

EVALUATION OF VISION LANGUAGE MODELS WITH ITEM RESPONSE THEORY

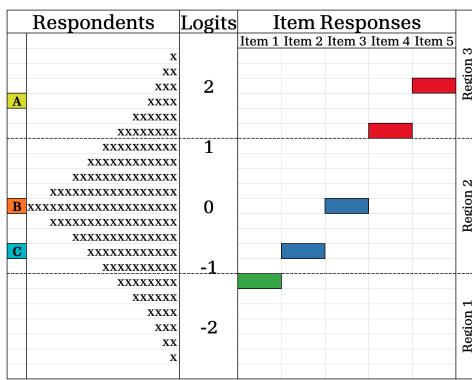
000
001
002
003
004
005 **Anonymous authors**
006 Paper under double-blind review
007
008
009
010

ABSTRACT

011
012 Evaluation of generative AI output is difficult because of the high dimensional
013 nature of the problem space. Accuracy-oriented benchmarks are often used to
014 assess the quality of outputs, but may not provide a complete picture because they
015 do not incorporate the difficulty of items relating to a particular task. We present
016 the use of Item Response Theory (IRT) to evaluate the output of a cohort of Vision
017 Language Models (VLMs) on two different tasks: image caption rating, and visual
018 reading comprehension. As a result, we show how to find meaningful differences
019 between popular state of the art VLMs, promoting meaningful interpretability.
020 IRT can be used in many areas of the ML workflow. We will cover some of the
021 prior work and ways IRT may be used in your research. Our aim is to encourage
022 the adoption of IRT by the ML community as a general tool for evaluation.
023
024

1 INTRODUCTION

025 In this work, we showcase the use of IRT for evaluating latent abilities of ML models. We will first
026 review the basics of IRT and Wright Map Analysis, and highlight some previous work related to
027 ML. We will then apply IRT to two case studies involving evaluation of VLMs on different tasks:
028 caption rating for visually impaired and blind people Scott et al. (2023); Narins et al. (2023), and
029 visually-informed comic reading comprehension Blum et al. (2020). We will then suggest other
030 ways that IRT may be used in ML workflows and generative AI research. Our hope is that these
031 techniques will be adopted and broadened by the ML community to aid interpretability, validation,
032 trust, fairness, alignment, and bias of machine learning models, datasets, and scales.
033
034



046 Figure 1: General form of Wright Map, here shown with 5 items and several respondents. The
047 area on the left is the distribution of the respondent’s ability, and the area on the right is the item’s
048 difficulties. They are both plotted on the same logit scale, providing probabilistic interpretation.
049

050 Traditional ML metrics for LLM and VLM development focus on optimizing the performance of
051 the model, and assumes that each sample in the dataset is equally difficult to infer. When we want to
052 find out which model is best at a natural language processing (NLP) task such as question answering
053 (QA), visual question answering (VQA), captioning, image-caption rating, semantic similarity, or
sentiment analysis, we often use classification metrics or NLP metrics. Metrics such as accuracy,

precision, recall, F1-score, BLEU Papineni et al. (2002), METEOR Denkowski & Lavie (2014), ROUGE Lin (2004), and CIDEr Vedantam et al. (2015) are used to cross-compare the performance of top models on common datasets, but they do not consider the differences in individual samples; IRT does account for those individual differences when assessing the model performance.

The main contribution of this paper is a new methodology and benchmark for evaluating VLMs using IRT and Wright Map Analysis Wright (1982). The methodology enables us to determine how well VLM respondents perform relative to humans, and the benchmark establishes a quantitative and interpretable process for evaluating progress in VLM research. We cover two specific case studies and analyze them from an IRT lens. The results of the paper provide useful information to researchers working on evaluating model outputs, and for generating synthetic data for augmenting datasets. The results may also be useful for researchers who are collecting data from human raters and want to determine validity of using VLMs as supplemental raters, as a possibly faster, alternative to human-annotated large datasets.

2 ITEM RESPONSE THEORY

Item Response Theory is a methodology used in Psychometrics to study psychological traits and abilities Lord & Novick (2008). It uses the terms *items* and *respondents* to refer the individual samples and persons. As the name implies, IRT takes into account the characteristic behavior of the item when assessing performance. Specifically, IRT measures respondent's *ability* in the presence of or with regard to an item's *difficulty*, providing a more nuanced view of performance. IRT is often used in standardized testing, such as the GMAT and GRE tests Murodulloyevich (2025), and is desired when optimal decisions are necessary, and interpretability is essential.

For example, imagine giving a test of 100 multiple-choice problems to 100 students. Each student has their own accuracy measured by the percentage of problems they got correct, but the problems also have an accuracy measured by the percentage of students that got the problem correct. A problem's accuracy reveals something about its difficulty: problems that everyone got correct are too easy, and problems everyone got wrong are too hard. It is possible that, when grouped by difficulty, there may be an imbalance in the percentage of easy problems that were correct versus the percentage of hard problems that were correct. This could lead to a situation shown in Table 1.

If we select the best model based on Total Accuracy, then we would select Student A, but that would also mean we would inadvertently be optimizing for Easy problems.

Student	Total Acc.	Easy Acc.	Hard Acc.
A	93%	100%	86%
B	92%	91%	93%

Table 1: Accuracy breakdown by item difficulty. Which student did better?

IRT considers the item difficulty along with the person's ability and places them on the same probabilistic, interval scale (i.e. log-odds, or logits). Placing both items and persons on the same interval scale allows for a more detailed level of analysis because the magnitudes of the differences between items and persons have probabilistic meaning, and may be considered in context of the latent trait being modeled. To do this, we apply a one parameter logistic model (1PL), also known as a *Rasch Model*. The 1PL equation is given by:

$$P(X_i = 1 | \theta, \delta_i) = \frac{1}{1 + e^{-(\theta - \delta_i)}} \quad (1)$$

where θ is the person's *ability*, and δ_i is the *difficulty* of item i Rasch (1960); Wilson (2023). The values for θ and δ are found by using statistical inference methods such as Maximum Likelihood Estimation (MLE), Monte-Carlo Estimation, or Variational Inference, which also generates item "fit" statistics that indicate how reliable the estimates are Wu & Adams (2013). In this work, we used the partial-credit model (PCM), which is an extension to the 1PL that preserves the levels in polytomous data. That equation is given by:

108
109
110
111
112
113

$$P(X_i = k | \theta, \delta_i) = \frac{\exp\left(\sum_{j=0}^k (\theta - \delta_{ij})\right)}{\sum_{h=0}^m \exp\left(\sum_{j=0}^h (\theta - \delta_{ij})\right)} \quad (2)$$

where δ_{ij} is the difficulty of item i at step level j Wright (1982); Wilson (2023). We focus specifically on the 1PL, rather than 2PL and 3PL, because of its interpretative power of placing respondents and items on the same logit scale, and the fact that the WrightMap is only defined for 1PL.

Since θ and δ are calibrated on the same logit scale, they can be plotted relative to one another on an *item-response map*, also known as a *Wright Map* Wilson (2017), which is a compact and very informative visualization for communicating Rasch-family model output. Mark Wilson coined the term “Wright Map” in honor of Ben Wright who was a proponent of their use and expressive ability, and had extended them in many ways Wilson (2011). We use them for their expressive interpretability, identifying meaning in the latent space we are studying.

The Mean Square fit statistics Wu & Adams (2013) is a quantitative way to evaluate the validity of the Wright Map. The ideal range for the fit statistics is between 0.77 and 1.33, where 1.0 is the ideal score. If an item has a fit statistic greater than 1.33, it would suggest too much randomness in the item; that is, people with high ability are getting it wrong, and people with low ability are getting it right. A fit statistic below 0.77 suggests overfit, that the item is too predictable. These fit statistics are checked to be within acceptable range before accepting the validity of the Wright Map.

Figure 1 shows a generic Wright Map. The key features of the Wright Map are the person abilities, plotted on the left as a histogram or distribution, the mean item difficulties, plotted on the right at their individual locations, and the interval-level logit scale marked in the middle and showing the y-axis of the graph. The dotted lines are *waypoints* that signify theoretical or substantive regions of interest. They are a visual indicator showing certain items correspond to certain regions. The letters “A”, “B” and “C” on the far left show particular respondents in which we are interested in studying. The “x” marks represent the entire sample studied, including respondents of interest.

Once the Wright Map is generated, interpreting the result is like reading a map, and inferences may be made about the differences in locations between respondents and items (including levels within items). A respondent’s location on the scale is often interpreted as a “proficiency” or “ability”, a term from academic testing, since in that case it represents their likelihood of responding correctly to more items. The zero value on this scale represents the average location of all respondents. The “location” of an item is a real number that indicates the difficulty to rate that item correctly. Higher values correspond to higher difficulty. If the person location is equal to the item location, then that person has an ability equal to that item’s difficulty. If a person location is higher it has a greater ability than that item’s difficulty. If the person location is lower, it has a lesser ability than that item’s difficulty. The item-characteristic curve (ICC) for the 1PL is shown in Figure 2.

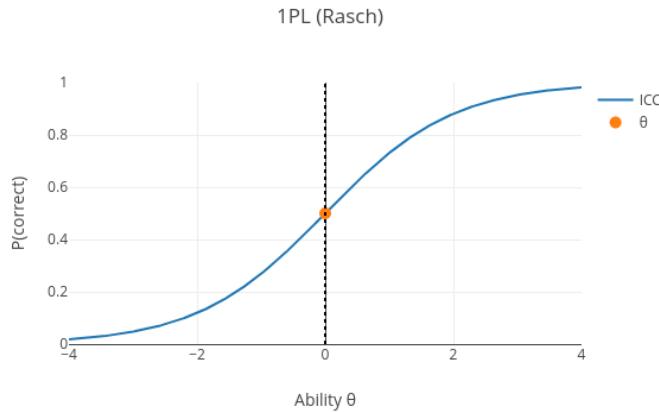
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159

Figure 2: The Item Characteristic Curve (ICC) for the 1PL Rasch Model. It shows the Ability, θ , plotted against the probabilistic scale, estimating the likelihood of getting an item correct.

162

3 RELATED WORK

164 We are not the first to notice the utility that IRT could have in the ML space. An early, notable work,
 165 from 2023, Narins et al. (2023), used IRT in the NLP space for validating an image-caption rating
 166 dataset. Another work took place at a workshop at the 2024 European Chapter of the Association
 167 for Computational Linguistics (EACL) conference. At the conference, Lalor, et. al. covered their
 168 Python library, `py-irt`, and applying IRT to the NLP space in multiple ways, including classifi-
 169 cation, data validation, a way to find annotation error, and using θ as an alternative to accuracy for
 170 ranking Lalor et al. (2024a;b); Lalor & Rodriguez (2023). Another work, Zhou et al. (2025), also
 171 uses θ for ranking, and compares it with popular benchmarks, finding that in some cases ranking by
 172 θ changes the ordering on a leaderboard. Both, Lalor et al. (2024a) and Zhou et al. (2025) discuss
 173 how benchmarks may be misleading and how IRT can provide more information than traditional
 174 metrics alone.

175 In the last couple years, there have been several new and intriguing research developments applying
 176 Psychometric practices and IRT to NLP problems for evaluation, calibration, validation, and align-
 177 ment Liu et al. (2025); Sharpnack et al. (2024); Song et al. (2025); Veeramani et al. (2024); Jain et al.
 178 (2025); Narins et al. (2023). We believe this is a result of the plethora of models we have to choose
 179 from, the need to discern between them, and because traditional evaluation metrics and benchmarks
 180 that do not include item information could be misleading Zhou et al. (2025); Schilling-Wilhelmi
 181 et al. (2025); Sarkar et al. (2025); Ding et al. (2024). Many researchers are discovering and explor-
 182 ing the use of IRT when using LLM-as-a-Respondent and LLM-as-a-Judge Scarlatos et al. (2025);
 183 Gurdil et al. (2024). And some are using Pyschometrics and IRT to explore the psychological aspect
 184 of AI Federiakin (2025); Vogelsmeier et al. (2025); Chollet (2019); Zhuang et al.; Ye et al. (2025).

186

4 METHODS

188 In the following sections, we present two case studies utilizing VLMs-as-a-Respondent, comparing
 189 and analyzing against a ground-truth of humans. Using a 1PL Rasch model with Wright Map not
 190 only allows us to find the best VLM for a ranking task, but also provides an interpretable analysis
 191 of our item difficulties, person abilities, construct map, and scale. Combining all of these allows us
 192 to ensure the internal structural validity of the latent trait we are modeling. In our work, all Rasch
 193 model estimations were generated by using ACER Conquest 5.0 Adams et al. (2012).

194 Performing a complete IRT and Wright Map analysis involves collecting data, running statistical
 195 inference to find θ and δ , checking item fit statistics, checking internal reliability and structure of the
 196 scale, generating a Wright Map with waypoints and plotting individual locations of respondents of
 197 interest, then interpreting the results in the context of the latent ability that is being assessed Wilson
 198 & De Boeck (2004); Wilson (2023).

201

5 VISION LANGUAGE MODELS

203 We used 10 popular VLMs from 6 vendors. We wanted to use a variety of models based on parameter
 204 scale, architecture, and training strategies, but we also had to work within budget and hardware
 205 availability. The models used in the study are below. We'll describe each model briefly. Note that
 206 not all models were able to perform in both case studies ¹.

207 *SmolVLM-Instruct* Research (2024) This model was developed by HuggingFace to be an open, com-
 208 pact model. It has 2.25 billion parameters, and was optimized for instruction-following and fast
 209 inference. It was released on November 18, 2024.

211 *Phi-3.5-Vision-Instruct* Microsoft (2024) This model was developed by Microsoft to be an open,
 212 mid-sized model. It has 4.2 billion parameters, and was trained on synthetic and filtered public data.
 213 It was optimized for instruction-following with images. It was released on August 20, 2024.

214 ¹SmolVLM-Instruct was unable to process text-only input, and LLAMA3.2-11B-Vision-Instruct was unable
 215 to handle multiple images, both in the visual reading comprehension study.

216 *Qwen2-VL-7B-Instruct* Team (2024) This model was developed by Alibaba Cloud to be an open,
 217 large model with 7 billion parameters. It was optimized for instruction-following with images. It
 218 was released on October 3, 2024.

219 *Gemini-1.5-flash* AI (2024) This model was developed by Google. It is proprietary but we know it
 220 is based on a sparse mixture-of-experts (MoE) transformer architecture. The parameter count has
 221 not been disclosed. It was released on May 14, 2024.

222 *Llama3.2-11B-Vision-Instruct* Meta Platforms (2024) This model was developed by Meta to be an
 223 open, large model with 11 billion parameters. It was optimized for instruction-following with im-
 224 ages. It was released on September 25, 2024.

225 *GPT-4o-2024-08-06*, *GPT-4o-2024-05-13*, *GPT-4o-2024-11-20*, *GPT-4o-mini*, *GPT-4-turbo-2024-04-09* OpenAI (2025) These models represent different iterations and parameter scales of the GPT-4 family. The “o” (omni) series allows for processing image embeddings alongside textual prompts, while the “mini” variant serves as a parameter-efficient version. The “turbo” variant has optimizations in speed and context handling compared to earlier GPT-4 models. GPT-4o-mini was released on July 18, 2024.

233 6 CASE STUDY 1: IMAGE CAPTION RATING

234 Automatic Generation of Image Captions and Video Descriptions is essential for visual retrieval,
 235 recommendation engines, and accessibility for Blind and Low Vision (BLV) individuals Karpathy & Fei-Fei (2015); Datta et al. (2008); Crudden & McBroom (1999); Kirchner & Smith (2005);
 236 McDonnell & Sui (2019); McMillen & Alter (2017); Shaw et al. (2007). Assessing the quality of
 237 the image captions and video descriptions is an increasingly important area of research as more
 238 generative solutions come online. Given the limitations of existing metrics, direct human evaluation
 239 remains the gold standard, but it is costly, subjective, and time-consuming to run data collection with
 240 humans-in-the-loop. The Validated Image Caption Rating (VICR) study addressed some of these
 241 challenges by introducing a high quality dataset and a well-defined rating scale, calibrated with IRT,
 242 which reflects human assessments of the quality of image-caption pairs, specifically for the BLV
 243 community Narins et al. (2023); Scott et al. (2023). We extend the VICR work with the addition of
 244 VLM raters to do the same task as the human raters. We then analyze the VLMs with IRT and a
 245 Wright Map with waypoints in conjunction with the original human raters.

246 The VICR dataset uses a detailed and robust rating scale, which was designed to capture the extent of
 247 four essential aspects of image captions: accuracy, completeness, local context, and global context
 248 and inferential information. The authors of the VICR scale claim it is based on sociocognitive
 249 theories Blum et al. (2020); Kintsch (1988); Smith & Hancox (2001). In their study, they selected
 250 25 image-caption pairs from their dataset, originally used as tutorial examples in a gamified, rating
 251 task known as the “Rating Game.” The 25 items were comprised of 5 distinct images, each paired
 252 with 5 distinct captions that reflected the 5 levels of coherence on the scale, ranging from level 1
 253 (lowest coherence) to level 5 (highest coherence). All captions were carefully selected by multiple
 254 domain experts to represent the full spectrum of quality.

255 The original Rating Game collected 132 human responses. We then prompted 10 VLMs to rate the
 256 same 25 image-captions pairs. We combined the human and VLM-generated ratings for the 1PL
 257 Rasch analysis, and investigate not only whether VLMs can serve as valid substitutes for human
 258 raters, but also whether the 25 items are sufficiently informative for evaluating rating behavior. The
 259 prompt given to the VLMs is shown in Appendix A. Essentially, they were prompted to play the role
 260 of an expert image-caption rater for visually impaired and blind users, and requested to rate each
 261 image-caption pair using the scale provided. Also see Appendix A for the scale.

262 Since prompts are highly subjective and can significantly influence model performance, we designed
 263 our prompts by following the “Hierarchical Prompt” methodology from the ShareGPT4V Chen et al.
 264 (2024) prompt template as a foundation. Each VLM was assigned a specific role, treating them as
 265 a “character” within the prompt. The prompts are structured to optimize the model’s performance
 266 in assessing caption quality. First, we define the role of the model, explicitly guiding it toward the
 267 task of rating captions. Then, we provide the rating scale as a reference framework to standardize its
 268 evaluations. Finally, we constrain the model’s task to solely generate ratings. We also provide the

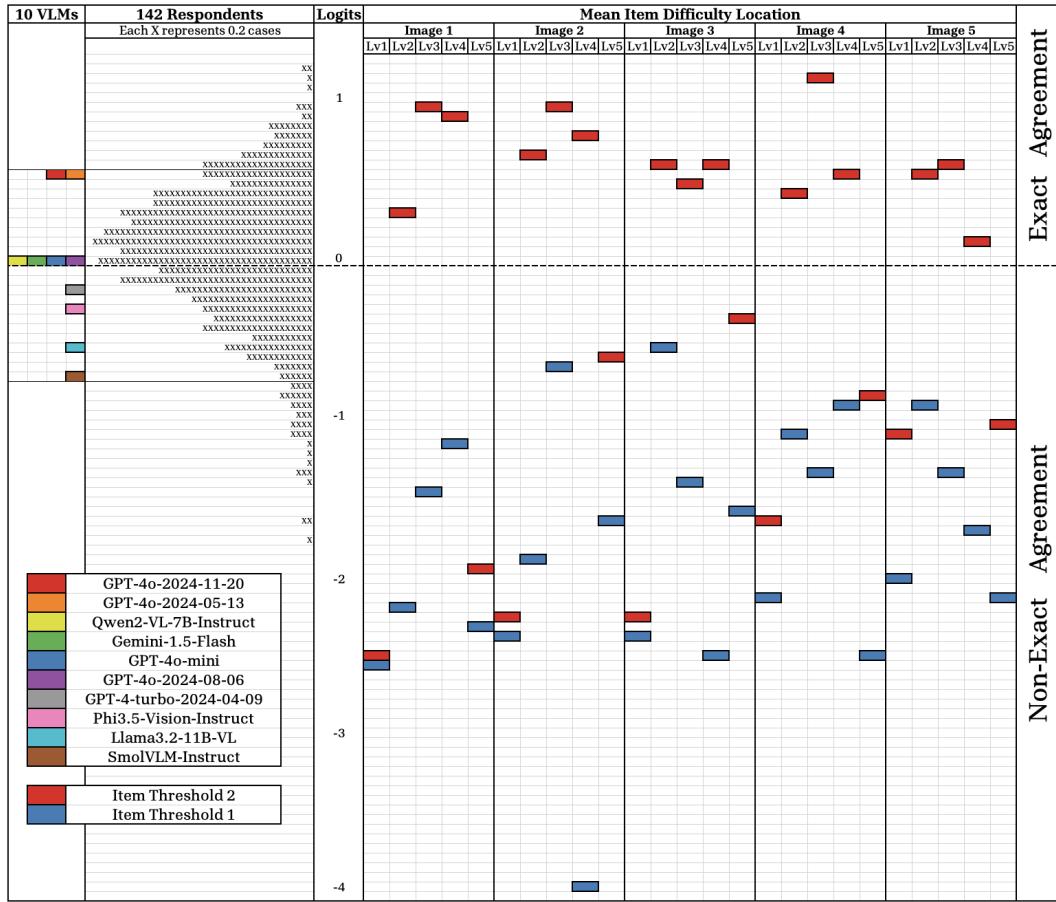


Figure 3: The Wright Map for the Image-Caption Rating task. The distribution of θ abilities for respondents is shown under the column “142 Respondents”. The VLMs are included in this group and additionally pulled-out into their own column, labeled “10 VLMs” so they can be seen in isolation. The item-map, on the right side, shows each of the 25 items (image-caption pairs) and their location. You can see “banding” of red and blue items. The waypoint separates the bands showing where respondents had Exact Agreement with the ground-truth or Non-Exact.

base64-encoded image and associated caption in each prompt. Each rating was generated independently, without the use of any memory, to avoid any carryover effects.

6.1 IRT AND WRIGHT MAP ANALYSIS

Figure 3 shows the Wright Map for the results we obtained from the 1PL for the image-caption rating task. In our analysis, all 25 items had fit statistics within the acceptable range so we can trust the validity. The Partial Credit Model (PCM) was used to label the responses Wright (1982) at 3 levels of agreement: level 0 is called “Distal Agreement”, meaning a rating two units or more away from the correct rating. Level 1 is called “Adjacent Agreement”, meaning a rating exactly one unit above or below the correct rating. Level 2 is called “Exact Agreement”, meaning that the respondent gave the correct numerical rating as the value provided by the expert. Consequently, there are 2 cumulative thresholds, between these 3 levels Wilson (2023). The Wright Map empirically represents the hypothesis that the levels of agreement defined above (distal, adjacent, exact) behave consistently across items.

Respondents aligned with *Item Threshold 1* for a given item have a 50% chance of scoring in the lowest category (e.g., level 0 representing Distal Agreement) and a 50% chance of scoring into any of the upper levels (Adjacent Agreement and Exact Agreement) for that item. Respondents aligned with *Item Threshold 2* for an item have a 50% chance of scoring level 0 and 1, and 50% chance of scoring into level 2 (Exact Agreement). Respondents who are above a threshold have more than 50%

Vision Language Models	Ability (θ)	Percentile
GPT-4o-2024-11-20	0.52807	89%
GPT-4o-2024-05-13	0.52807	89%
Qwen2-VL-7B-Instruct	0.01775	54%
GPT-4o-mini	0.01775	54%
GPT-4o-2024-08-06	0.01775	54%
Gemini-1.5-flash	-0.09786	41%
GPT-4-turbo-2024-04-09	-0.21015	33%
Phi-3.5-Vision-Instruct	-0.31962	24%
Llama3.2-11B-Vision-Instruct	-0.53158	16%
SmolVLM-Instruct	-0.73605	12%

Table 2: Evaluation of each VLM on the image caption rating task. The columns labeled “Ability” and “Percentile”, indicating the logit location for each VLM (its ability) and the percentage of the human raters that the VLM matched or exceeded in ability.

chance of being correct on that item at the level corresponding to that threshold; whereas respondents who are below that threshold have less than a 50% chance. Overall, respondents that are higher on the scale have more ability at rating image-caption pairs. Items and their associated threshold, that are high on the scale are considered more difficult for respondents to achieve “Exact Agreement” (e.g. level 2). The inverse holds true for respondents and item thresholds that are comparatively lower.

The column labeled “10 VLMs” shows the location of the ten VLM respondents on the scale. The column labeled “142 respondents” shows the location of all 142 respondents, which includes the 10 VLMs and the 132 human respondents. As a reminder, a higher location for a respondent indicates higher proficiency with respect to latent ability of rating image-caption items (image-caption pairs). The 25 columns on the right, grouped under the heading “Mean Item Difficulty Locations” correspond to the 25 items that we used in constructing this map. Each column corresponds to one item, i.e., one image-caption pair. Each column indicates with a red rectangle the location of Item Threshold 2 for that item, and with a blue rectangle the location of Item Threshold 1 for that item.

One of the observations we see on the Wright Map is that Items representing the middle categories (i.e. items with VICR ratings of 2, 3, or 4) are harder to get exact agreement than the extremes (e.g., items with VICR ratings of 1 or 5). This shows that the appropriate score for mid-quality captions is harder for individuals to discern than for very high or very low quality captions.

GPT-4o-2024-11-20 and GPT-4o-2024-05-13 had the highest θ (ability) out of all VLMs; these top two VLMs also performed better than or equal to 89% of the human respondents. There are 4 other VLMs Qwen2-VL-7B-Instruct, Gemini-1.5-Flash, GPT-4o-mini, and GPT-4o-2024-08-06 that scored close to 0, suggesting that those VLMs are as good as the average human respondent. Moreover, GPT-4-turbo-2024-04-09, Phi-3.5-Vision-Instruct, Llama3.2-11B-Vision-Instruct, and SmolVLM-Instruct performed below the average respondent suggesting they are not as good at the image-caption rating task, relative to the other respondents.

7 CASE STUDY 2: VISUAL READING COMPREHENSION

In Blum (2019), the author set out to investigate the impact of visual reading material on comprehension, more specifically, degree of information integration Blum et al. (2020). Notions of accessibility bring to question the perceived deficits in narrative comprehension for autistic people. This deficit has been positioned as having a cognitive processing disposition towards local inference (taken directly out of the material), rather than global inference (requiring greater integration of world knowledge). Instead of a unitary deficit in the individual, reduced performance on inferential narrative comprehension tasks was hypothesized to be an issue of modality; that is, does the presence of visual information impact the results? In the study, the impact of modality on inferential reasoning was compared between autistic and non-autistic adolescents.

The original study included $N = 130$ human respondents (18 autistic, and 112 non-autistic). The author constructed narratives that focused on moral and ethical issues involved in human relationships (e.g., friendship, stealing, lying). These narratives were presented in two formats: unimodal text-only and multimodal image plus text. They are both word-for-word the same, except one is a

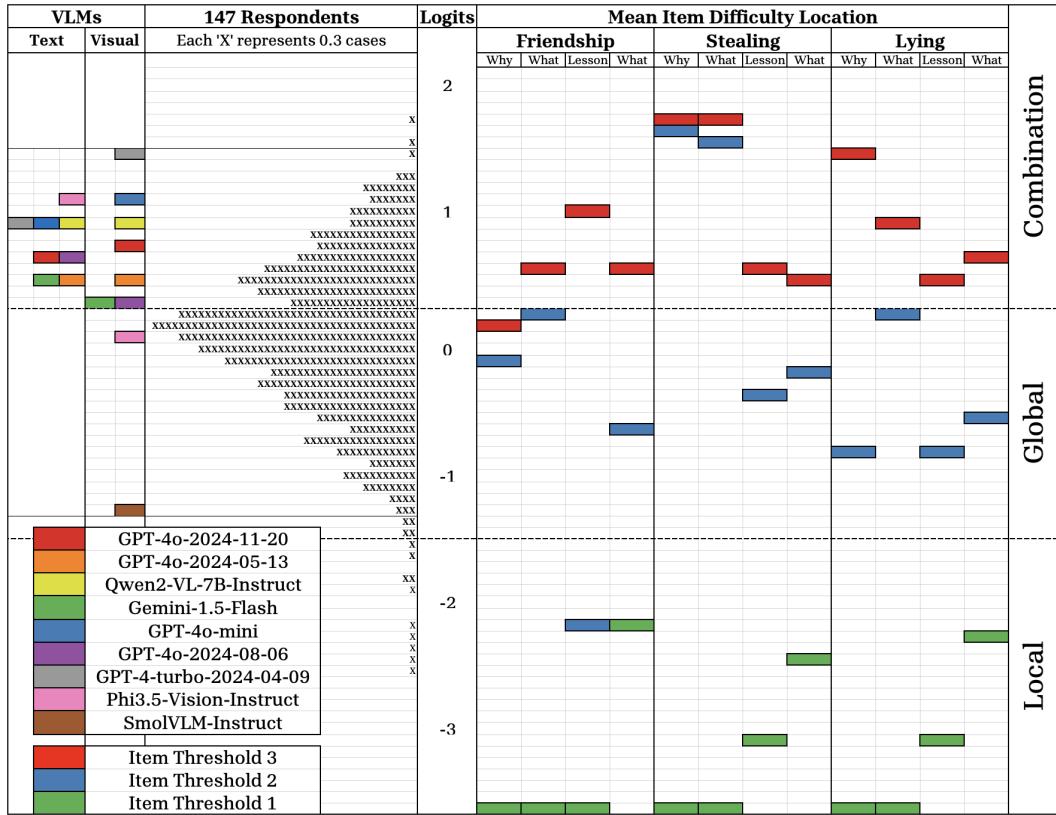


Figure 4: The Wright Map for the visual reading comprehension task. The column “147 Respondents” shows the distribution of the θ abilities at information integration for all respondents, humans and VLMs, and in both formats. The VLMs are pulled-out into their own column and format so they can be seen in isolation. The item-map on the right side shows the three narratives, and their corresponding item locations. There is fairly clear “banding” showing the validity of the latent ability’s internal structure.

text format and the other has the same words spread across panels within narration boxes, speech bubbles, and thought bubbles within the images.

We prompted the VLMs with the prompt in Appendix B to engage in the same task as the human respondents. For each narrative story, the prompt had the VLMs adopt the role of a grade-school student, had them “read” the narrative and then answer the subsequent chain of questions, presented in text-only and visual plus text formats. The three narratives were stories about a boy’s friendship with a cat, a boy stealing someone’s wallet for a bike, and a friend lying to protect their friend’s cheating.

The initial question asked of the respondents was situated in the narrative, and the subsequent questions delved into their reasoning. The questions involved were: 1. A motivational inference question about why a character engaged in an intentional action. 2. A meta-cognitive question, asking *What made them think of the answer?* 3. An evaluative inference question, asking *What lesson could someone learn from this story?* 4. Another meta-cognitive inference question, asking *What made them think of the answer?* Human and VLM responses were encoded as Local Inference, Global Inference, or a Combination of them both.

In the study, each human group was randomly assigned to a sub-group (groups A, B, C, and D). Groups A and B consisted of randomly assigned autistic individuals. Group A read the narratives in visual format. Group B read the narratives in text-only format. Groups C and D consisted of randomly assigned non-autistic individuals. Groups C read the narratives in visual format and group D read it in text-only. The VLMs were assigned both formats.

Vision Language Models	Ability (θ)	Percentile
GPT-4-Turbo-2024-04-09 (comic)	1.55903	100%
GPT-4o-mini (comic)	1.18214	97%
Phi-3.5-Vision-Instruct (text)	1.18214	97%
GPT-4-turbo-2024-04-09 (text)	1.00795	93%
GPT-4o-mini (text)	1.00795	93%
Qwen2-VL-7B-Instruct (comic)	1.00795	93%
Qwen2-VL-7B-Instruct (text)	1.00795	93%
GPT-4o-2024-11-20 (comic)	0.84269	88%
GPT-4o-2024-08-06 (text)	0.68559	82%
GPT-4o-2024-11-20 (text)	0.68559	82%
GPT-4o-2024-05-13 (comic)	0.53525	78%
GPT-4o-2024-05-13 (text)	0.53525	78%
Gemini1.5_flash (text)	0.53525	78%
GPT-4o-2024-08-06 (comic)	0.38988	67%
Gemini1.5_flash (comic)	0.38988	67%
Phi-3.5-Vision-Instruct (comic)	0.10637	48%
SmolVLM-Instruct (comic)	-1.25288	5%

Table 3: Evaluation of each VLM on the visual reading comprehension task. The columns labeled “Ability” and “Percentile” indicate the logit location for each model (its ability) and the percentage of human raters that the model matched or exceeded in ability.

7.1 IRT AND WRIGHT MAP ANALYSIS

From analyzing the output of the statistical inference, we find that the fit statistics were within acceptable range, and we could trust the validity of the Wright Map for this latent ability. The results of the IRT and Wright Map for the Visual Reading Comprehension study are shown in Figure 4. We see on the map that most VLMs are above average, compared to human respondents, and that most VLMs are in the same region of interest. Since they are all above the 0 logit, the results suggest that all of the VLMs are good at the comprehension task, regardless of modality, with the exception of SmolVLM and Phi3.5.

Although most of the VLMs are in a similar region, regardless of format, three of them stand-out as anomalous: GPT-4-Turbo-2024-04-09, Phi3.5-Vision-Instruct, and SmolVLM-Instruct. GPT-4-Turbo-2024-04-09 has the highest latent ability with the visual format part of the task, even exceeding the human respondents, but performed marginally lower on the text format part – but not so low as to be in a different level of information integration. Whereas, Phi-3.5-Vision-Instruct measured as the first highest in text-only format, but performed significantly lower in the visual format, placing at a different region. SmolVLM-Instruct performed far below average, compared to human respondents and other VLMs. It would not be a good candidate, relatively speaking, at the visual reading comprehension task.

8 DISCUSSION AND CONCLUSIONS

In this paper we showcase the evaluative and interpretative power of Item Response Theory for the ML and NLP communities by using 10 Vision Language Models on two case studies: image caption rating and visual reading comprehension. Traditional accuracy-focused metrics show the general overview of model performance but lack detail provided by considering item difficulty when assessing respondent ability, such as θ , δ , fit statistics, and the validity of internal structure of the latent ability.

By employing the Rasch model and Wright Map analyses, we demonstrated a comprehensive and informative evaluation strategy that extends beyond aggregate performance, providing valuable insights into the latent abilities of VLMs compared to human respondents. Through our IRT investigation, we were not only able to uncover how well each of the 10 VLMs performs on each task but also how hard the items are to the respondents. Incorporating the principles and practices of IRT in the NLP problem space for the ML Community could offer significant methodological advantages for validating both items and respondents. We hope that our work may prove a valuable bridge between the measurement and machine learning communities.

486 REFERENCES
487

488 Raymond J Adams, M L Wu, and Mark Wilson. ACER ConQuest: Generalised item response
489 modelling software, 2012. QWERT.

490 Google AI. Gemini 1.5 flash. <https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash>, 2024. Accessed: 2025-01-22.

491

492 Alexander Mario Blum. *Deficit or difference? Assessing narrative comprehension in autistic and*
493 *typically developing individuals: comic vs. text*. University of California, Berkeley, 2019.

494

495 Alexander Mario Blum, James M Mason, Jinho Kim, and P David Pearson. Modeling question-
496 answer relations: the development of the integrative inferential reasoning comic assessment.
497 *Reading and Writing*, 33(8):1971–2000, 2020.

498

499 Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong
500 Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and genera-
501 tion with better captions. *arXiv preprint arXiv:2406.04325*, 2024.

502 François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

503

504 A. Crudden and L. W. McBroom. Barriers to employment: A survey of employed persons who are
505 visually impaired. *Journal of Visual Impairment and Blindness*, 93(6):341–350, June 1999. doi:
506 10.1177/0145482X9909300602.

507

508 Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and
509 trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008.

510

511 Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for
512 any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pp.
513 376–380, 2014.

514

515 Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild,
516 Tianyi Zhou, Tom Goldstein, John Langford, Animashree Anandkumar, et al. Easy2hard-bench:
517 Standardized difficulty labels for profiling llm performance and generalization. *Advances in Neu-
518 ral Information Processing Systems*, 37:44323–44365, 2024.

519

520 Denis Federiakin. Improving llm leaderboards with psychometrical methodology. *arXiv preprint*
521 *arXiv:2501.17200*, 2025.

522

523 Hatrice Gurdil, Yesim Beril Soguksu, Salih Salihoglu, and Fatma Coskun. Integration of artificial
524 intelligence in educational measurement: Efficacy of chatgpt in data generation within the scope
525 of item response theory. *arXiv preprint arXiv:2402.01731*, 2024.

526

527 Yoshee Jain, John Hollander, Amber He, Sunny Tang, Liang Zhang, and John Sabatini. Explor-
528 ing the potential of large language models for estimating the reading comprehension question
529 difficulty. In *International Conference on Human-Computer Interaction*, pp. 202–213. Springer,
530 2025.

531

532 Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descrip-
533 tions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.
534 3128–3137, 2015.

535

536 Walter Kintsch. The role of knowledge in discourse comprehension: a construction-integration
537 model. *Psychological review*, 95(2):163, 1988.

538

539 C. Kirchner and B. Smith. Transition to what? education and employment outcomes for visually
540 impaired youths after high school. *Journal of Visual Impairment and Blindness*, 99(8):499–504,
541 August 2005. doi: 10.1177/0145482X0509900806.

542

543 John P. Lalor, Pedro Rodriguez, Joao Sedoc, and Jose Hernandez-Orallo. Item response theory for
544 natural language processing. In *Proceedings of the 18th Conference of the European Chapter of*
545 *the Association for Computational Linguistics: Tutorial Abstracts*, Malta, 2024a. Association for
546 Computational Linguistics.

540 John P Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. Item response theory for
 541 natural language processing. In *Proceedings of the 18th Conference of the European Chapter of*
 542 *the Association for Computational Linguistics: Tutorial Abstracts*, pp. 9–13, 2024b.

543

544 John Patrick Lalor and Pedro Rodriguez. py-irt: A scalable item response theory library for python.
 545 *INFORMS Journal on Computing*, 35(1):5–13, 2023.

546 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization*
 547 *branches out*, pp. 74–81, 2004.

548 Yunting Liu, Shreya Bhandari, and Zachary A Pardos. Leveraging llm respondents for item eval-
 549 uation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052,
 550 2025.

551

552 Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. IAP, 2008.

553

554 M. C. McDonnell and Z. Sui. Employment and unemployment rates of people who are blind or
 555 visually impaired: Estimates from multiple sources. *Journal of Visual Impairment and Blindness*,
 113(6):481–492, November 2019. doi: 10.1177/0145482X19887620.

556

557 R. McMillen and F. Alter. Social media, social inclusion, and museum disability access. *Museum*
 558 *and Society Issues*, 12(2):115–125, July 2017. doi: 10.1080/15596893.2017.1361689.

559

560 Inc. Meta Platforms. Llama-3.2-11b-vision-instruct. [https://huggingface.co/](https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct)
 561 meta-llama/Llama-3.2-11B-Vision-Instruct, 2024. Accessed: 2025-01-22.

562

563 Microsoft. Phi-3.5 vision instruct. [https://huggingface.co/microsoft/Phi-3.](https://huggingface.co/microsoft/Phi-3.5-vision-instruct)
 564 5-vision-instruct, 2024. Accessed: 2025-01-22.

565

566 Rasulov Ulugbek Murodulloyevich. Analysis of test systems used in classical and modern test
 567 theories for assessing students' knowledge. *International Journal of Pedagogics*, 5(03):150–154,
 568 2025.

569

570 Lothar D Narins, Andrew Scott, Aakash Gautam, Anagha Kulkarni, Mar Castanon, Benjamin Kao,
 571 Shasta Ihorn, Yue-Ting Siu, James M Mason, Alexander Blum, et al. Validated image caption
 572 rating dataset. *Advances in Neural Information Processing Systems*, 36:61292–61305, 2023.

573

574 OpenAI. Openai models documentation. [https://platform.openai.com/docs/](https://platform.openai.com/docs/models)
 575 models, 2025. Accessed: 2025-01-22.

576

577 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic
 578 evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association*
 579 *for Computational Linguistics*, pp. 311–318, 2002.

580

581 Georg Rasch. Studies in mathematical psychology: I. probabilistic models for some intelligence
 582 and attainment tests. 1960.

583

584 Hugging Face TB Research. Smolvlm-instruct. [https://huggingface.co/](https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct)
 585 HuggingFaceTB/SmolVLM-Instruct, 2024. Accessed: 2025-01-22.

586

587 Soujatya Sarkar, Manikandan Ravikiran, and Rohit Saluja. Scorecliq: A dynamic llm-based frame-
 588 work for item difficulty estimation. In *International Conference on Artificial Intelligence in Edu-
 589 cation*, pp. 169–176. Springer, 2025.

590

591 Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan.
 592 Smart: Simulated students aligned with item response theory for question difficulty prediction.
 593 *arXiv preprint arXiv:2507.05129*, 2025.

594

595 Mara Schilling-Wilhelmi, Nawaf Alampara, and Kevin Maik Jablonka. Lifting the benchmark ice-
 596 berg with item-response theory. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

597

598 Andrew Taylor Scott, Lothar D Narins, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta
 599 Ihorn, Yue-Ting Siu, and Ilmi Yoon. Improved image caption rating–datasets, game, and model.
 600 In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pp.
 601 1–7, 2023.

594 James Sharpnack, Phoebe Mulcaire, Klinton Bicknell, Geoff LaFlair, and Kevin Yancey. Autoirt: Calibrating item response theory models with automated machine learning. *arXiv preprint arXiv:2409.08823*, 2024.

595

596

597

598 A. Shaw, D. Gold, and K. Wolffe. Employment-related experiences of youths who are visually impaired: How are these youths faring? *Journal of Visual Impairment and Blindness*, 101(1): 599 7–21, January 2007. doi: 10.1177/0145482X0710100103.

600

601

602 Elliot Smith and Peter Hancox. Representation, coherence and inference. *Artificial Intelligence Review*, 15:295–323, 2001.

603

604

605 Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. Irt-router: Effective and interpretable multi-lm routing via item response theory. *arXiv preprint arXiv:2506.01048*, 2025.

606

607

608 Qwen Team. Qwen2-vl-7b-instruct. <https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct>, 2024. Accessed: 2025-01-22.

609

610

611 Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

612

613

614 Hariram Veeramani, Surendrabikram Thapa, Natarajan Balaji Shankar, and Abeer Alwan. Large language model-based pipeline for item difficulty and response time estimation for educational assessments. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 561–566, 2024.

615

616

617

618 Leonie VDE Vogelsmeier, Eduardo Oliveira, Kamila Misiejuk, Sonsoles López-Pernas, and Mohammed Saqr. Delving into the psychology of machines: Exploring the structure of self-regulated learning via lm-generated survey responses. *arXiv preprint arXiv:2506.13384*, 2025.

619

620

621

622 M Wilson. Some notes on the term:“wright map.”. *Rasch Measurement Transactions*, 25(3):1331, 2011.

623

624

625 Mark Wilson. Things I learned from ben. In Mark Wilson, Fisher, Jr., and William P. (eds.), *Psychological and Social Measurement: The Career and Contributions of Benjamin D. Wright*, chapter 8, pp. 75–81. Springer International Publishing, Cham, 2017.

626

627

628 Mark Wilson. *Constructing measures: An item response modeling approach*. Routledge, 2023.

629

630

631 Mark Wilson and Paul De Boeck. Descriptive and explanatory item response models. In *Explanatory item response models: A generalized linear and nonlinear approach*, pp. 43–74. Springer, 2004.

632

633 Benjamin D Wright. Rating scale analysis. *Measurement, Evaluation, Statistics, and Assessment Press*, 1982.

634

635 Margaret Wu and Richard J Adams. Properties of rasch residual fit statistics. *Journal of Applied Measurement*, 14(4):339–355, 2013.

636

637

638 Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*, 2025.

639

640

641 Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *arXiv preprint arXiv:2505.15055*, 2025.

642

643

644

645 Yan Zhuang, Qi Liu, Zachary Pardos, Patrick C Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang, and Enhong Chen. Position: Ai evaluation should learn from how we test humans. In *Forty-second International Conference on Machine Learning Position Paper Track*.

646

647

648 **A APPENDIX A: CASE STUDY 1 PROMPT**
649650 The “<Image>” and “<Caption>” tags are replaced with the base64 encoded image, and the cap-
651 tion.
652653
654 You are an expert at determining how good a description is for an image. You are able to
655 rate the given description and image on a scale of 1 to 5, corresponding to how well the
656 description fits the image. Using the following rubric of rating levels:657
658 1. Objects are incorrectly identified. The caption gives the wrong idea about what is
659 happening in the image.
660 2. Objects are partially correctly identified with some errors, but the caption is accu-
661 rate enough to give an idea of what is happening in the image. The caption identifies
662 most of the objects but might not identify everything. There is no interpretation of
663 what anything means.
664 3. Relevant objects are correctly identified. The caption describes what is seen but
665 not where objects are in space. There is no description of the overall setting and no
666 interpretation of an event.
667 4. Objects and/or a general scene and/or an action are correctly identified but not
668 every element is completely identified. The caption describes what is seen and
669 where things are in space. There is no interpretation of an event.
670 5. Objects, a general scene, and actions are correctly identified if present in the image.
671 The caption describes what is seen and where things are in space. Interpretation of
672 overall setting and/or event is included.673 Please rate the following description on how well they describe the given image. Just the
674 number (ratings from 1 to 5).675 <Image>
676 <Caption>
677680 **B APPENDIX B: CASE STUDY 2 PROMPT CHAIN**
681682 The prompt used for Case Study 2 was a scripted, multi-prompt. First we would prompt with the
683 initial context, ask a question, wait for response, and then asked another question. Here is the chain
684 of questions.
685686
687 You are a middle school or high school student. Please answer the following questions
688 based on the given narrative in 1 sentence.
689

690 <Comic Panels>

691 **Question:** Why did the red headed boy keep the cat?
692 **Assistant:** <Response 1>693
694
695 ↓
696697
698 **Question:** What made you think of that answer?
699 **Assistant:** <Response 2>700
701 ↓

702

703

Question: What lesson could someone learn from this story?

Assistant: <Response 3>

704

705

706

707

↓

708

709

Question: What made you think of that answer?

Assistant: <Response 4>

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755