# Automated Detection of Invalid Responses to Creativity Assessments

Antonio Laverghetta Jr.,[a*] Simone A. Luchini,[b] Jimmy Pronchick,[b] Roger E. Beaty[b]

[a]Quantitative Analysis Center

Wesleyan University

Allbritton Center 108, 222 Church Street, Middletown, CT 06459, USA

[b]Department of Psychology

The Pennsylvania State University

140 Moore Building, 138 Fischer Rd, University Park, PA 16802, USA

alaverghetta@wesleyan.edu

**Declaration of Interest:**

The authors declare no competing interests.

**Funding:**

This work was supported by a seed grant from the Pennsylvania State Center for Language Science. R.E.B. is supported by grants from the National Science Foundation [DRL1920653; DRL-240078; DUE-2155070].

**Author Note:**

All data, materials, and code are available at this OSF repo:

https://osf.io/u9knp/?view_only=0725705748fc4c0bb92913ec281fc84d

**Abstract**

Participants in creativity studies sometimes produce invalid data that is unusable for analysis, such as nonsensical or incomplete responses to idea generation tasks. Identifying such responses is a time-consuming yet necessary process to ensure robust results but also remains challenging to automate. We explore the efficacy of transformer language models (TLMs) for automatically detecting invalid creativity responses, using only the text prompt and response and no other metadata about the experimental session. We train a suite of transformers to detect invalid data for two creativity assessments: the Alternate Uses Task (AUT) and a design problems task (DPT). We find that transformers generally outperform other baselines for both tasks. Further, we show that TLMs' predictions are well calibrated to the quality of the participant response, ensuring that model failures will occur in a predictable way and that high-quality responses are unlikely to be labeled invalid. Finally, we conduct a fairness analysis based on language background—using an adversarial study where participants attempt to "break" the model by coming up with invalid responses that are nonetheless labeled valid. Our results demonstrate the potential of deep learning methods for cleaning creativity assessment data, using solely participant responses, in a reliable and unbiased way.

*Keywords: creativity, survey data quality, machine learning, natural language processing*

## 1. Automated Detection of Invalid Responses to Creativity Assessments

In creativity research, free response surveys are the norm for measuring divergent thinking and creative potential (Lubart et al., 2013; Qian & Plucker, 2021; Patterson et al., 2023), among numerous other creative constructs. A key challenge with any such study, however, is ensuring the gathered data is high quality, necessitating that participants complete the task faithfully and accurately. But this is far from a trivial task: participants can (and often do) deliver low quality data due to lack of understanding, poor effort, or attempting to "game" the system to receive payment as quickly as possible (Chmielweski & Kucker, 2020; Eyal et al., 2021). These problems are even more salient in online data collection, where ensuring compliance with instructions has historically been challenging (Barends & De Vries, 2019). Critically, even when low quality responses only constitute a small fraction of the collected data --- which is often the case --- failure to identify or filter them out can weaken statistical power or even nullify effects (Stostic et al., 2024), carrying important implications when such data is used when reporting results. Detecting such invalid (NA) responses both accurately and efficiently is a crucially important component of running creativity studies. Historically, the most viable method for identifying invalid free responses has involved manual coding by trained human raters (Brühlmann et al., 2020; Sischka et al., 2022). While effective, this method also has numerous limitations, such as the potential for rater disagreements and high costs for rater training (Forthman et al., 2023). Initial attempts to automatically evaluate response validity leveraged response times to detect low-effort responses (Huang et al., 2012). Despite their promise, these methods require analysis of study metadata, which is not always accessible. Further, past work has indicated that relying on response times alone is often ineffective, especially when the task requires memory retrieval or domain expertise (Meade & Craig, 2012; Leiner, 2019). Stronger

correlations have been found between human ratings of validity and predictions drawn from character or word count (Yeung & Fernandes, 2022). Given that many low-effort responses will consist of few letters or words with nonsensical meanings, this approach for NA detection uses word counts to filter out excessively short responses (Etz et al., 2018). Nevertheless, it is limited by the arbitrary selection of the cut-off point typically decided without support from empirical evidence (Kaczmirek et al., 2017; Banks et al., 2018; Gangnon et al., 2019). More recently, methods based on machine learning and word or sentence embeddings of participant responses have demonstrated promising results for detecting NA free responses (Yeung & Fernandes, 2022; Cibelli Hibben et al., 2024). Notably, Kucwaj and Kroczek (2025) recently evaluated the efficacy of ChatGPT for judging the validity of responses to an alternate uses task, finding that AI could improve objectivity in this crucial yet under-reported aspect of data preparation. Despite these advances, a comprehensive assessment of methods for NA response detection, comprising multiple models and tested against multiple types of creativity assessment, appears to be missing in the literature, which hinders the ability of researchers to screen such responses in an objective and consistent manner.

In the present work, we seek to bridge this gap by training and validating transformer language models (TLMs) to detect NA responses to two creativity assessments. TLMs are the backbone of modern natural language processing systems, as they have achieved state-of-the-art performance on tasks ranging from sentiment analysis (Wankhade et al., 2022) to language modeling (Devlin et al., 2019) and word sense disambiguation (Loureiro et al., 2021), among many others (Aftan & Shah, 2023). Their strength comes from the ability to represent semantically rich and context specific word and sentence meanings in their embeddings, unlike prior bag-of-words representations like TF-IDF (term frequency – inverse document frequency)

where words have the same representation regardless of context. In creativity research, finetuned TLMs have become a standard tool for automated originality and effectiveness scoring; they have been used to score creativity on tasks as diverse as metaphors (DiStefano et al., 2024), creative problem solving (Luchini et al., in press a), short stories (Luchini et al., in press b) and creative sentence completions (Organisciak et al., 2023). They are the fundamental building block of large language models (LLMs) like ChatGPT (Brown et al., 2020), which have proven effective at both automated scoring in their own right (Luchini et al., in press a; Organisciak et al., 2023), and automatically generating new items for creativity tests (Laverghetta Jr. et al., 2024).

We developed a suite of finetuned TLMs for NA detection across two creativity tasks, the alternative uses task (AUTs) and a design problems task (DPT). This approach allowed us to study invalid detection across multiple domains of creative thinking tasks. In two studies, we analyzed the performance of our TLMs, checking for their calibration with respect to criterion-related metrics, and performed bias analysis on our models to ensure they are fair with respect to L2 English speakers.

## 2. Study 1: Training and Validation of TLM NA Detectors

### 2.1. Methods

***2.1.1. Dataset Preparation.*** We curate data for both the AUT and the DPT (also abbreviated as CDT), using archival data for these tasks that were rated for NA. The AUT is a well-established task in creativity research that requires participants to think of alternative uses for common objects, such as a brick or a paperclip (Guilford, 1967). The DPT was proposed as a test of creativity in engineering (Luchini et al., 2025); a participant is given an engineering

problem, such as how to increase use of renewable energy or reduce traffic congestion in cities, and must develop novel and effective ideas to solve it. While both the AUT and DPT are divergent thinking tests, the AUT is more domain-general than the DPT, as the DPT specifically deals with scientific and engineering problem-solving, and benefits from knowledge of the subject matter. However, DPT items are still designed to not require domain expertise to solve. For the AUT, we use the English subset of data from Patterson et al (2023), who gathered more than 40,000 AUT responses across a multilingual sample. For the DPT, we use data from Luchini et al. (2025) who both validated the original items and gathered participant response data. Detailed sample statistics for both datasets can be found in their respective papers. For both datasets, a minimum of three raters annotated each response as being invalid or not. Raters obtained an ICC absolute agreement of 0.783 (95% CI: [0.779, 0.788]) for the AUT and 0.714 (95% CI: [0.700, 0.726]) for the DPT, indicating moderate to good agreement. We aggregate these labels by assigning each response as being invalid (1) if at least 60% of raters rated an AUT response as such, and 80% of raters for the DPT, marking all other responses as valid (0). We adopt these cutoffs to avoid rater-specific biases from influencing our models by ensuring that the majority of raters agree that the response is NA. Notably, Kucwaj and Kroczek (2025) found that human judges tend to have poor agreement with each other on the validity of AUT responses, with all judges only reaching perfect agreement for 58% of responses. We thus selected agreement thresholds that did not necessitate perfect agreement --- which is unlikely to occur --- while also preventing classifications based on only a minority of raters' judgements --- which could reflect noise in the rating process. Further, because DPT responses tend to be more complex than those for the AUTs, we reasoned that misclassifications were more likely to occur for this dataset and chose to adopt a more stringent threshold. While we do not claim that our

selection of agreement thresholds is optimal, testing classification methods for multiple agreement levels would dramatically increase the number of models to train, and we chose to leave this analysis to future work. All raters were trained research assistants who were familiar with the creativity tasks in question. Raters were given these instructions to rate for NAs:

> In each row, read the prompt in the PROMPT column and the response in the RESPONSE column. In the NA column, write "invalid" if a response is nonsensical, gibberish, or completely unrelated to the question. For the responses that are valid, don't write anything in that column.

We dropped any items that did not contain at least 1 valid and 1 NA response, as well as any rows with missing data, and converted all items to lowercase (certain items appeared in both upper and lowercase in the data). To construct the final dataset splits, we first held out ~15% of all items (three for the DPT and two for the AUT) given to participants to create an out-of-item (OOI) test set. We used the remaining data to construct train, validation, and out-of-response (OOR) test sets, using a 70 / 15 / 15 ratio and stratifying based on the item given to participants. Because OOI contains items not present in the training data, it is meant to be a stronger test of model generalization than OOR.

After performing the initial splits, we found that the training set for both tasks was severely imbalanced, with NAs comprising less than 10% of the training data. To mitigate issues arising from this class imbalance, we applied data augmentation to both training sets. We designed a series of transformations that, when applied to valid responses, we reasoned would consistently render them NA. These transformations were developed in consultation with other experts in the field, as well as RAs who routinely rate for NA in creativity data. Given a valid response with N words and M characters, we performed the following transforms:

1. **Nonsensical:** Replace between one to N words with a random English word from a dictionary, with the exact number varied randomly. A link to the dictionary is provided in the supplementary materials.
2. **Misspelled, Incomplete, or Ungrammatical:** Replace characters in the words of the response at random, doing so for between one to N words. Strip off the last M characters of the response at random, in the range one to M.
3. **Don't Know:** We manually created a list of variations of the response "I don't know", which we then augmented using ChatGPT-3.5. We replaced the response with one from this list at random. This list of responses is available in the supplementary materials.
4. **Didn't Understand or Follow Instructions:** Replace the response with a valid response from the other task (e.g., for the DPT, swap out a valid DPT response with a valid one from the AUT).

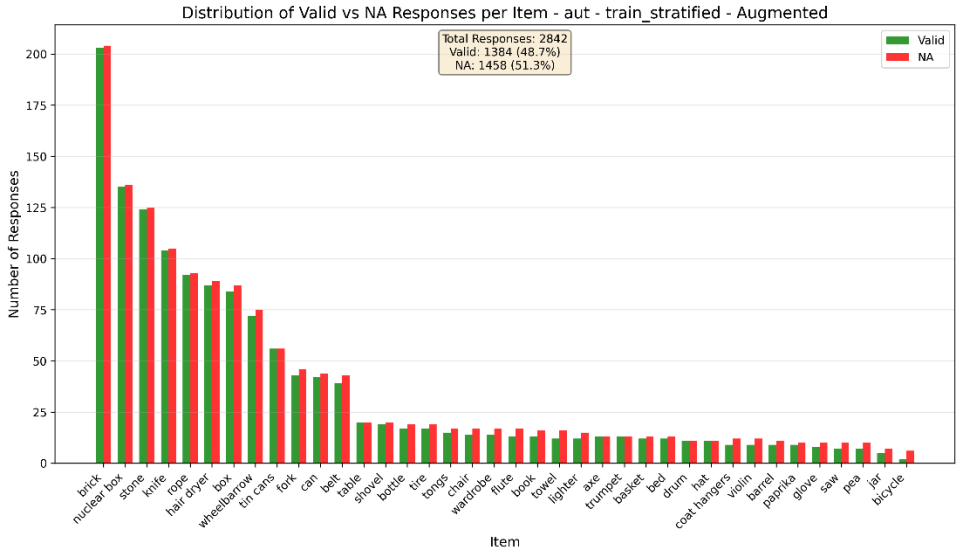
We parsed responses using the NLTK word tokenizer (Bird et al., 2009) and applied each invalid transformation to every valid response once. We do not claim that these transformations are comprehensive --- there are likely others that are more context-specific (e.g., irony or jokes about the task). However, we believe they reflect the types of NAs that one would expect a classifier to detect. We applied data augmentation to both task train sets, leaving the validation and test sets as is. This yielded 5504 NAs for the DPT train set and 3251 for the AUT train set.

Full dataset statistics are included in Table 1, with the statistics for the train set being after augmentation. Additionally, Figure 1 shows the number of responses broken down by both item and label for both training sets. With augmentation, the number of NAs is roughly balanced

across items. We include graphs for item level statistics for all other folds in Appendix A. High-resolution versions of all figures in the paper are available in the supplementary materials.

| Task | Train (Valid / NA) | Validation | OOR | OOI |
|---|---|---|---|---|
| DPT | 4372 / 4395 | 929 /23 | 934 / 18 | 1652 / 20 |
| AUT | 1374 / 1637 | 298 / 25 | 295 / 28 | 72 / 9 |

Table 1. Number of samples in each split of every augmented dataset. Numbers before the slash denote valid responses, numbers after the slash denote invalid responses. Please refer to the original publications for raw dataset statistics. DPT = design problems task; AUT = alternate uses task



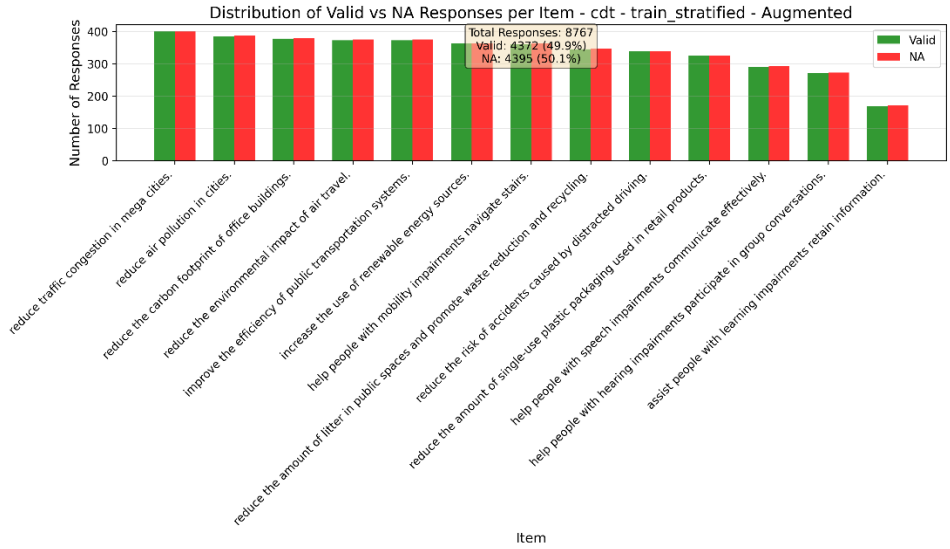



Figure 1. Item-level statistics for AUT (top) and DPT (bottom) training sets, after applying augmentation. The full text for every item is included on the x-axis. Please refer to Appendix A for all other splits.

***2.1.3. Modeling Problem Setup*** For all classifiers, we assume access to a set of *prompts* (*p*) administered to participants in a creativity experiment that contains items they provide *responses* (*r*) to. These responses may be NA, and the goal of the task is to determine whether to rate the response as such:

$$\mathcal{F}(\{X_i\}_{i=1}^{N}) = \begin{cases} 1 & \text{invalid} \\ 0 & \text{valid} \end{cases} \tag{1}$$

Where $x_i$ in Equation 1 is a prompt-response pair from the dataset $X$ and $\mathcal{F}$ is a classifier trained to predict whether $x_i$ is NA. Determining what constitutes an NA response can be challenging and varies depending on the goals of the study and desired level of data quality. We

define an NA as any response that is unusable as data for the creativity task in question, as judged by human raters examining $x_i$. Specifically, such responses are rated NA by the minimum number of raters needed to reach the agreement threshold, which in our case is 60% for the AUT and 80% for the DPT. We purposefully include the prompt when training models because we theorize this classification will benefit from context: what is invalid for one prompt in one task could be perfectly valid for another. Note that this task will generally be imbalanced; because many survey platforms have some degree of participant quality checks built in, most responses can be assumed to be valid.

***2.1.4. Models Used*** We explore how well models of varying complexity perform at detecting NA responses to each task, pulling from categories of increasing levels of model complexity.

**2.1.4.1 Length Only** Prior work has often used response length as a heuristic for NA (Yeung & Fernandes, 2022). We include as a baseline a logistic regression model that predicts validity based on response length alone, where we calculate length as the number of tokens in the response extracted using the NTLK word tokenizer. We use the sci-kit learn implementation of this model (Pedregosa et al., 2011).

**2.1.4.2 TF-IDF Classifiers** We explore a suite of models trained using TF-IDF embeddings of each prompt-response pair, and include all models used by (Yeung & Fernandes, 2022) as well as feedforward neural networks. We use Sci-kit Learn to train both TF-IDF embeddings and the classifiers, and we perform hyperparameter optimization using Optuna (Akiba et al., 2019). We implemented all neural networks using PyTorch. For each model, we first performed hyperparameter optimization using validation set for a maximum of 100 hyperparameter combinations, which included relevant hyperparameters for both the TF-IDF

embedding model and the classifier trained using these embeddings. The hyperparameter grid for every model can be found in the code released with the supplementary materials. During training, we first constructed a TF-IDF embedding matrix for each word in the respective training set. This involved creating a document-term matrix, where each training example was treated as a separate document, and the vector representation of each document was based on the counts of all the unique words the document contained normalized by their TF-IDF scores (the frequency of a word's occurrence in a document multiplied by its inverse document frequency across all documents). This created a vector space that we used to embed all examples across a given dataset, and we created separate vector spaces for the DPT and AUT. We then used these TF-IDF embeddings as features to train logistic regression, decision tree, support vector machine, random forest, naïve bayes, adaboost, and neural network models. For neural networks, we initialized the embedding layer of the model using the learned TF-IDF embeddings and allowed the network to continue finetuning the embeddings during training.

**2.1.4.3 Word Embedding Models** We use the same classifiers as the TF-IDF models but replace the TF-IDF embeddings with word embeddings extracted from a FastText embedding model (Bojanowski et al., 2017) implemented in SpaCy. Unlike TF-IDF, FastText provides word representations that are pre-trained using a large corpus of English language, specifically Wikipedia and an English web crawl. This means that FastText embeddings can represent a richer set of semantic information about words than is possible with TF-IDF. Our experimental setup is the same as for TF-IDF, with the critical difference being that FastText embeddings do not need to be fit to our training set before making predictions. Because this reduces the number of hyperparameters we need to optimize, we elected to reduce the number of Optuna steps to 50.

**2.1.4.4 Sentence Embedding Models** We again use the same classifiers as both TF-IDF and FastText, but this time using a transformer embedding model from the Sentence Transformers package (Reimers and Gurevych, 2019). Unlike prior approaches, this embedding model is designed to embed sentences rather than individual words, and Sentence Transformer models are found to perform well at clustering, semantic search, and related tasks involving multi-word queries. We use the all-MiniLM-L6-v2 model available on Hugging Face (Wolf et al., 2020), which functions as a stand-in replacement for FastText. All other experimental protocols are kept the same as in FastText.

**2.1.4.5 Finetuned Transformers** We finetune DistilBERT-base-uncased (Sanh et al., 2019), DeBERTa-v3-base (He et al., 2020), DeBERTa-v3-large, ModernBERT-base (Warner et al., 2024), ModernBERT-large, gemma-2-2b-it, gemma-2-9b-it, and gemma-2-27b-it. As training these models require significantly more time and computing than was the case for prior model types, we only performed limited hyperparameter optimization using the validation set with the DeBERTa and DistilBERT models and applied the optimal hyperparameters to all models across the remaining trials. We trained the gemma models using LoRA (Hu et al., 2022) with a rank of 64, an alpha of 128, a lora dropout of 0.05, and 8-bit quantization, selecting these hyperparameters based on the recommendations listed on Hugging Face. We used Wandb to log results, hyperparameters, and performance metrics. We add a separator token between the prompt and response, for all models that use this token.

**2.1.4.6 Few-Shot LLMs** We evaluate gemini-2.5-flash and gpt-5-nano by prompting the model with between 0 to 10 examples selected at random from the training set to classify all examples in the test sets. We sample a balanced number of valid and NA examples, using a seed to select the same shots for every trial. We set the temperature to 0 for gemini-2.5-flash and to 1

for gpt-5-nano (this model does not support a temperature of 0), and we keep all other text generation parameters at their defaults. We include the prompt used in the supplementary materials.

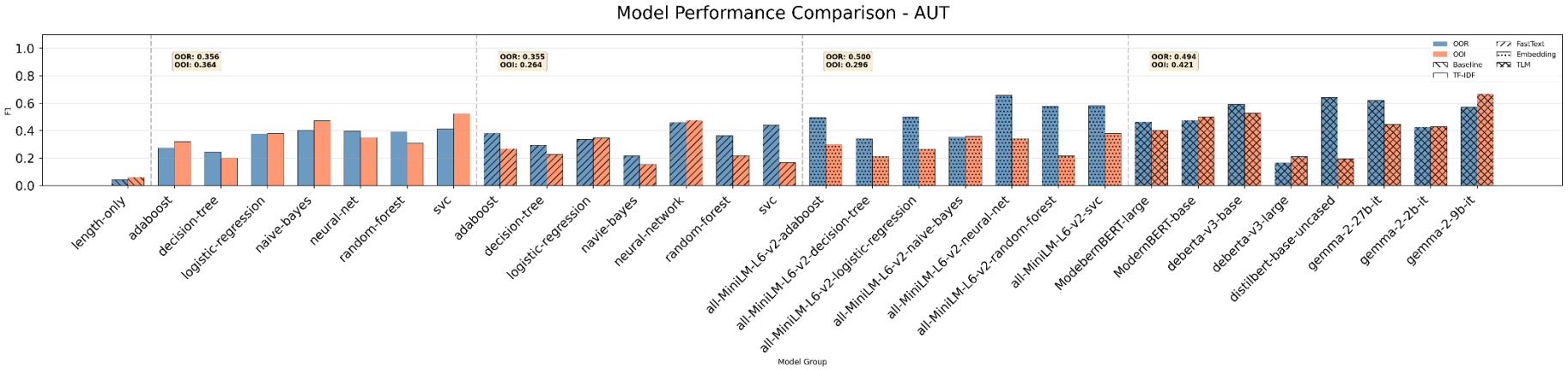
Finally, to further address data imbalance, we applied class weighting for all models to be inversely proportional to class frequencies. We trained all TLMs models using three RTX A6000 GPUs with 49GB of video memory each, except for few-shot models, which we access using the respective APIs.

**2.2. Results**

We begin by examining the precision, recall, and F1-score of models on the test sets. Precision is the ratio of the true positive predictions to all positive predictions and intuitively represents a model's ability to not predict a valid response as NA. Recall is instead the ratio of true positive predictions to the sum of true positive and false negative ones and is a measure of the model's ability to correctly classify all NA responses. The F1 score is the harmonic mean between precision and recall; it ranges between 0-1 and higher a higher F1 indicates a better model. Achieving a high F1 requires that models balance both precision and recall, making it a good summary metric for our analysis. We thus expect a performant model to achieve a high F1 score.

Figure 2 lists the F1 scores for all datasets and models (TF-IDF, FastText, sentence embedding, and TLMs, but excluding LLMs), including both results for individual models and the average F1 for each model type. Results for precision and recall are available in Appendix C. For the AUT, we find that models based on TLMs --- sentence embedding and finetuned TLMs --- generally outperform TF-IDF and FastText models on either the OOR or OOI set, in some cases both. While TLMs slightly underperform on the OOR set relative to sentence embedding

models, they achieve the highest average F1 on the OOI set of all model types. This isn't surprising, given that TLMs are pre-trained on large corpora prior to finetuning and tend to generalize well to unseen tasks. Not surprisingly, larger transformers tend to perform better than smaller ones, with gemma-2-9b achieving an F1 score above 0.6 on the OOI set. We see similar trends for the DPT, with finetuned TLMs again achieving better performance on the OOI set, though in this case the gulf in performance between TLMs and sentence embedding models is much larger for the OOR set. Importantly, performance is lower on average for the DPT than it was for the AUT, despite the fact that we set a more stringent agreement threshold for this task. Additional performance statistics broken down by item are listed in Appendix E. Briefly, we kept only items for which there was at least one valid and one NA response, and plotted performance metrics for each model type on every remaining item from the OOR set from both datasets, excluding LLMs. We observed significant disparities in performance as a function of item, with models achieving near perfect F1 scores for some items yet near zero for others, with this pattern holding across both tasks. Due to the very poor performance of the length only baseline, we chose not to include it in subsequent analyses.


Model Performance Comparison - AUT
F1
1.0
0.8
0.6
0.4
0.2
0.0
OOR: 0.356
OOI: 0.364
OOR: 0.355
OOI: 0.264
OOR: 0.500
OOI: 0.296
OOR: 0.494
OOI: 0.421
OOR
OOI
Baseline
TF-IDF
FastText
Embedding
TLM
length-only
adaboost
decision-tree
logistic-regression
naive-bayes
neural-net
random-forest
svc
adaboost
decision-tree
logistic-regression
navie-bayes
neural-network
random-forest
svc
all-MiniLM-L6-v2-adaboost
all-MiniLM-L6-v2-decision-tree
all-MiniLM-L6-v2-logistic-regression
all-MiniLM-L6-v2-naive-bayes
all-MiniLM-L6-v2-neural-net
all-MiniLM-L6-v2-random-forest
all-MiniLM-L6-v2-svc
ModebernBERT-large
ModernBERT-base
deberta-v3-base
deberta-v3-large
distilbert-base-uncased
gemma-2-27b-it
gemma-2-2b-it
gemma-2-9b-it
Model Group

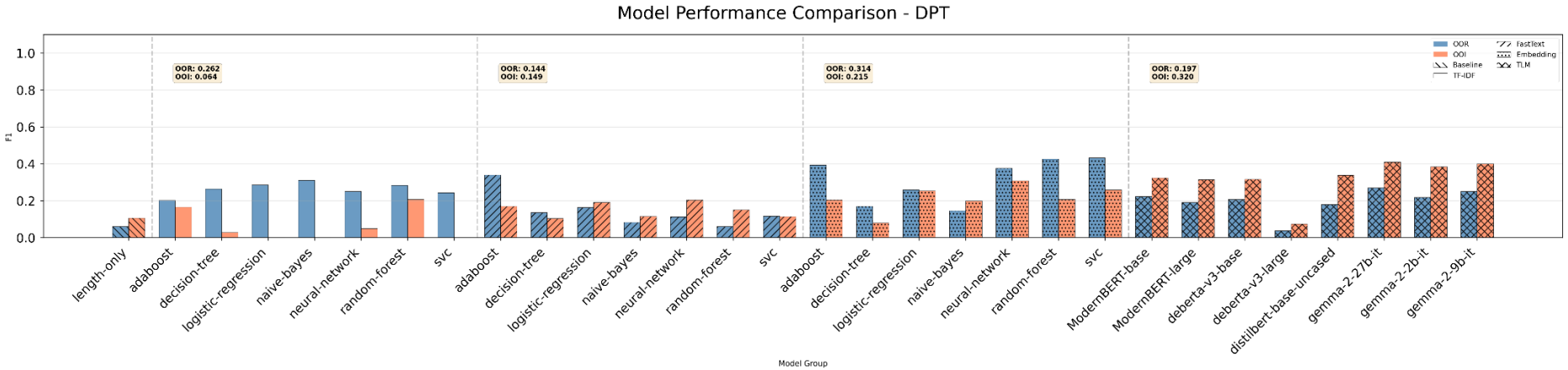

Model Performance Comparison - DPT
F1
1.0
0.8
0.6
0.4
0.2
0.0
OOR: 0.262
OOI: 0.064
OOR: 0.144
OOI: 0.149
OOR: 0.314
OOI: 0.215
OOR: 0.197
OOI: 0.320
OOR
OOI
Baseline
TF-IDF
FastText
Embedding
TLM
length-only
adaboost
decision-tree
logistic-regression
naive-bayes
neural-network
random-forest
svc
adaboost
decision-tree
logistic-regression
naive-bayes
neural-network
random-forest
svc
adaboost
decision-tree
logistic-regression
naive-bayes
neural-network
random-forest
svc
ModernBERT-base
ModernBERT-large
deberta-v3-base
deberta-v3-large
distilbert-base-uncased
gemma-2-27b-it
gemma-2-2b-it
gemma-2-9b-it
Model Group

Figure 2. F1 scores for all models on the OOR and OOI sets of the AUT (top) and DPT (bottom). Results are grouped by model type: TF-IDF, FastText (word embedding), sentence embedding, and TLM. Wheat colored boxes for each group are the average OOR and OOI F1 score for that group. For example, TLMs obtain an average OOR F1 on the DPT of 0.197.

We include few-shot LLM F1 scores in Figure 3, broken down by the number of shots. Additional results for precision and recall are listed in Appendix D. We find that gemini-2.5-flash is highly effective at identifying NAs on the AUT, with an OOI F1 scores of 0.5 and an OOR F1 of ~0.65 in the best performing trial, making it the best performing mode on the AUT overall. We find that gpt-5-nano performs much more poorly, and that increasing the number of shots has the surprising effect of reducing this model's performance for both sets. On the DPT, we again see weaker performance from both models, though gemini-2.5-flash again demonstrates stronger performance among the few-shot models. We emphasize that our analysis here is intentionally limited: due to budget constraints and the large number of existing models, we could not run an exhaustive analysis of LLMs to examine other important hyperparameters (prompt optimization, thinking mode, the effect of other text generation settings, etc.). Nevertheless, our results do indicate that few-shot models can at least match, if not surpass, the performance of their finetuned counterparts.

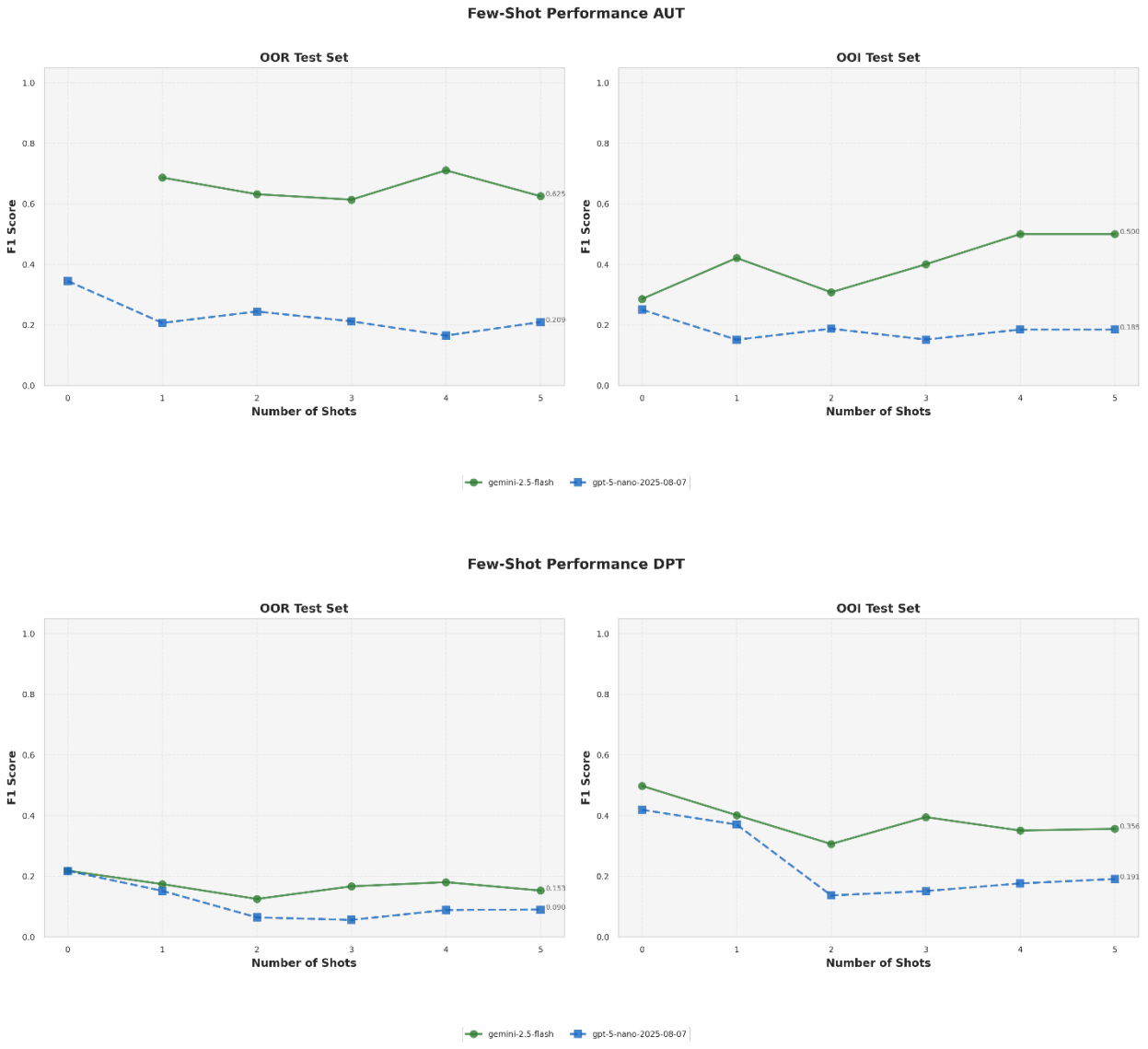



Figure 3. Few-shot F1 scores for the AUT (top) and DPT (bottom). Due to an error during data cleaning, zero shot results for gemini-2.5-flash for the AUT were not recorded. Each shot includes both a valid and invalid response. For example, 1 shot means that 1 valid and 1 invalid response were provided in the prompts.

To further analyze model performance, we include the Type I and Type II error rates for all model types (excluding few-shot models) in Figures 4, 5, 6, and 7. Type I errors are false positives (a response is predicted NA when it is valid) while Type II errors are false negatives (a response is predicted valid when it is NA). We again see a tradeoff between the sentence embedding and finetuned TLMs approaches: TLMs tend to have a very low Type I error rate, but

this comes at the expense of a higher Type II error rate relative to sentence embedding models. The test set again influences model performance: on the DPT, TLMs almost universally outperform sentence embedding models on the OOI set but comparatively perform much more poorly on the OOR set, though we do not see the same trends for the AUT. TF-IDF models consistently exhibit a high disparity in error rates on the OOI sets, which implies that this embedding approach failed to generalize to unseen items.

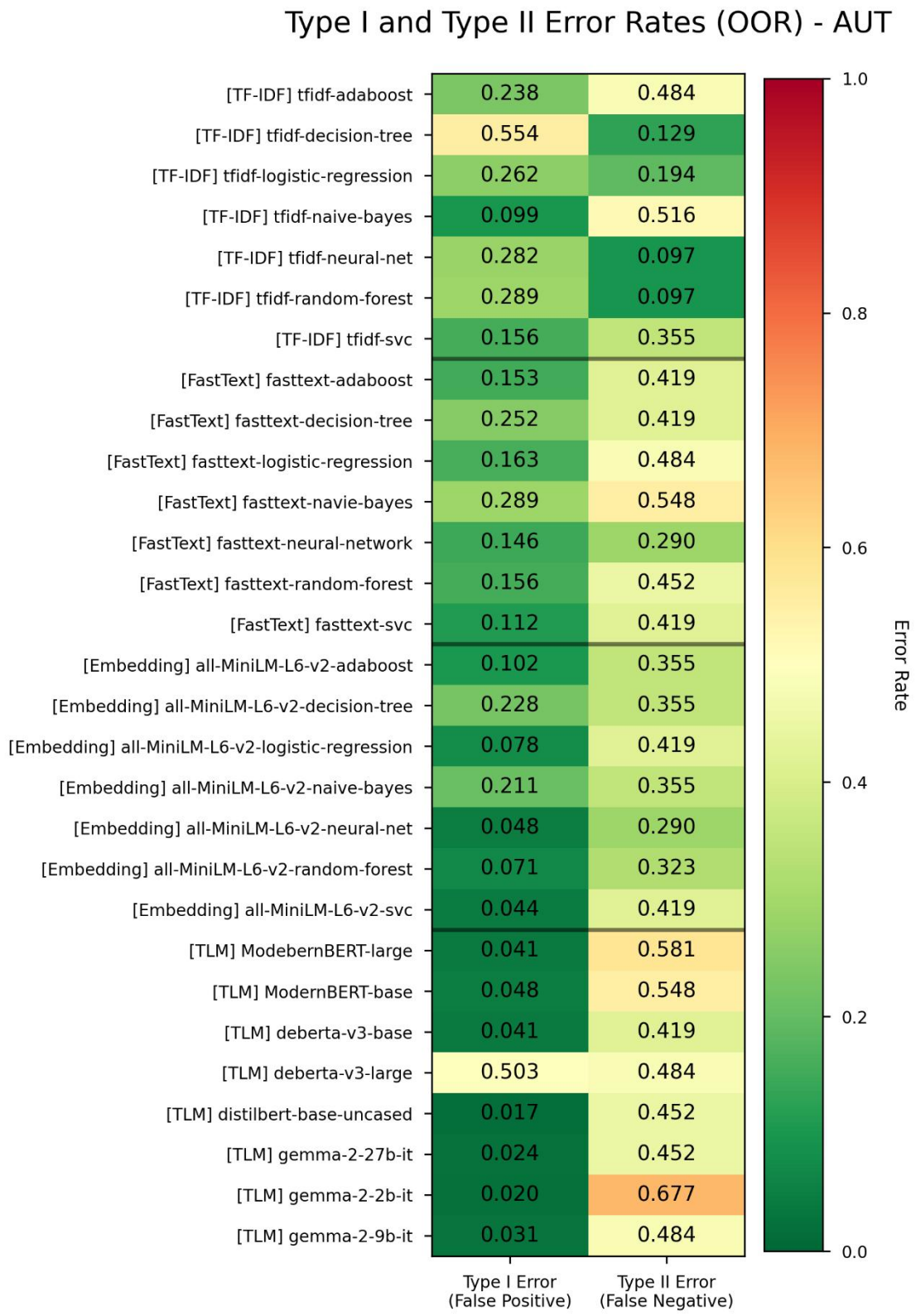



Figure 4. Type I and Type II errors broken down by model type for the AUT OOR set. Lower is better.

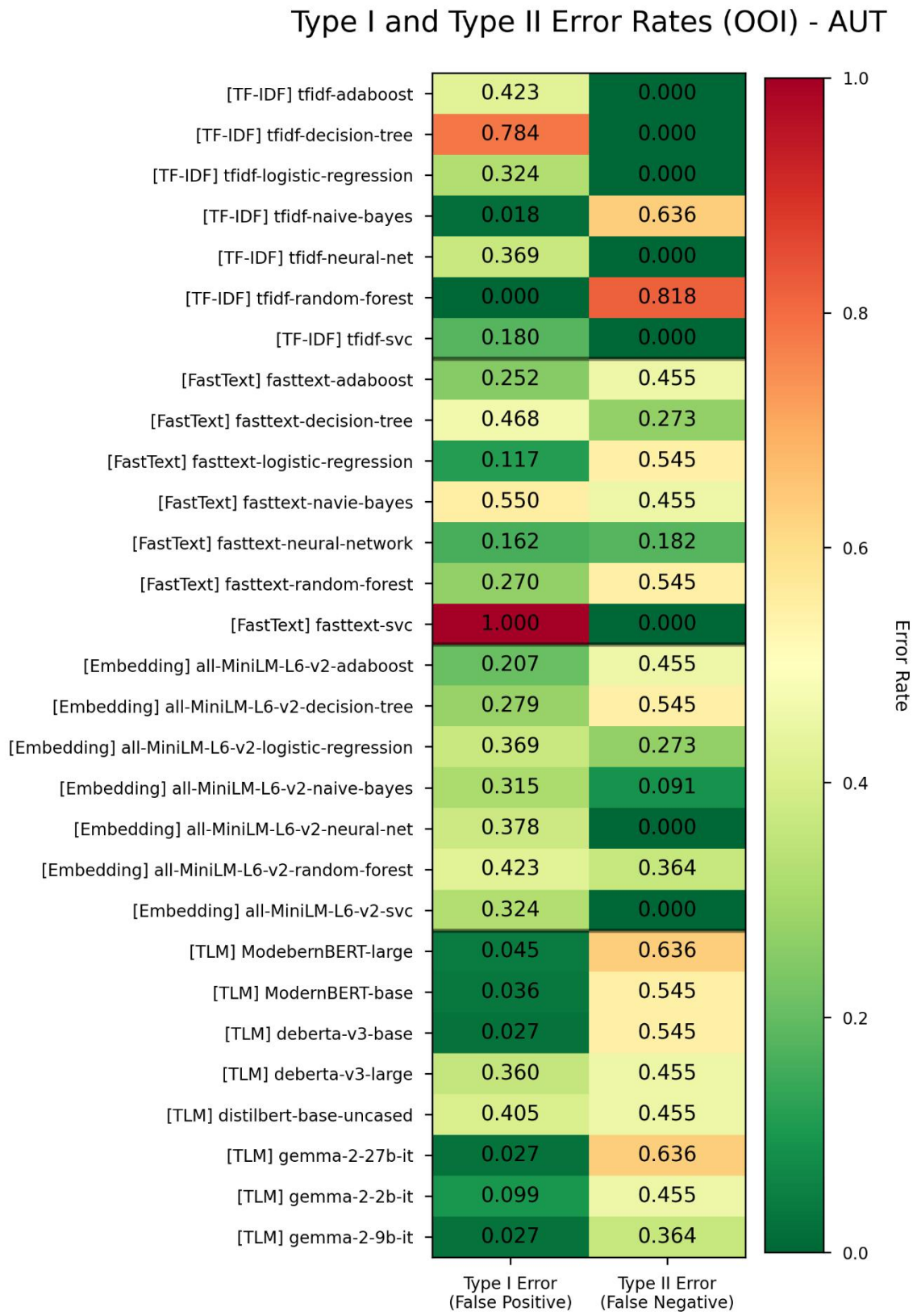



Figure 5. Type I and Type II error rates broken down by model type for the AUT OOI set.

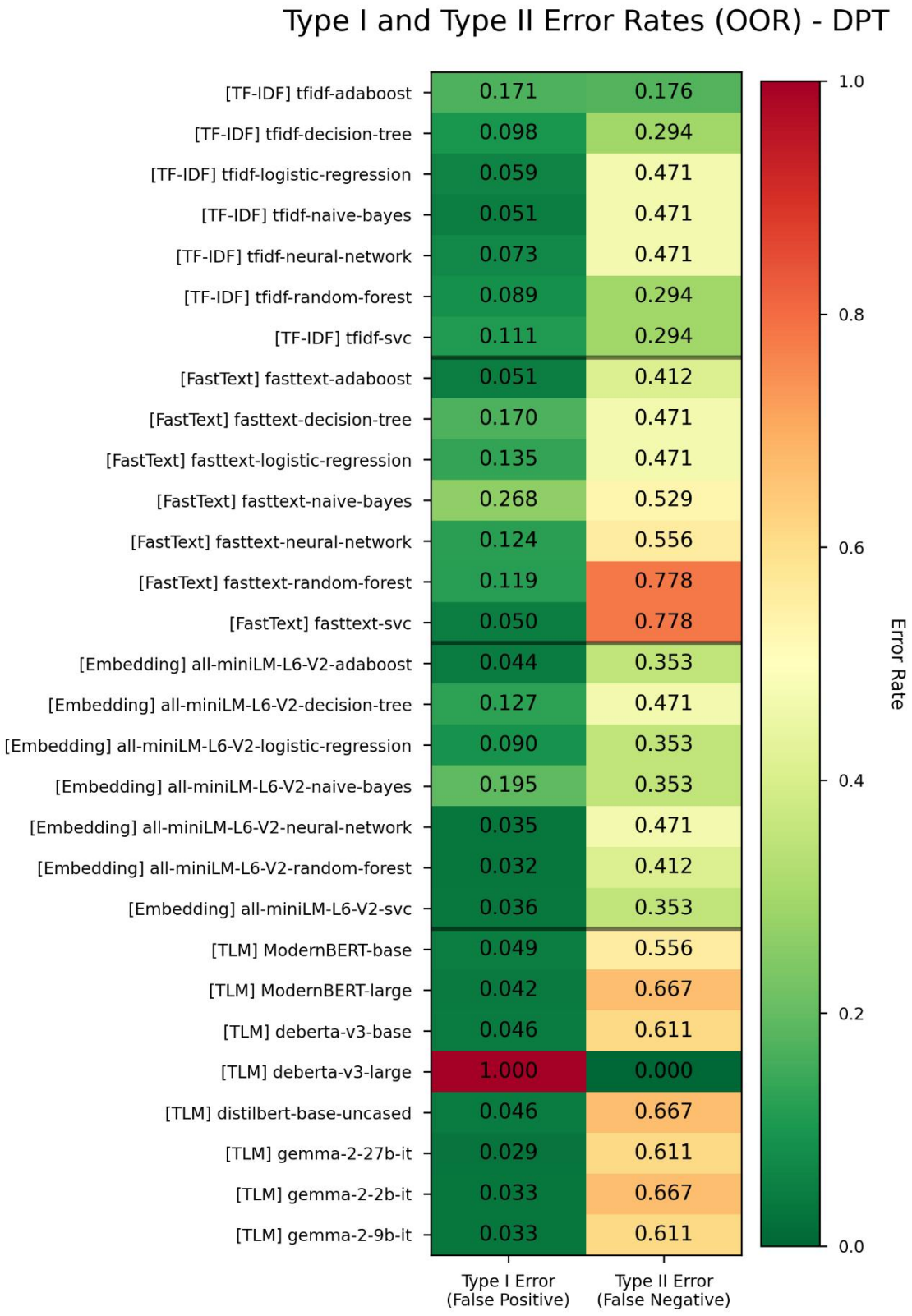



Figure 6. Type I and Type II error rates broken down by model type for the DPT OOR set.

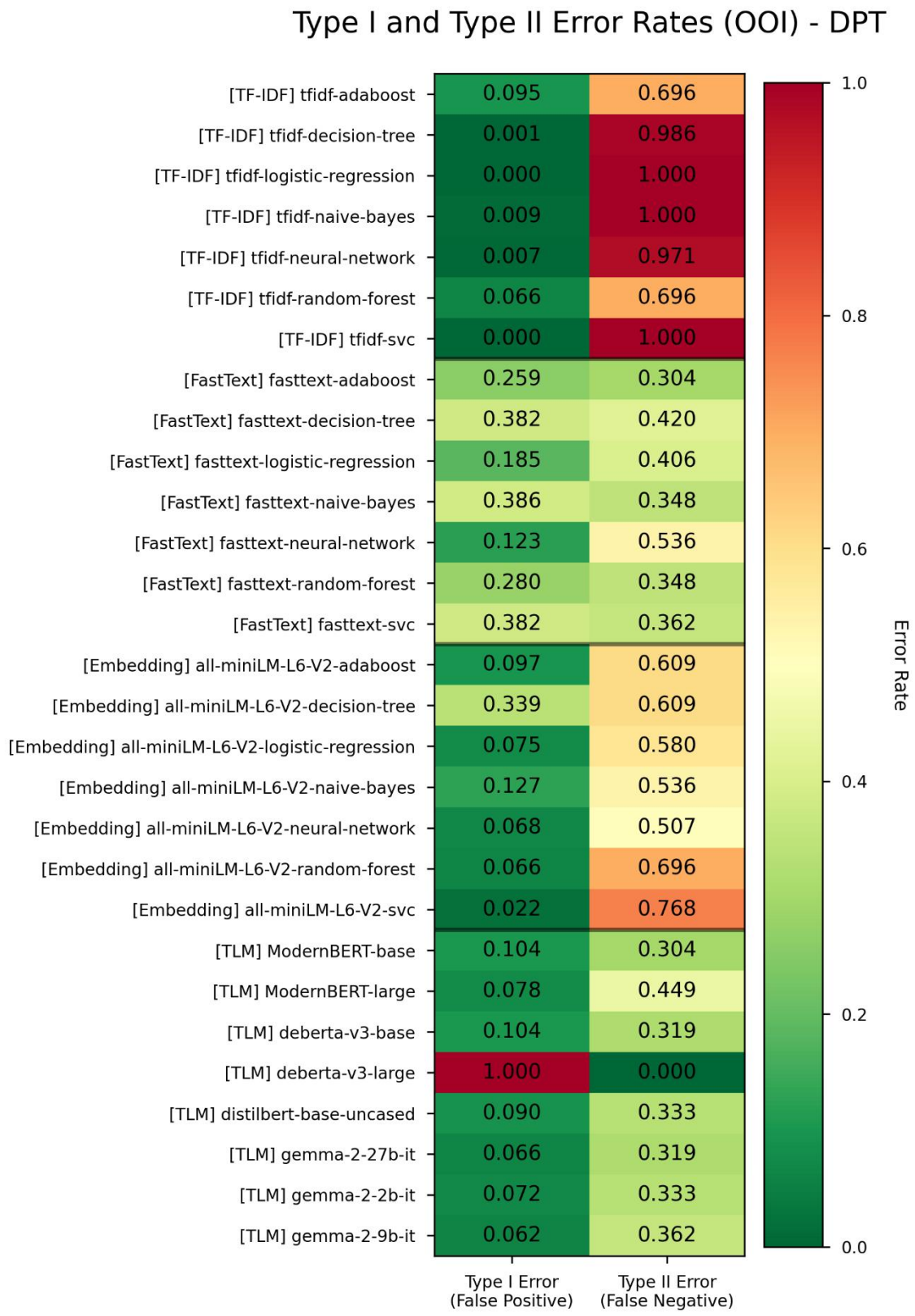



Figure 7. Type I and Type II error rates broken down by model type for the DPT OOI set.

### 2.3. Discussion

As expected, both TLMs and LLMs perform well at detecting NAs, exceeding the performance of most baselines. We found that they performed better on the AUT than they did for the DPT, which makes sense given that the AUT had more NA responses in the train set that weren't synthetic. The length-only classifier obtains the weakest performance of all models, which is in line with prior findings that word count alone is a poor predictor of NA responses (Yeung & Fernandes, 2022).

We found notable disparities in performance between the DPT and the AUT, with models universally performing worse in terms of F1 at predicting NAs on the DPT. This could be driven by differences in the difficulty of identifying NA responses on the DPT, as the prompts for this task are much longer and tend to elicit more elaborate responses from participants. This, coupled with the fact that the DPT emphasizes the practicality of solutions more than the AUT, could drive greater disagreements among raters on the validity of responses, lowering the ceiling of model performance. Alternatively, this might be an artifact of the more stringent agreement threshold we set for the DPT compared to the AUT. Since a higher threshold would reduce the number of NAs in the training set and increase the number of synthetic NAs that would need to be generated to correct for imbalance, it is possible that models failed to generalize as well due to over-reliance on synthetic NAs. However, TLMs still performed much better on average on the OOI set than any baseline, and they at least exceeded the length-only baseline for the OOR set, which implies that they still generalized beyond what is possible with shallow response heuristics. While one could further test model performance at different levels of rater agreement, we believe doing so would be ill-fated without significantly more data collection. First, the number of NAs was already quite small at 80% agreement, so an increase in the threshold

beyond this would likely to result in too few NAs to train a performant model. Second, given that raters seldom fully agree with each other on the validity of responses for divergent thinking assessments (Kucwaj and Kroczek, 2025), it is possible that this would be true for other types of creativity assessments, and reducing the agreement threshold could introduce significant noise in the dataset. Notably, raters reached only moderate agreement for DPT NAs, giving further credence to this possibility. While our models perform well compared to the baselines explored, it is likely that headroom remains in improving model performance, though achieving this may require more data than we presently have available.

Also of note is the sensitivity we observed in model performance for differences in item content, with certain items having markedly better detection accuracy than others. This was true across models and tasks and does not appear to be an artifact of any specific modeling approach. Importantly, we dropped any items for which there were no NA responses --- not all items in the test sets had NAs --- meaning that models could not obtain a trivially high score by simply predicting every response to be valid. We suspect this behavior is driven by differences in the support of each item in the data. Especially for the AUT, certain items have many more responses than others, and it may not be reasonable to expect models to perform equally well across all items given this constraint. Nevertheless, this finding points to an important limitation of our detectors which we leave to future work to explore in more depth.

## 3. Study 2: Analysis of Cross-Task Transfer

A *multitask* detector, trained to detect NAs across multiple assessments, could prove to be an invaluable resource for creativity researchers, especially when working with new tasks where there may not be data available to finetune new models. However, it is unclear whether such a model would achieve similar detection accuracy as models trained only on the task of

interest. This largely depends on how similar the operationalization of validity is in common between tasks; significant differences in this operationalization could make cross-task transfer difficult. We imagine this might be the case for our tasks. For the DPT, because more emphasis is placed on the quality of the response (Luchini et al., 2025), and because the solution must clearly address the problem, it is more likely that raters will label a response as invalid on the grounds of failing to comply with instructions (Ex. while building roads made of candy could be a possible solution to reduce traffic congestion in cities, it does not comply with the DPT's instruction to only consider solutions that are feasible). The AUT, on the other hand, tends to give participants much more liberty in the types of ideas they can generate (Ex. While using a brick as a shoe might be a questionable alternative use from the perspective of practicality, it is arguably still valid). Such differences in task instructions directly affect how raters handle NAs for these tasks, as the tasks have differing standards for the types of creative responses that are desired. Regardless, we are still interested in the extent to which TLMs can learn from a multitask mixture during training. In Study 2, we assessed this by training a TLM on both the DPT and AUT training sets, and comparing its performance to the same TLM trained on one task and assessed on the other task.

### 3.1. Methods

We use gemma-2-2b-it for all multi-task experiments and use the same hyperparameters as were used when training single-task models. Notably, the gemma series of models were trained via supervised finetuning on a large mixture of natural language processing tasks, covering topics as diverse as math, coding, and user conversations with chatbots. Since this was not the case for our other TLMs, and since the gemma models are also much larger, we reasoned that this model would be a good choice for this experiment. We first train a multitask variant of

gemma on the union of the training sets from both tasks and evaluate the multitask model on the OOR and OOI sets from both tasks. Using the previously trained single-task models, we evaluate on the OOR and OOI sets of whichever task the model was not trained on. Specifically, we evaluate the gemma-2-2b-it trained on the AUTs on the DPT test sets, and vice-versa for the model finetuned on the DPT. All other training details are the same as they were for the single-task experiments.

### 3.2. Results

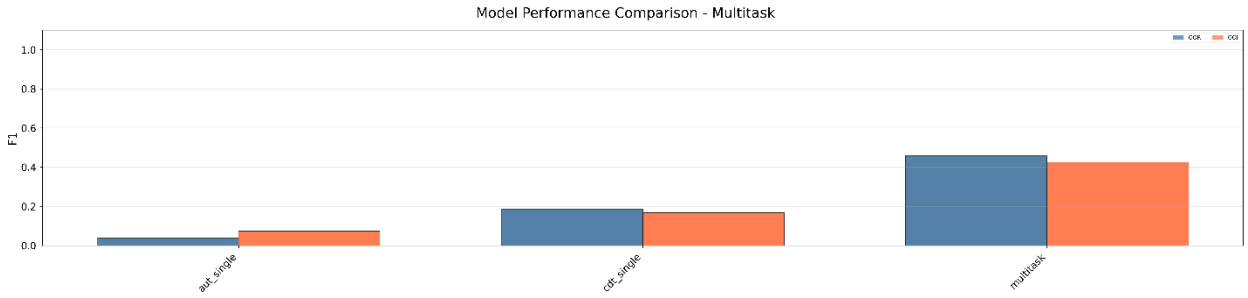



Figure 8. F1 results from multitask experiments. AUT and DPT single-task results are from whichever task the model was not trained on (Ex. for the AUT, the model is evaluated on the DPT).

F1 scores from the multitask trials are listed in Figure 8. Compared to the single task models evaluated cross task, the multitask model performs considerably better, achieving nearly double the F1 score. However, we note that the multitask model's performance is on par with the single task model evaluated on the *same* task it was trained ont: gemma-2-2b achieves and F1 score of about 0.4 for both test sets on the AUT and for the OOI set of the DPT. This implies that, while a single task model may not generalize to unseen tasks, a multitask model does not necessarily perform better when data is available to finetune a model on the task in question.

### 3.3. Discussion

We find that single task TLMs do not generalize well to unseen tasks, and that the multitask model does so much more effectively. The relationship between task transfer is not

symmetric: DPT TLMs generalize better on the AUT than AUT TLMs do for the DPT, though this implies that there must be some features of NA responses common to both tasks. While gemma-2-2b-it may outperform single task models evaluated out of domain, evidence is mixed on its performance when evaluated against within-domain single task models. Although the multitask model does not perform significantly worse, it also does not perform significantly better, making this modeling setup hard to justify given that it takes more time to train with the added data. Clearly our analysis is limited by the fact that we evaluated only one model, and that we only have two tasks to evaluate with. While other models could fare better, since gemma was among the best performing finetuned models, we suspect that results using this model are likely an upper bound on the performance we would expect from a multitask model. Another approach for improving multitask performance would be to include data from additional creativity tests, though that would require additional data collection that is out of scope for this paper. Given that single-task TLMs already outperform baselines, and are the main focus of this paper, we chose not to include multitask models in additional experiments.

## 4. Study 3: Are Models Well Calibrated NA Detectors?

In practice, creativity researchers are often concerned with the quality or effectiveness of a response and not just its originality. Indeed, the standard definition of creativity states that products must be both novel and effective (or useful) for them to be creative (Runco and Jaeger, 2012). Given the importance of quality, it is interesting to consider how quality relates to response validity, as this would inform if validity should be considered a separate construct from quality. One way to explore this is to test if a NA detector is *calibrated* with respect to a quality rating, such that low-quality responses are assigned a lower probability of being valid than

responses that have a higher quality. Intuitively it appears that this might be the case: if raters give a response a quality score of 1 (on a 5-point Likert scale), it could be that the response did not follow task instructions or contained gibberish, since a response with those characteristics would not be useful for the creativity task. However, quality as defined in the standard definition does not necessarily encompass the same considerations as validity; it is generally assumed that participants at least followed the instructions provided in the task (Runco and Jaeger, 2012). From this perspective, an NA detector might not be calibrated with respect to quality because assessing validity encompasses a different set of criteria. With these considerations in mind, we designed our third study to test if our best-performing models from each type were calibrated with respect quality, as assessed by human raters.

### 4.1. Methods

We employed a separate calibration set for the DPT held out from all folds of training and testing. We built this set from DPT responses collected as part of a pre-registered study on the effect of AI feedback on human creativity (de Chantal et al., 2025). We scored the originality of all the responses collected for this pre-registration using the Creativity Assessment Platform (CAP), which provides validated originality scoring models (Patterson et al., 2025). Specifically, the CAP uses a RoBERTa-base (Liu et al., 2019) model trained on 8477 DPT responses with originality scores from 20 expert raters that were collected using a missing data design. Each rater was assigned approximately 2000 responses, with each response receiving five ratings. This model was found to strongly predict human creativity ratings and was deemed suitable for our analysis. We selected 100 responses at random from the bottom 10% of originality scores in the calibration set, and 100 from the top 90%, giving us a diverse array of

responses. As this data was collected as part of a separate study from that used to create the training set, no model was trained on any responses from the calibration set.

We then had three research assistants rate the responses for quality. Raters would first determine if a response was valid or invalid. If the response was deemed valid (using a 66% agreement threshold), raters then rated the quality of the response using a five-point Likert scale, further details on rater instructions are provided in Appendix F. We obtained an ICC2k of 0.90 for the quality scores of valid responses and 0.78 for agreement on NAs, indicating good rater agreement. Finally, we extracted the model predicted probability of valid for each label, doing so for the best performing TF-IDF, FastText, and sentence embedding models, gemma-2 2b and 9b, and for all LLMs (zero shot). Importantly, our LLMs do not provide access to the probabilities of each class, so we elected to assign a probability of 0 if the response was classified as NA and 1 otherwise.

### 4.2. Results

We plot the probability distributions for each model in Figure 9. We see a clear trend where lower quality responses are assigned to increasingly lower probabilities of being valid, with the only exception to this trend being the word embedding (FastText) model. A Kruskal-Wallis H-test (Sidney, 1957) revealed a significant difference in median probability of valid for each quality group for gpt-5-nano ($H=39.91$, $p < 0.001$), gemini-2.5-flash ($H=88.42$, $p < 0.001$), gemma-2-2b ($H=65.81$, $p < 0.001$), gemma-2-9b ($H=35.28$, $p < 0.001$), the TF-IDF random forest ($H=82.24$, $p < 0.001$), the FastText neural network ($H=38.72$, $p < 0.001$) and the sentence embedding neural network ($H=12.31$, $p < 0.05$), indicating that these differences in probability scores were significant.

### 4.3. Discussion

Our analysis indicates that most models are calibrated with respect to response quality: responses that are lower in quality are assigned to lower probabilities of being valid, which algins with our expectations. The only exception to this trend was for the FastText model, where invalid responses were assigned a *higher* probability of valid than even those with a quality rating of 5. We suspect that this driven by the poor performance of this model when generalizing to unseen responses, as the model misclassified most of the NA in the calibration set, and obtained a low OOR F1 score relative to the other model types in earlier trials. The strength of this calibration effect also appears to increase with model complexity: for TF-IDF models, NAs are on average assigned a 60% probability of valid, whereas for gemma models the probability is near 0%. Importantly, due to the relatively small size of the calibration set, only 7 of the 100 responses were assigned to be invalid by our raters, which could bias our results if these responses had similar characteristics (e.g., they were invalid for the same reasons). We note, however, that the calibration effect remains even when excluding NAs from consideration, as the probability of being valid for quality scores of 2 or 1 is consistently lower than it is for a quality score of 5. Collectively, our results point to the existence of a relationship between quality and validity, though further study is needed to fully characterize the strength of this relationship across creativity tasks and different rating protocols.

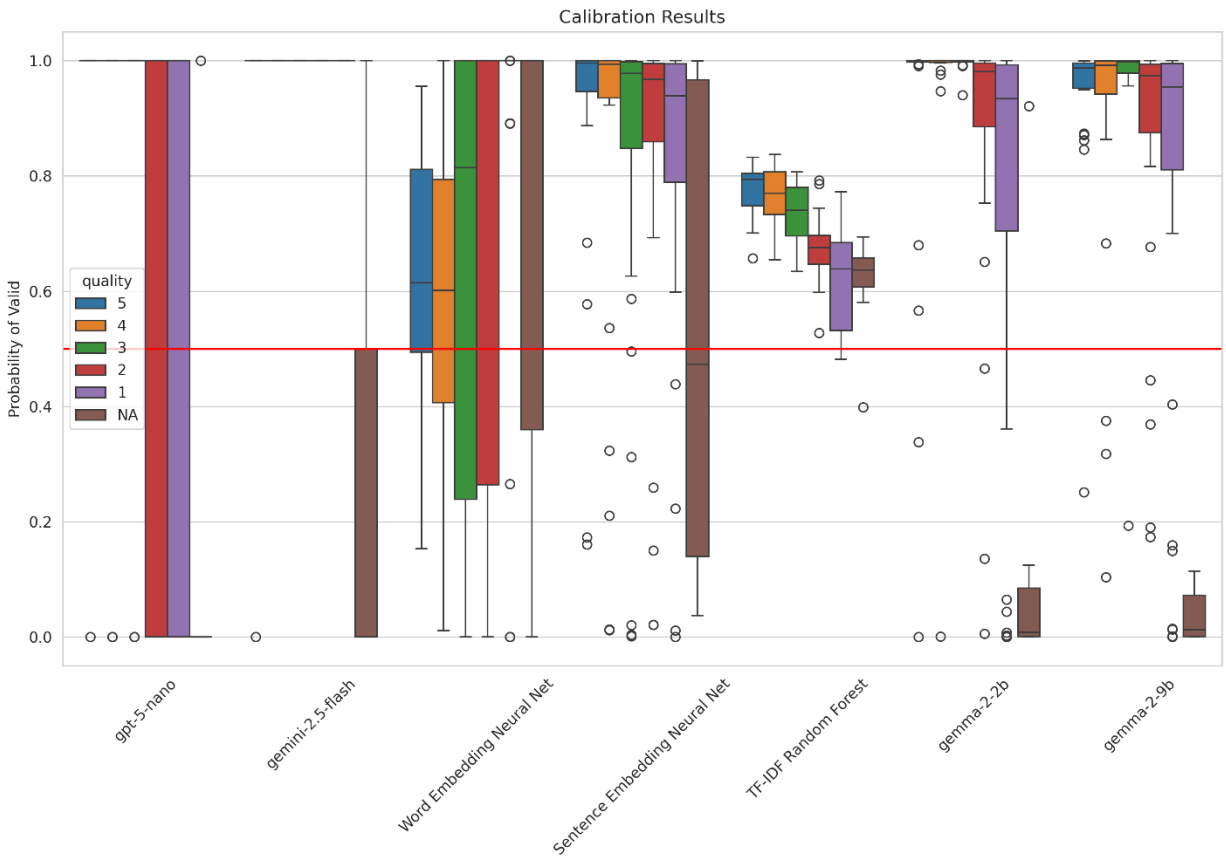



Figure 9. The model derived probabilities of the response being valid for each quality level. For quality levels 1-5, we use calibration examples which 1) all raters agreed were valid, and 2) at least 66% of raters agreed on the chosen quality score. For NAs, we required that all raters agree the response was invalid. The probability of invalid is 1 – probability of valid. Note that few-shot LLMs (gemini and gpt-5) do not output true probabilities, and the distributions shown here are instead pseudo-probabilities constructed based on the predicted label.

## 5. Study 4: Do Models Penalize Non-Native English Speakers?

For a detector to be deployed in actual creativity studies, we must also have confidence that it is fair with respect to protected characteristics (Weinberg, 2022). As all our tasks are linguistic, we are specifically interested in fairness with respect to language status, and whether our TLMs unfairly penalize L2 English speakers as opposed to L1. We reason that L2 speakers, by virtue of not being English natives, may exhibit different semantics in their responses to creativity tasks, which our models may unfairly classify at NA at a higher frequency than

English speakers. Our final study sought to test this in a controlled setting, where we ask L2 and L1 speakers to deliberately produce NA responses to the DPT. If models are fair in their predictions, and both groups are equally competent at fooling the model, then we expect the error rates for the model to be roughly the same in both conditions. Our choice to examine this in an adversarial setting is driven by practical considerations. Recall that NAs rarely occur in datasets, they make up less than 8% of both of our training sets. While a fairness analysis would ideally call for participants to complete the DPT in a naturalistic setting, it is reasonable to assume that the NA rate would be similar in this setting as it was for our archival data. This implies that we would need to collect more than 1,000 responses to obtain even 100 NAs, which we did not have the resources for. By asking participants to deliberately produce NAs, we can increase the rate of NAs collected dramatically, even if it comes at some expense to ecological validity.

### 5.1. Methods

We recruited 205 participants (113 English natives, 92 Spanish-English bilinguals with English as their second language) on Prolific to complete the DPT. We used an adversarial protocol where participants were instructed to write NA solutions to a DPT prompt. Participants were asked to respond to the item "Develop as many design ideas as you can to improve the safety of pedestrian crossings." and they provided a single response to the item. Participants were instructed that their response would be scored by an AI, and that they should try to "fool" the model by coming up with an NA response that would be misclassified as valid, but they were not given any information about the model itself or the scoring process. Prior to completing the experiment, participants were provided detailed instructions, as well as examples of five NA responses to ensure they understood the task: (1) "idk"; (2) "_qwefnqw qwf qw" (3) "brick" (4) "Once upon a time." (5) "Traffic lake still". Additionally, we asked participants if they used AI

during the survey and administered them the short version of the Language History Questionnaire (Li et al., 2020), which we used to confirm their language status. After dropping participants who ran out of time, reported use of generative AI, or reported that Spanish was not their native language on the LHQ despite reporting as such on Prolific, we obtained a final sample of 96 participants (74 English native and 22 Spanish-English bilingual). The study was approved by the local institution's IRB. We then scored each response using the same set of models from the calibration study to obtain both the predicted NAs.

### 5.2. Results

We report in Figure 10 the type II error rates for all models for both the reference (monolingual English) and focal (bilingual) groups. Additionally, we report the predicted positive condition rate (PPCR), which is simply the fraction of positive (NA) predictions. Non-TLMs appear brittle to adversarial responses under this condition, with the sentence embedding model failing to classify even 40% of the responses as valid, and the TF-IDF random forest failing to classify any response as NA. All TLMs perform comparatively much better, obtaining a PPCR of at least 0.5, with gemma-2-2b obtaining a PPCR above 0.8 for both groups. This provides evidence that our TLMs can flag responses as invalid even in this test of far generalization, with responses that are unlikely to occur in typical creativity datasets. However, we also observe stronger disparities in the PPCR and error rates for TLMs as opposed to non-TLM models, with gemini-2.5-flash having about double the Type II error rate for the reference group as compared to the focal group. Pairwise t-tests revealed that the difference in Type II error rates was significant for gemini-2.5-flash ($t=-2.97$, $p < 0.01$), but not for any other TLM ($p > 0.05$ for all models). This implies that differences in the number of NA predictions between the groups is significant only for gemini.

### 5.3. Discussion

In summary, we find that TLMs may classify more L2 responses as invalid than responses from English natives, given the higher PPCR for the focal group across nearly all TLMs. Gemma-2-2b is the only TLM that does not obey this trend, with the reference group obtaining a slightly higher PPCR. Statistical testing revealed that these differences were generally not significant, though because participants responded to only one item, we did not have enough data to perform a Mantel-Haenszel test or other type of fairness analysis. Furthermore, TLMs also performed universally better at detecting NAs in this experiment, obtaining significantly lower error rates than all non-TLM models.

What could be driving this disparity? First, we note that the sample size for the L1s was much larger than it was for the L2s, which was driven both by the difficulty of recruiting bilinguals online and the fact that many were found to not truly be bilingual after completing the LHQ --- perhaps only identifying as such on Prolific to quality for more studies. This difference in sample size could have been problematic for this analysis, it is less likely that the L2 sample is representative of the general L2 population than was the case for the L1s, which would impact any claims made on the fairness of our models. It is also plausible that the L2s were simply better at fooling the models than the L1s, in which case these results reflect genuine differences in ability rather than a source of bias. We note that our L2 sample was somewhat better educated, with their highest degree attainment being a doctorate (compared to a master's for the L1 sample), and with more L2s reporting a graduate degree or higher than the L1s. This disparity in education could also have been the source of the difference in detection accuracy, though it remains unclear whether education alone could have influenced performance on this task to such an extent. English and Spanish were chosen as the focus of this analysis due to their prevalence

in the USA, making the recruitment of bilinguals in both languages easier compared bilinguals for other language pairs. While we find no strong evidence of bias in our models for this important population, such fairness analysis should continue to be performed in future work, especially if our detectors are applied to L2 English speakers for other languages.

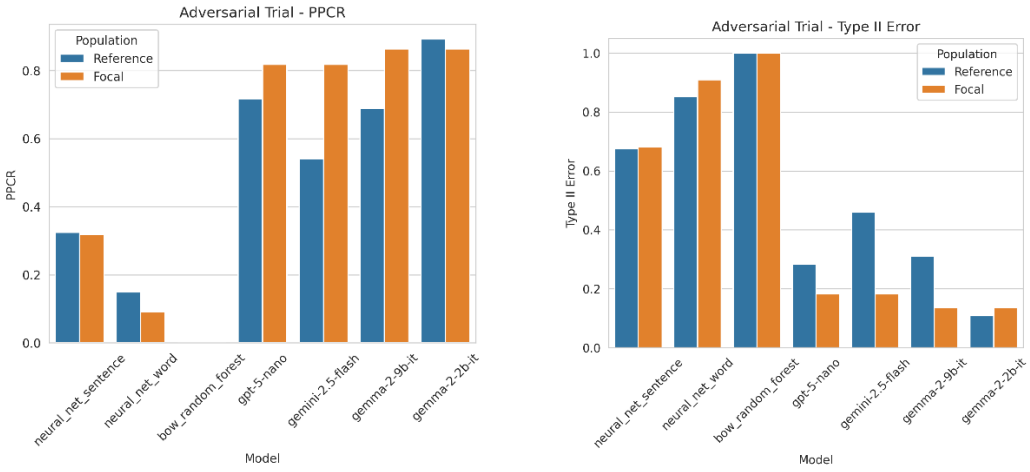



Figure 10. Type II error rate (right) and PPCR (left) for each model on the adversarial task. The reference group is English native speakers, and the focal is English-Spanish bilinguals with English as their L2 language. PPCR = predicted positive condition rate.

## 6. General Discussion

Collectively, our findings indicate that TLMs, be they finetuned or few-shot, generally provide better performance than models that use length heuristics or various other types of features, without exhibiting strong bias towards underrepresented populations. This is in line with past work which showed that TLMs can effectively detect invalid free responses in other domains (Cibelli Hibben et al., 2024), as well as more recent work that has explored invalid detection for divergent thinking tasks (Kucwaj and Kroczek, 2025). We found that the F1 score of TLMs matched or exceeded that of other models across most tasks, though there remained few cases where they exceeded 0.6. It appears that there are many invalids that TLMs can

consistently detect accurately, especially for the AUT, but also for the DPT to a lesser extent. Because TLM probabilities appear well calibrated with respect to response quality, the type II error rate can be further reduced by reassigning predicted but low confidence valids to NA, albeit at the cost of more annotator effort to correct type I errors.

The rich contextual language representations offered by TLMs also gives them the advantage of strong transfer learning capabilities. We found this capability held for our tasks as well, with gemma-2-2b-it outperforming single task models evaluated on unseen tasks, when trained on the multitask training set. However, we also found that this multitask model did not perform significantly better than single task models evaluated on the same task they were trained on, calling into question the utility of transfer learning approaches when training data is available for the creativity task of interest. It is conceivable that the multitask model would be a good choice for completely new tasks for which no NA data exists, though we leave it up to future work to test this. Given that we were limited to only two tasks, transfer learning performance can likely increase by training on a more diverse task mixture, though we leave it to future work to curate such a dataset. Our general recommendation for creativity researchers seeking to apply NA detection for tasks not studied in this paper would be to start with few-shot models, as they are the most likely to perform well zero-shot and are the easiest to update (since this only requires changing the prompt and hyperparameters). Should few shot models fail to perform well, our multitask model could be a suitable choice, though we caution against assuming that our model will perform well on tasks we did not evaluate.

While we hope that our models will be useful for data cleaning, we stress that they do not completely obviate the need for manual analysis. Model predictions should never be used as the sole basis to reject a participant's submission, as we intend our work to serve only as a first-pass

filter to prevent the need for exhaustively rating all responses. Additionally, we must emphasize that whether a response is considered NA can be affected by cultural factors that often result in substantial rater disagreement. While we mitigated this using a conservative agreement threshold, future creativity research must examine these dynamics more carefully to ensure NA detectors do not make unfair classifications. Our goal in this work is not to provide a definitive set of standards for identifying all the ways in which a response can be invalid, nor do we intend for researchers to blindly use our models during data cleaning. As discussed in Kucwaj and Kroczek (2025), checking for invalid responses is a necessary but woefully underreported step in data preparation for creativity studies. By developing and reporting on the performance of a broad set of machine learning models for this task, we hope to both draw attention to methodological considerations for this stage of data preparation in creativity studies, while also making the process of screening for invalids both more objective and less labor-intensive. We also note that our models do not replace other aspects of data quality control for online surveys, including IP address filtering and attention checks. It is conceivable that our models could be used as part of a larger ensemble for detecting suspicious participants, especially for mixed methods surveys that use both multiple-choice and free-response items, but we leave this to future work to explore. Similarly, while a response created using generative AI is NA if the participant was instructed not to do this, and an NA detector could coincidentally detect AI responses if it was trained on ones labeled as NA, we emphasize that NA detectors are not AI detectors and should not be used for this purpose. A comprehensive approach to survey data quality must include multiple flags for different types of suspicious behavior, which we believe our models help contribute to.

NA detection for free responses remains an open problem, which as Cibelli Hibben et al. (2024) notes is still in its infancy. Performance improvements for TLM models will require invalid data for a broader range of creativity tasks, more NA data collection for the tasks already explored, and the creation of a comprehensive set of methodological standards for how such responses are identified and screened for. Indeed, free responses can encompass other modalities, such as visual in the case of drawing tasks or auditory in the case of music production, yet we know of no work that has built detectors for non-linguistic assessments. Creativity is unique in that it can be studied and scored across modalities, making NA detectors for non-linguistic tasks an interesting avenue to explore.

Finally, an important theoretical question is how response validity is best operationalized in relation to the measurement scale. Our experiments have treated NAs as being separate from the construct measured: not high or low in creativity but something else to first be filtered out. Another possibility would be to place NAs on the same measurement scale as valid responses, and to then assign them the lowest possible score. This approach would have the added benefit of allowing all responses to be used in training a scoring model, possibly increasing model performance. The results from our calibration study suggest that this is feasible, given that many models already appear to predict lower quality responses as more likely to be NA. We argue that keeping NAs separate from the target scale aids in interpretability by making the data cleaning process more explainable, but future work should test competing operationalizations to evaluate other possible approaches for NA detection.

## 7. Conclusion

Developing easy to use and effective methods for detecting invalid free responses is becoming increasingly important for creativity researchers as more human studies are run online. The convenience of these platforms also brings with them added challenges for ensuring only high-quality data is collected, a problem worsened by the rapid adoption of generative AI. Yet relying on manual inspection alone is quickly becoming impractical, necessitating that researchers turn to automated solutions. In this paper, we have sought to fill this need by developing new models for detecting invalid free responses for creativity tasks. Our models are open source, based on easy-to-use libraries, and require minimal resources for inference mode, making them a viable option for most researchers.

## References


Aftan, S., & Shah, H. (2023). A survey on bert and its applications. *20th Learning and Technology Conference (L&T).* IEEE.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology, 33*, 445-459.

Barends, A. J., & De Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and individual differences, 143*, 84-89.

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the left-right scale a valid measure of ideology? Individual-level variation in associations with "left" and "right" and left-right self-placement. *Political Behaviour, 39*, 553-583.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media. https://doi.org/9780596516499

Boschini, M., Bonicelli, L., Porrello, A., Giovanni, B., Pennisi, M., Palazzo, S., Spampinato, C., & Calderara, S. (2022). Transfer Without Forgetting. *ECCV.* Springer Nature.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.

Brown, T., Mann, B., Ryder, N., Subbiah, M., ..., & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems.* 1877-1901: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology, 2*. https://doi.org/https://doi.org/10.1016/j.metip.2020.100022

de Chantal, P. L., Beaty, R., Laverghetta, A., Pronchick, J., Patterson, J., Organisciak, P., ... & Karwowski, M. (2025). Artificial intelligence enhances human creativity through real-time evaluative feedback.

Chmielweski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science, 11*(4), 464-473.

Cibelli Hibben, K., Smith, Z., Rogers, B., Ryan, V., Scanlon, P., & Hoppe, T. (2024). Semi-Automated Nonresponse Detection for Open-Text Survey Data. *Social Science Computer Review*. https://doi.org/https://doi.org/10.1177/08944393241249720

Delle Fave, A., Drdar, I., Freire, T., Vella-Brodick, D., & Wissing, M. P. (2011). Delle Fave, A., Brdar, I., Freire, T., Vella-Brodrick,The eudaimonic and hedonic components of

happiness: Qualitative and quantitative findings. *Social indicators research, 100*, 185-207.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics.

DiStefano, P. V., Patterson, J. D., & Beaty, R. E. (2024). Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, 1-15.

Etz, R. S., G. M., Eden, A. R., & Winship, J. (2018). Rapid sense making: A feasible, efficient approach for analyzing large data sets of open-ended comments. *International journal of qualitative methods, 17*(1).

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behaviour research Methods, 54*, 1-20.

Forthman, B., Goecke, B., & Beaty, R. E. (2023). Planning missing data designs for human ratings in creativity research: A practical guide. *Creativity Research Journal*, 1-12.

Gangnon, V., Labrie, A., Bhatnagar, S., & Desmarais, M. C. (2019). Filtering non-relevant short answers in peer learning applications. *Educational Data Mining.*

Gogami, M., Matsuda, Y., Arakawa, Y., & Yasumoto, K. (2021). Detection of careless responses in online surveys using answering behavior on smartphone. *IEEE Access , 9*, 53205-53218.

Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior, 1*(1), 3-14.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint*. https://doi.org/arXiv:2006.03654

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Huang, J. L., Curran, P. G., Keeny, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99-114.

Iiagan, M. J., & Falk, C. F. (2024). Model-agnostic unsupervised detection of bots in a Likert-type questionnaire. *Behaviour Research Methods, 56*(5), 5068-5085.

Kaczmirek, L., Meitinger, K., & Behr, D. (2017). Higher data quality in web probing with EvalAnswer: a tool for identifying and reducing nonresponse in openended questions.

Kucwaj, H., & Kroczek, B. (2025). ChatGPT as a Competent Enough Judge in Validating Responses from a Divergent Thinking Task. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 47).

Laird, R. D., Marks, L. D., & Marrero, M. D. (2011). Religiosity, self-control, and antisocial behavior: Religiosity as a promotive and protective factor. *Journal of Applied Developmental Psychology, 32*(2), 78-85.

Laverghetta Jr., A., Luchini, S., Linnell, A., Reiter-Palmon, R., & Beaty, R. (2024). he creative psychometric item generator: a framework for item generation and validation using large

language models. *CREAI 2024: Workshop on Artificial Intelligence and Creativity.* Santiago de Compostela (Spain): CEUR Workshop Proceedings.

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods, 13*(3), 229-248.

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition, 23*(5), 938-944. https://doi.org/doi:10.1017/S1366728918001153

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., ..., & Roberts, A. (2023). The flan collection: Designing data and methods for effective instruction tuning. *International Conference on Machine Learning.* PMLR.

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics, 47*(2), 387-443.

Lubart, T., Zenasni, F., & Barbot, B. (2013). Creative potential and its measurement. *International Journal for Talent Development and Creativity, 1*(2), 41-51.

Luchini, S. A., Maliakkal, N. T., DiStefano, P. V., Laverghetta Jr., A., Patterson, J. D., Beaty, R. E., & Reiter-Palmon, R. (in press). Automated Scoring of Creative Problem-Solving with Large Language Models: A Comparison of Originality and Quality Ratings. *Psychology of the Aesthetics, Creativity, and the Arts*.

Luchini, S. A., Moosa, M., Patterson, J. D., Johnson, D., Baas, M., Barbot, B., Bashmakova, I., Benedek, M., Chen, Q., Corazza, G. E., Forthmann, B., Goecke, B., Ibrahim, S.,

Karwowski, M., Kenett, Y. N., Lubart, T., Miroshnik, K. G., ..., & Beaty, R. E. (in press). Automated assessment of creativity in multilingual narratives. *Psychology of Aesthetics, Creativity, and the Arts*.

Luchini, S., Beaty, R., Boyce, A., Zappe, S., & Forthmann, B. (2025). Automating creativity assessment in engineering design: A psychometric validation of AI-generated design problems.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods, 17*(3), 437.

Miller, T. J. (2022). Assessing the desire to change personality across methods. *Journal of Personality Assessment, 104*(4), 447-457.

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity, 49*. https://doi.org/https://doi.org/10.1016/j.tsc.2023.101356.

Ozaki, K. (2024). Detecting inattentive respondents by machine learning: A generic technique that substitutes for the directed questions scale and compensates for its shortcomings. *Behavior Research Methods, 56*, 1-20.

Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S., Baker, B. S., ..., & Beaty, R. E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts, 17*(4), 495.

Patterson, J. D., Pronchick, J., Panchanadikar, R., Fuge, M., van Hell, J. G., Miller, S. R., ... & Beaty, R. E. (2025). CAP: The creativity assessment platform for online testing and automated scoring. *Behavior Research Methods*, *57*(9), 264.

Patterson, J., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2024). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior research methods, 56*(4), 3619-3636.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of machine Learning research, 12*, 2825-2830.

Qian, M., & Plucker, J. A. (2021). Creativity assessment. In J. A. Plucker, *Creativity and innovation* (pp. 223-234). Routledge.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ..., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research, 21*(140), 1-67.

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992).

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity research journal*, *24*(1), 92-96.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint*. https://doi.org/arXiv:1910.01108

Saris, W. E., & Gallhofer, I. N. (20214). *Design, evaluation, and analysis of questionnaires for survey research.* John Wiley & Sons.

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and psychological measurement, 82*(1), 29-56.

Sidney, S. (1957). Nonparametric statistics for the behavioral sciences. *The Journal of Nervous and Mental Disease, 125*(3), 497.

Sischka, P. E., Décieux, J. P., Mergener, A., Neufang, K. M., & Schmidt, A. F. (2022). The impact of forced answering and reactance on answering behavior in online surveys. *Social Science Computer Review, 40*(2), 405-425.

Stostic, M. D., Murphy, B. A., Duong, F., Fultz, A. A., Harvey, S. E., & Bernieri, F. (2024). Careless responding: Why many findings are spurious or spuriously inflated. *Advances in Methods and Practices in Psychological Science, 7*(1). https://doi.org/https://doi.org/10.1177/25152459241231581

Wankhade, M., Rao, A. C., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731-5780.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., ... & Poli, I. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research, 74*, 75-109.

Wolf, T., Debut, L., Sanh, V., Cahumond, J., Delangue, C., Moi, A., ..., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations.* ACL.

Yeung, R. C., & Fernandes, M. A. (2022). Machine learning to detect invalid text responses: Validation and comparison to existing detection methods. *Behaviour Research Methods, 54*(6), 3055-3070.

# Appendices

## A. Additional Statistics for Dataset Splits

Statistics for the validation, OOR, and OOI sets for both tasks are listed in Figures A through F. Note that, unlike for the train set, these sets were never augmented.

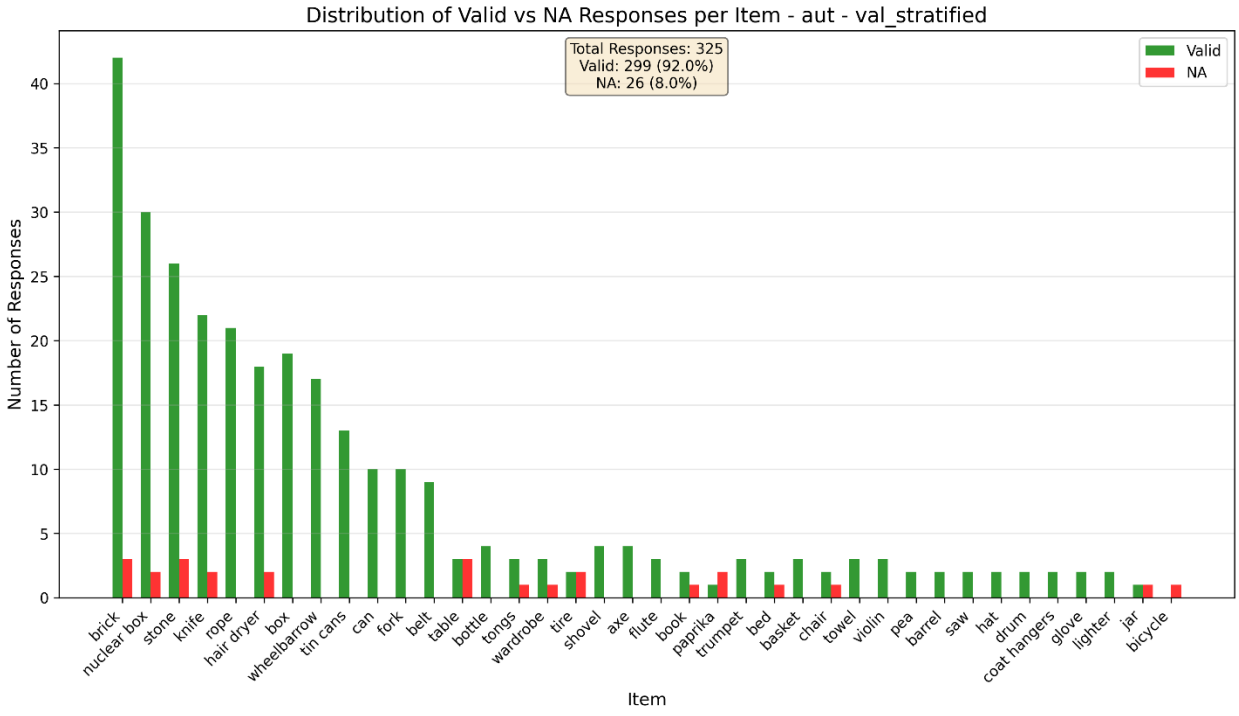



Figure A. Dataset statistics for the AUT validation set.

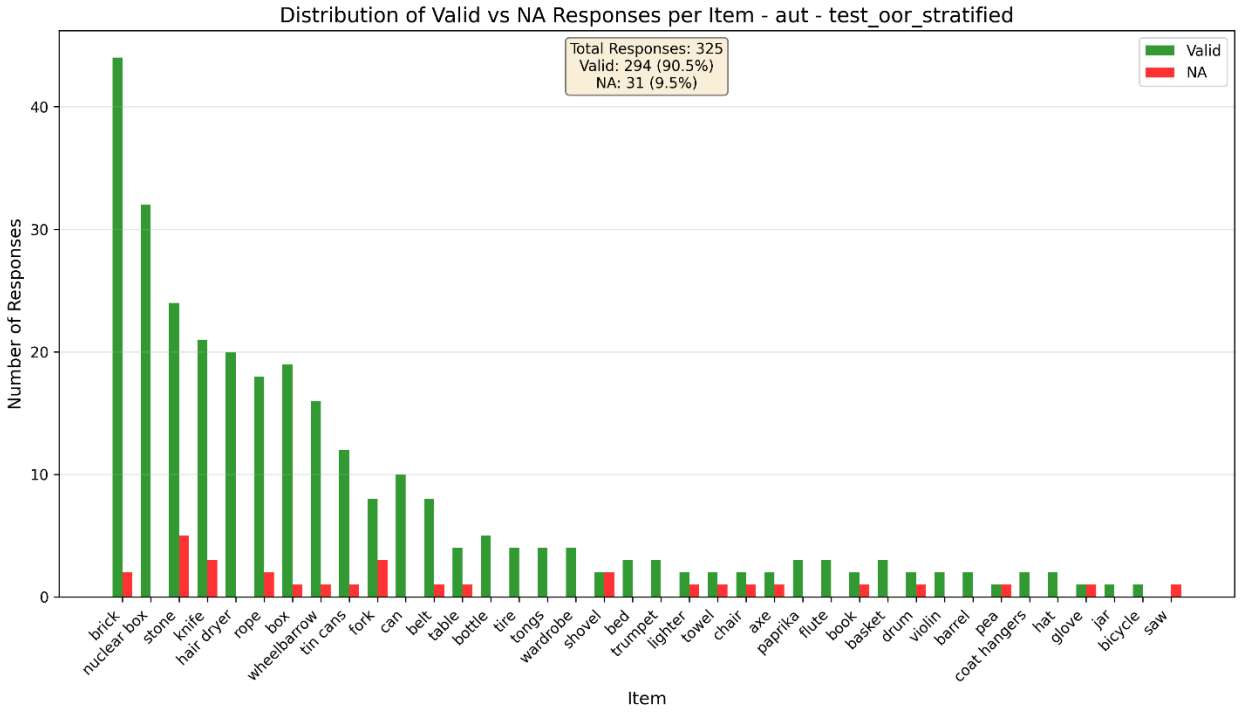



Figure B. Dataset statistics for the AUT OOR test set

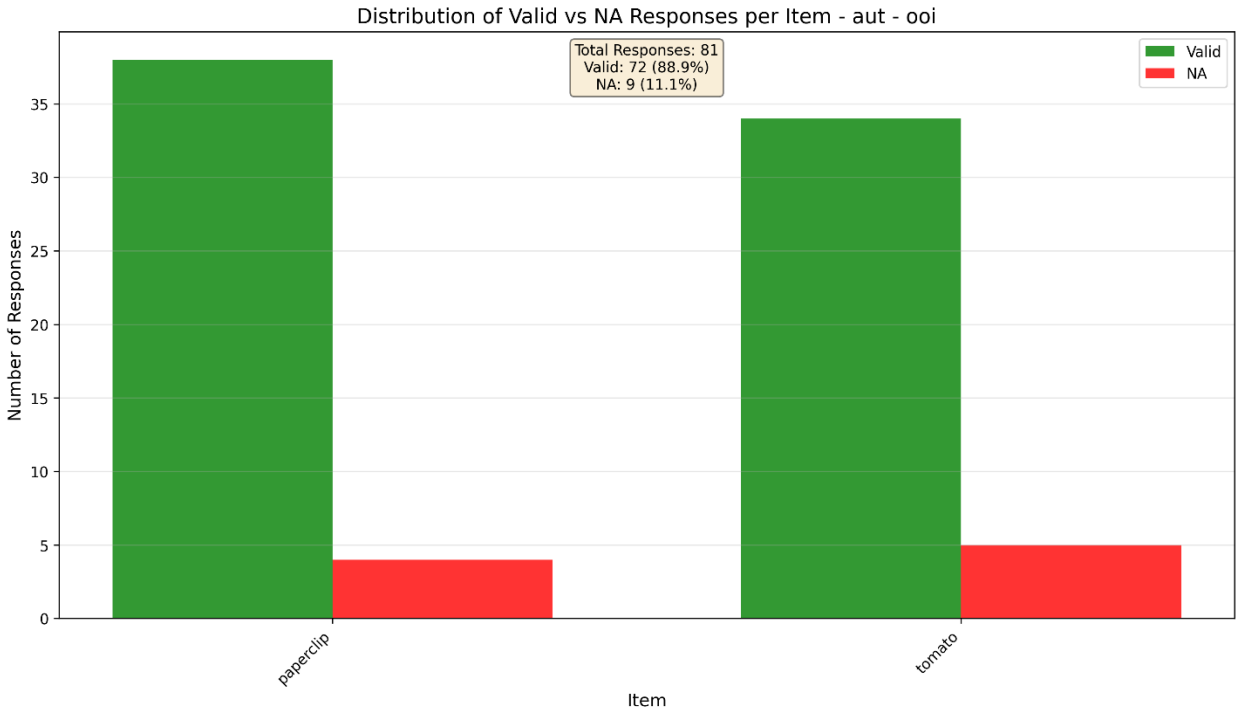



Figure C. Dataset statistics for the AUT OOI test set.

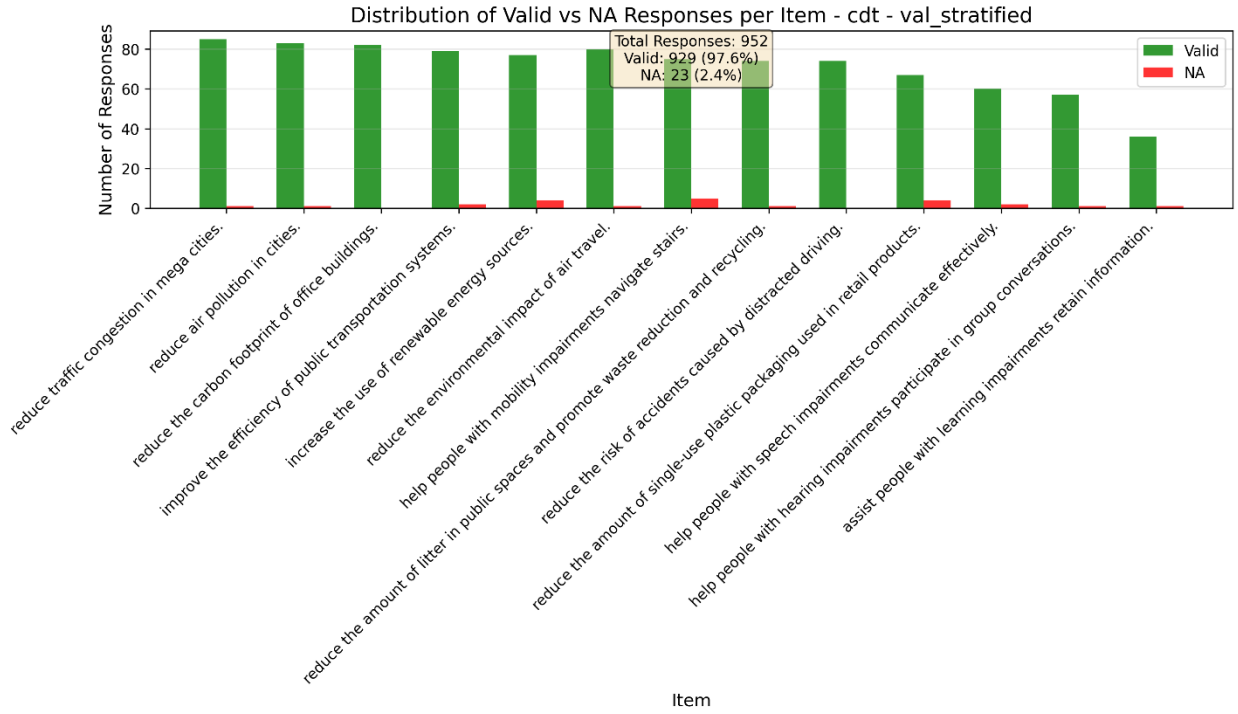



Figure D. Dataset statistics for the DPT validation set.

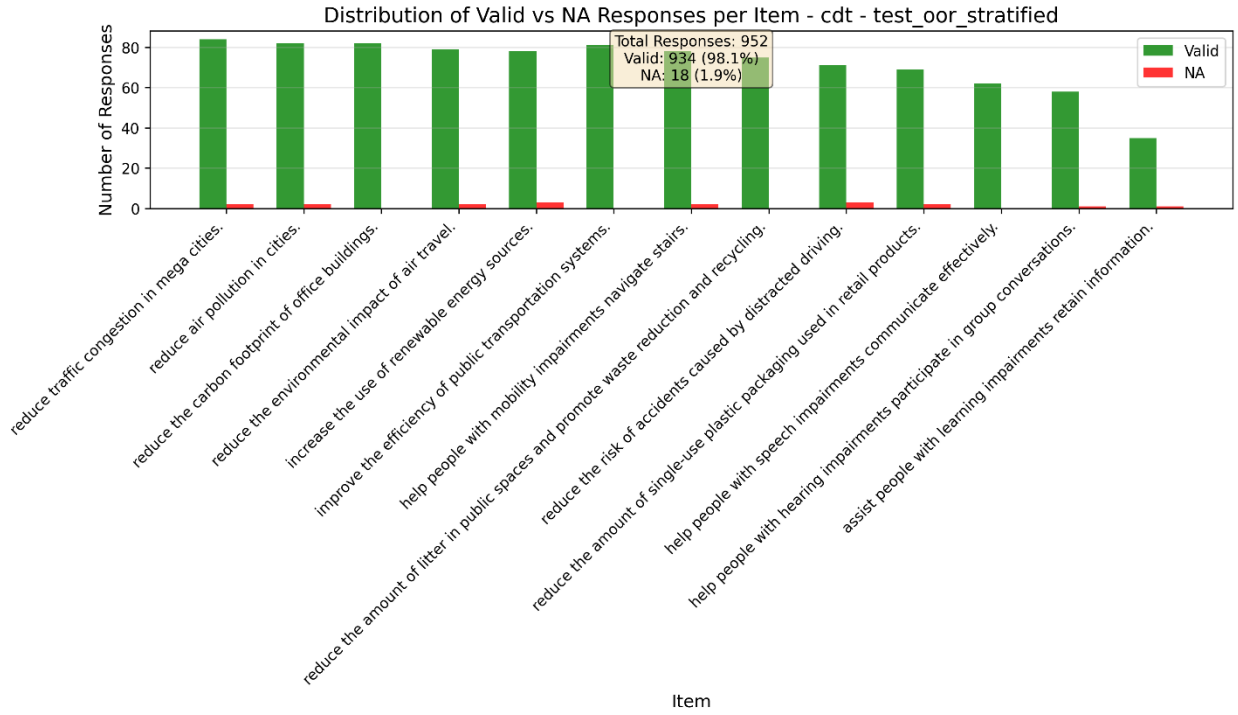



Figure E. Dataset statistics for the DPT OOR test set.

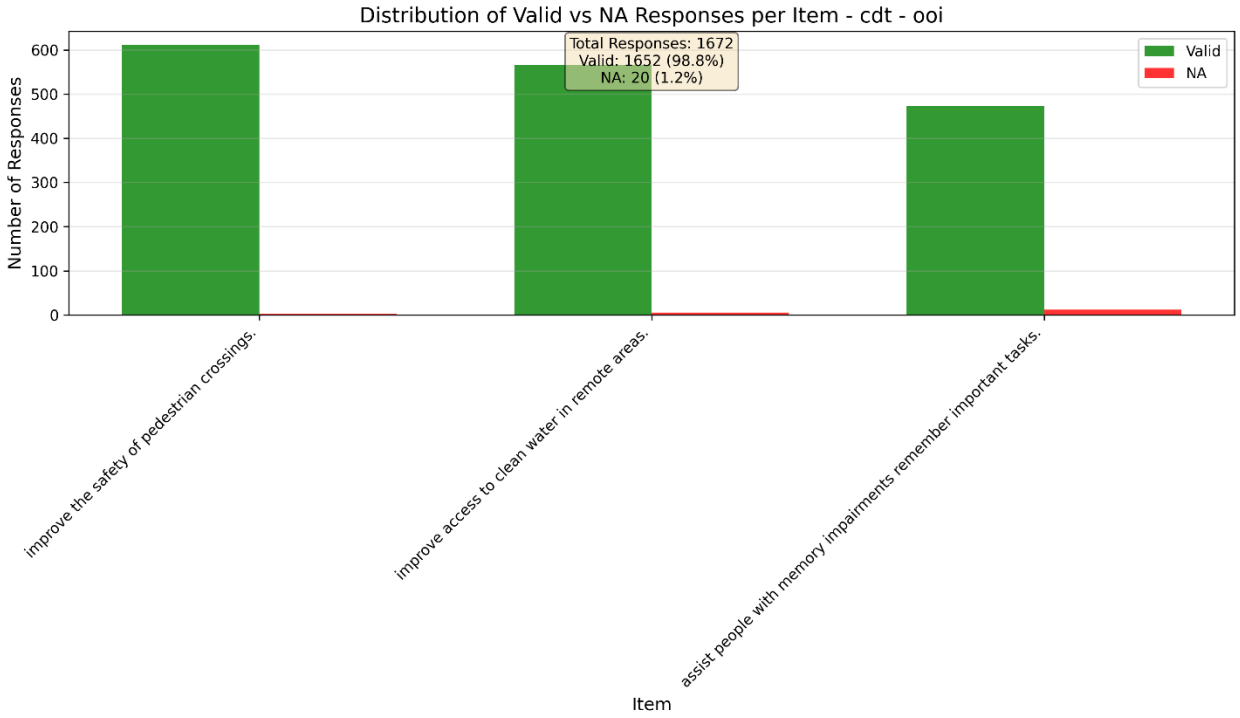



Figure F. Dataset statistics for the DPT OOI test set.

## B. Details on Hyperparameter Search for TLMs

We use Wandb to orchestrate our hyperparameter search and use Bayesian optimization to perform hyperparameter tuning. We perform a separate hyperparameter sweep for both tasks, using as a criterion the $F_1$ score on the respective task's dev set. Table A lists the hyperparameter ranges we searched for across all TLMs. We had each hyperparameter sweep run for a minimum of 100 steps and obtained the best-performing hyperparameters listed in Table B. We ran hyperparameter sweeps for only the DeBERTa and DistilBERT models and set hyperparameters for all other TLMs based on the optional ones found for these models. Hyperparameters for DeBERTa-v3-large were set to be the same as the optimal ones for DeBERTa-v3-base for the task in question. We used a learning rate, warmup ratio, and weight decay that were similar to these values for evaluating on the test sets. However, we elected to reduce the batch size to 1 and number of epochs to 2, as the gemma models required significantly more time and memory to

train, so keeping these hyperparameters set at higher values was unfeasible. Since we found that TLMs still performed well on OOI evaluations even with these modified hyperparameters, we believe they remain close to optimal.

| Hyperparameter | Values |
|---|---|
| Batch Size | 4, 8, 16, 32 |
| Learning Rate | 1e-07 – 1e-03 |
| Num Epochs | 1 – 20 |
| Warmup Ratio | 0 – 0.1 |
| Weight Decay | 0 – 0.01 |

Table A. Hyperparameter ranges for TLM models. TLM = transformer language models.

| Model | | DPT | AUT |
|---|---|---|---|
| DistilBERT | Learning Rate | 2e-6 | 5e-5 |
| | Epochs | 30 | 20 |
| | Batch Size | 16 | 8 |
| | Warmup Ratio | 0.001 | 0.001 |
| | Weight Decay | 0.001 | 0.001 |
| DeBERTa-v3-base | Learning Rate | 1e-5 | 1e-5 |
| | Epochs | 5 | 50 |
| | Batch Size | 16 | 4 |
| | Warmup Ratio | 0.1 | 0.01 |

| | | | |
|---|---|---|---|
| | Weight Decay | 0.01 | 0.0001 |

Table B. Best performing hyperparameters for all TLMs. Hyperparameters for DeBERTa-v3-large were set to be the same as DeBERTa-v3-base. DPT = design problems task, AUT = alternative uses task.

## C. Additional Precision and Recall Results

We include the precision and recall of all models (excluding LLMs) in Figures G though J. High resolution versions of these figures are also available in the supplementary materials.

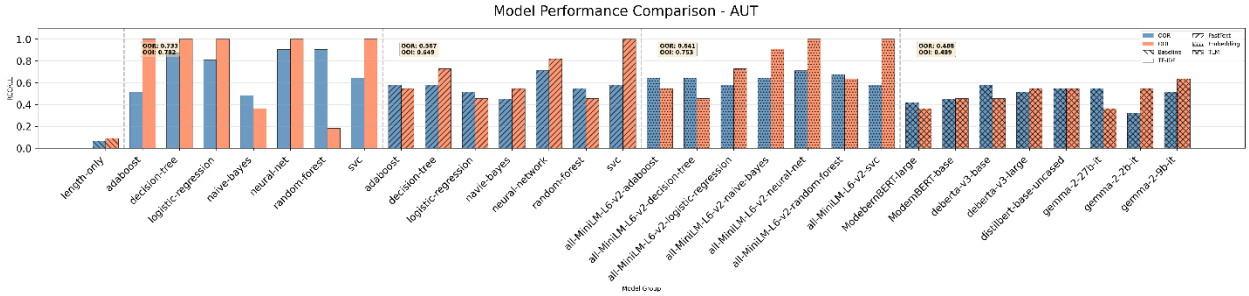



Figure G. Recall on the AUT test sets.

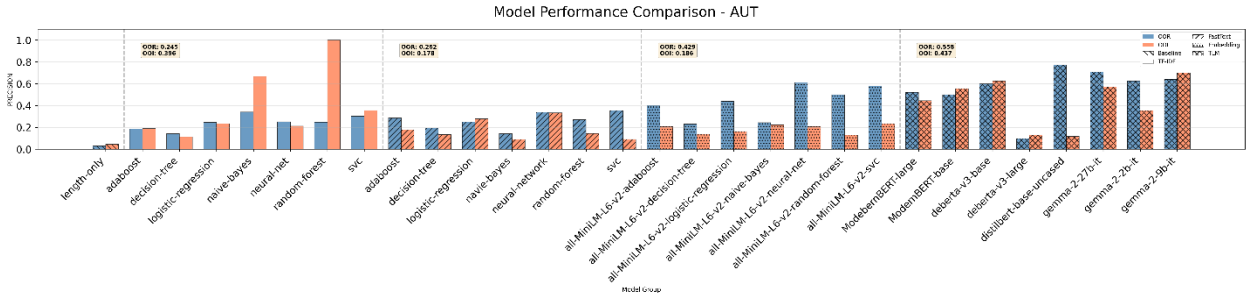



Figure H. Precision on the AUT test sets.

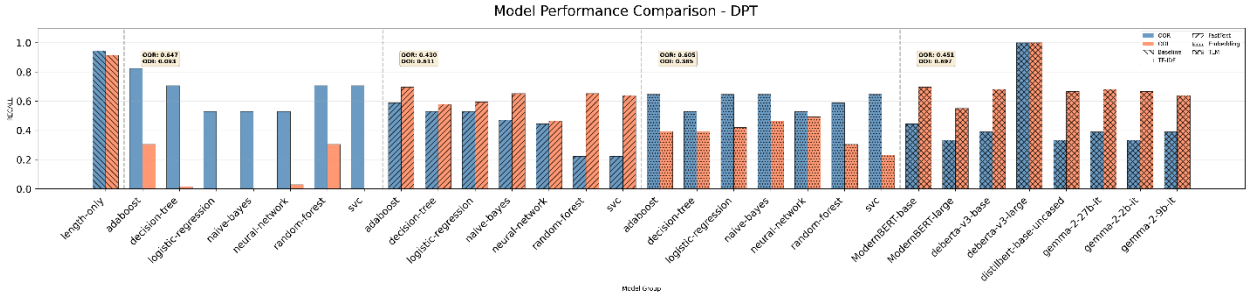



Figure I. Recall on the DPT test sets.

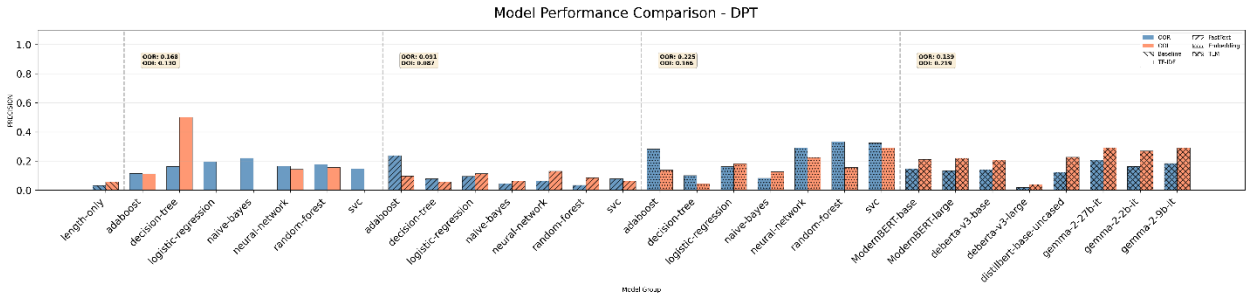



Figure J. Precision on the DPT test sets.

## D. Precision and Recall for LLMs

We include precision and recall scores from both gpt-5-nano and gemini-2.5-flash in Figures K through N.

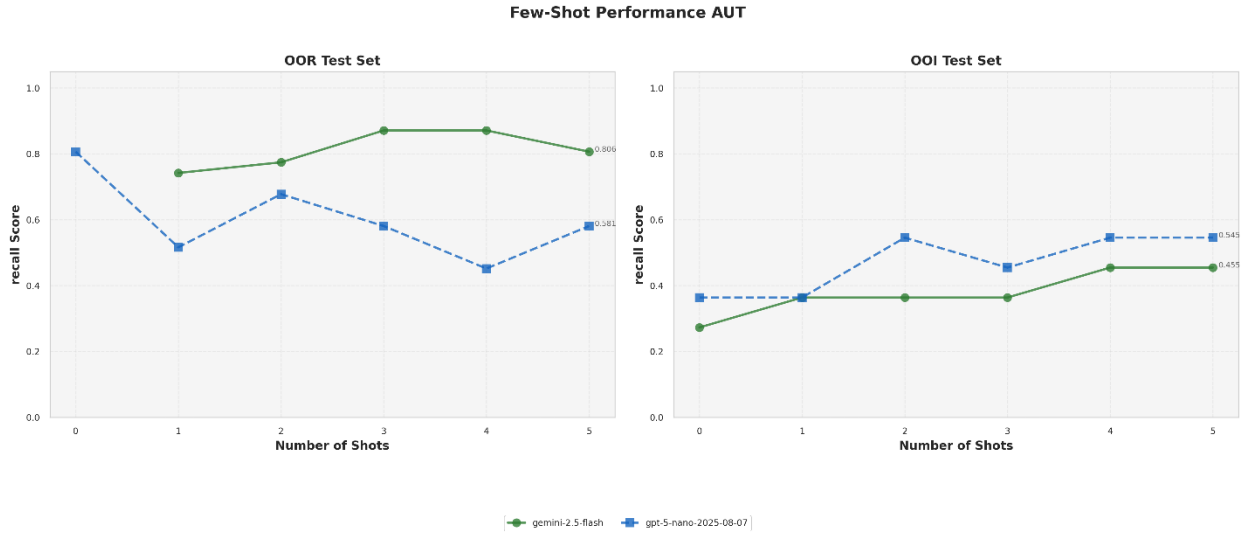



Figure K. Recall on the AUT test sets for few-shot LLMs.

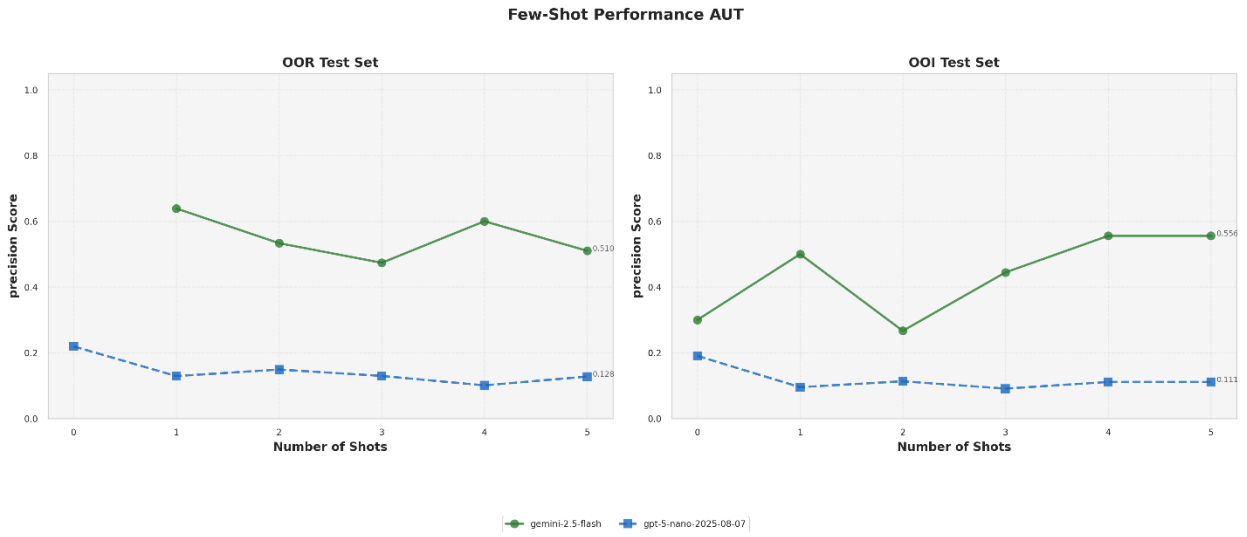



Figure L. Precision on the AUT test sets for few-shot LLMs.

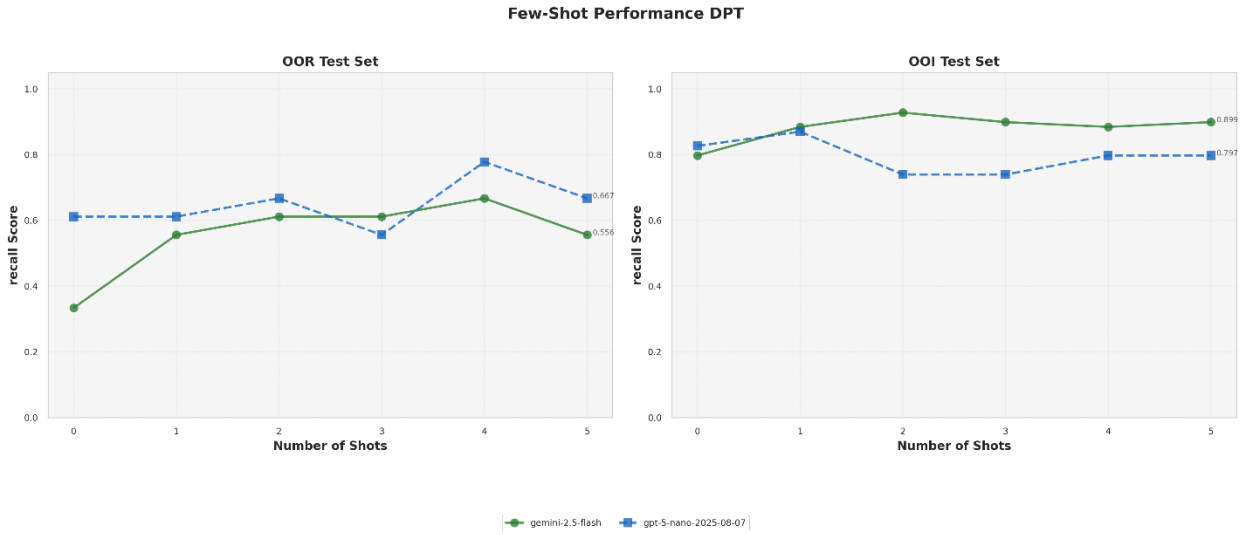



Figure M. Recall on the DPT test sets for few-shot LLMs.


Figure N. Precision on the DPT test sets for few-shot LLMs.

## E. Results at the Item Level

We include precision, recall, and F1 scores of models at the level of individual items, for both the AUT and DPT and for the OOR set, in Figures O and P. Each cell in the graphs corresponds to a specific model-item combination. High resolution versions of these figures are available in the supplementary materials.

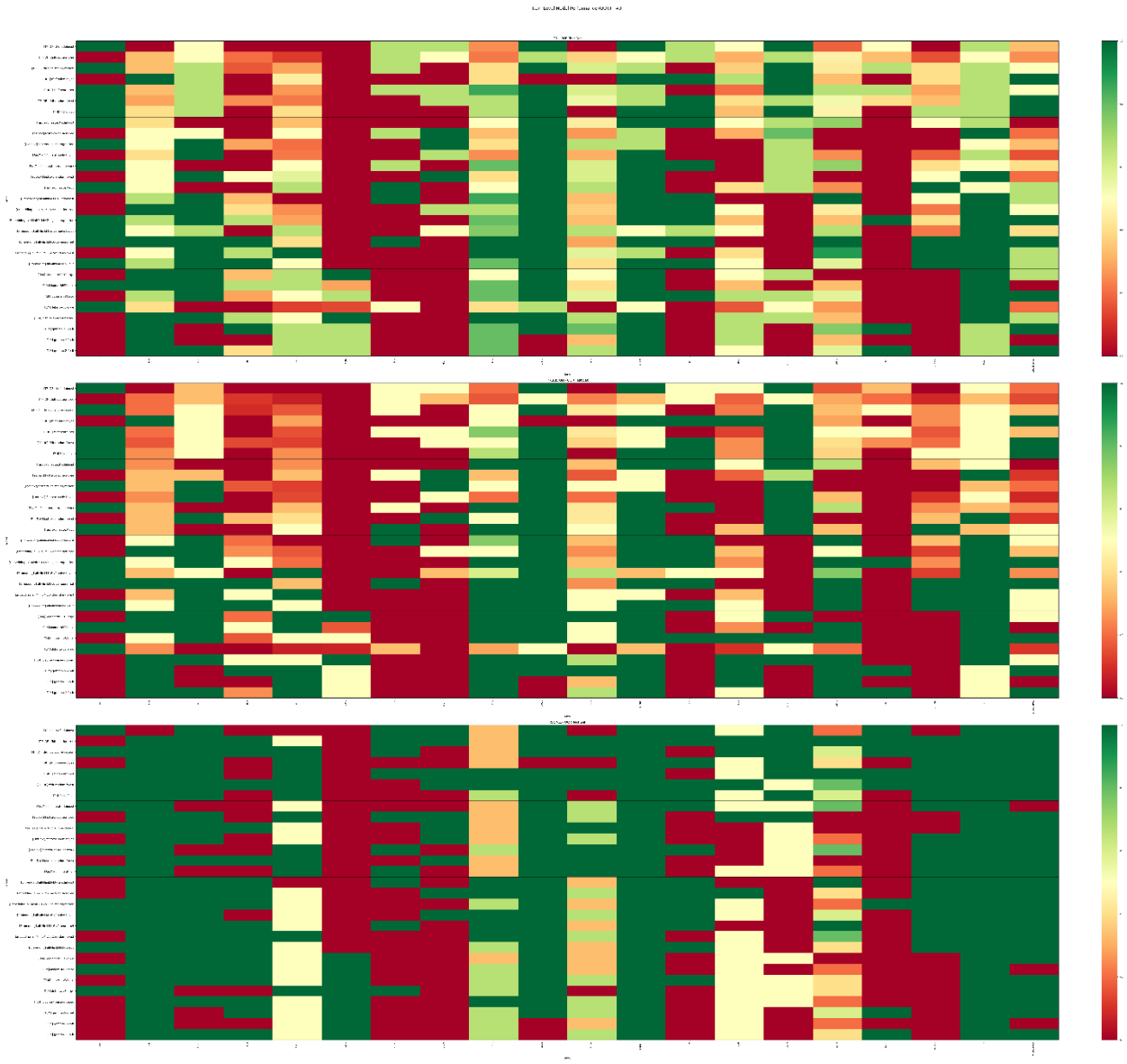


Figure O. Precision, recall, and F1 scores for the AUT OOR set. A high-resolution version of this figure is available in the supplementary materials.

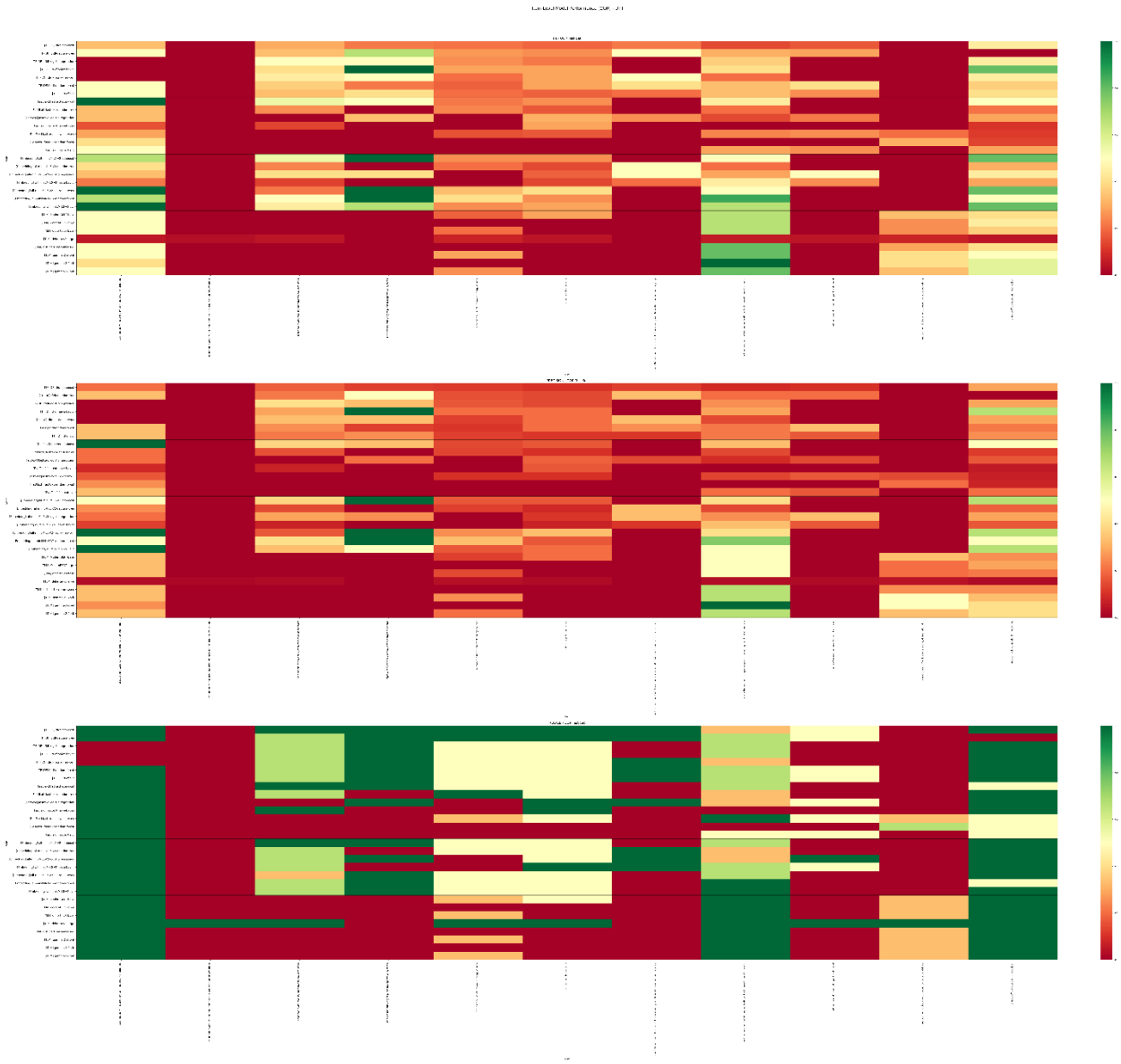


Figure P. Precision, recall, and F1 scores for DPT OOR set. A high-resolution version of this figure is available in the supplementary materials.

## F. Rating Procedure for Calibration Set

We recruited three undergraduate research assistants with prior experience performing quality scoring to perform the ratings for the calibration set. Before the rating procedure, the raters took part in a one-hour-long training session. To anchor raters onto a common understanding of invalid responses, they were first provided a definition: “A response is invalid

if for any reason it does not comply with the instructions of the task. In general, an invalid response does not address the task in any meaningful way." We also introduced them to the task requirements of the DPT and provided them with manually curated examples of valid and NA responses. Once raters had finished individually rating the example responses, one of the co-authors led a discussion for each individual practice item, ensuring that raters would understand the true rating of each example response. We collected ratings using Doccano and assessed rater agreement using the intraclass correlation coefficient implemented in the Pingouin library.[2] We manually assessed the rater agreement and conducted additional rounds of training until we reached a satisfactory agreement level. Due to high degrees of freedom, the rater agreement was consistently found significant ($p < 0.001$).