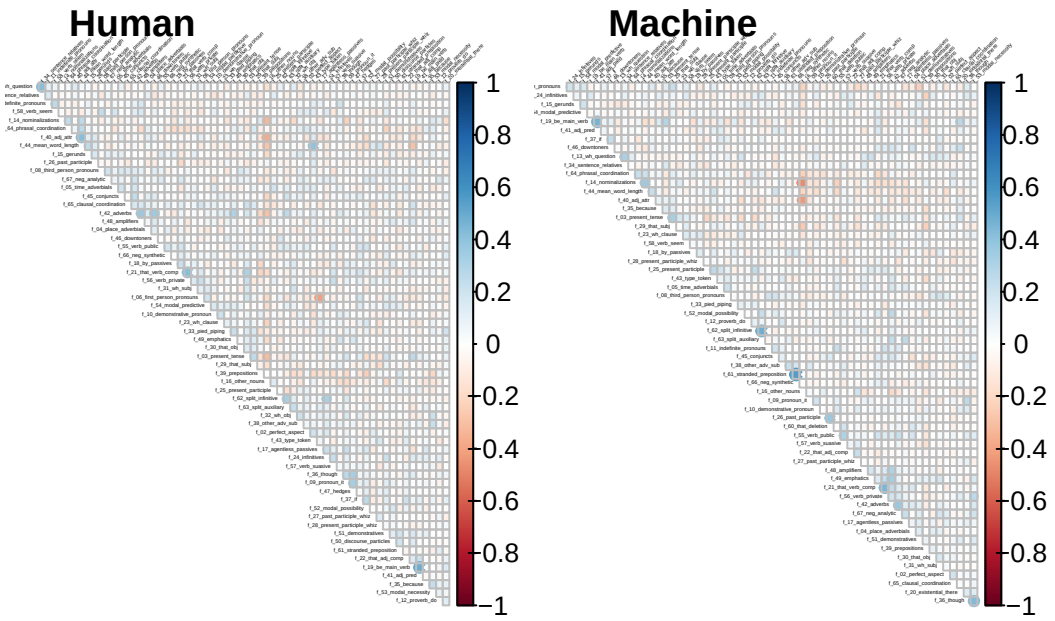

Figure 1: Correlation Matrices of Out-Sample Human and Machine

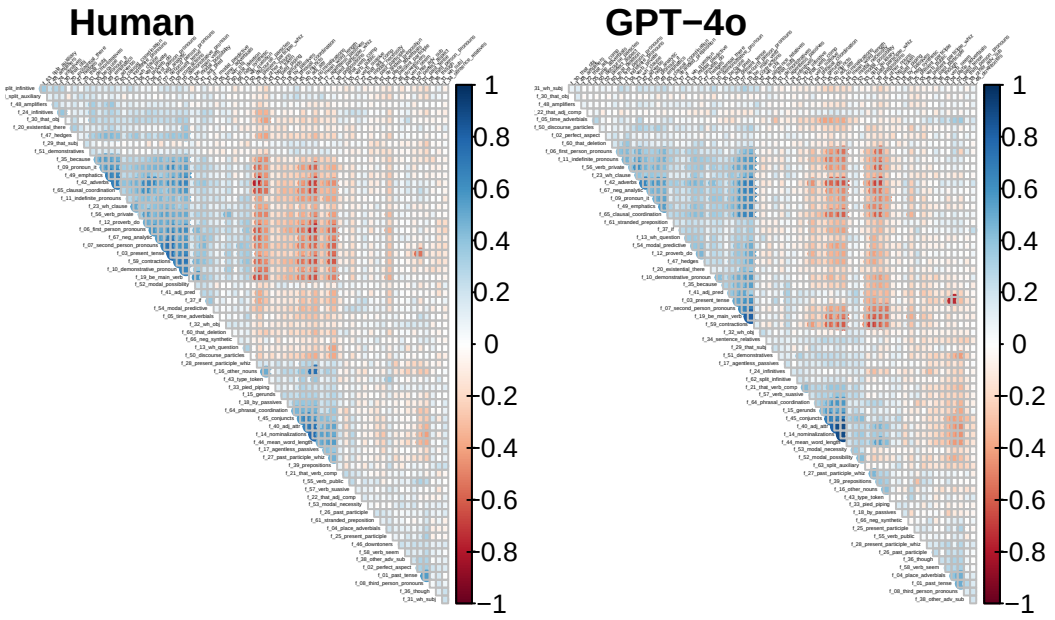

Figure 2: Correlation Matrices of In-Sample Human and GPT-4o

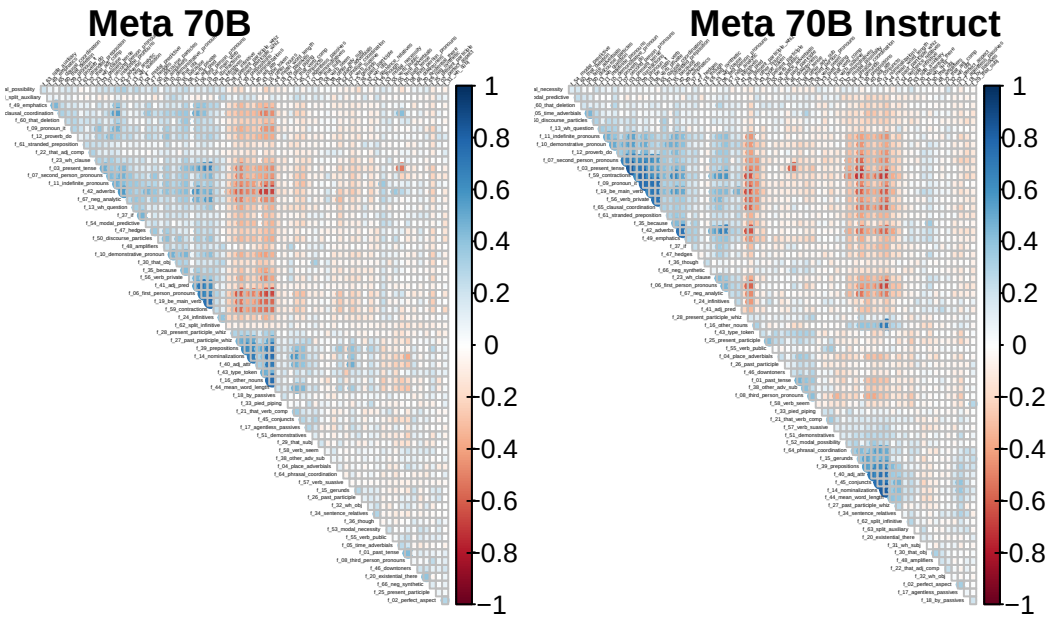

Figure 3: Correlation Matrices of Meta 70B and Meta 70B Instruct Models

Table 4: Highest Frequency Tokens in Meta Llama 70B

feature	frequency	rank
that	1272	1
with	574	2
they	501	3
have	411	4
what	405	5
this	404	6
move	364	7
like	358	8
then	331	9
about	309	10

Table 4: Highest Frequency Tokens in Meta Llama 70B Instruct

feature	frequency	rank
that	1272	1
with	574	2
they	501	3
have	411	4
what	405	5
this	404	6
move	364	7
like	358	8
then	331	9
about	309	10

Table 5: Highest Frequency Tokens in In Sample Human

feature	frequency	rank
that	1206	1
with	688	2
have	498	3
this	489	4
like	460	5
they	364	6
what	362	7
from	354	8
know	276	9
were	256	10

Table 6: Highest Frequency Tokens in GPT 4o

feature	frequency	rank
with	1230	1
that	1211	2
this	498	3
their	441	4
into	403	5
from	381	6
like	372	7
they	311	8