Testing the Depth of ChatGPT's Comprehension via Cross-Modal Tasks Based on ASCII-Art: GPT3.5's Native Abilities in Regard to Recognizing and Generating ASCII-Art Are Not Totally Lacking

David Bayani[000-0001-5811-6792]
Inpleo, Inc
david.bayani@inpleo.com

Abstract

In the months since its release, ChatGPT and its underlying model, GPT3.5, have garnered massive attention, due to their potent mix of capability and accessibility. While a niche industry of papers have emerged examining the scope of capabilities these models possess, language — whether natural or stylized like code — has been the vehicle to exchange information with the network. Drawing inspiration from the multi-modal knowledge we'd expect an agent with true understanding to possess, we examine GPT3.5's aptitude for visual tasks, where the inputs feature ASCII-art without overt distillation into a lingual summary. In particular, we scrutinize its performance on carefully designed image recognition and generation tasks.[1]

1 Introduction

ChatGPT has rapidly been adopted since its release in November 2022. This large language model (LLM) builds off of version 3.5 of the Generative Pre-trained Transformer model family developed by OpenAI, a child whose lineage has been marked by one massive step after another in regard to the size of LLM networks and their training data. Active utilization in industry (Marr, 2023) and education (Brown, 2023) are already a reality, though with growing concerns on the impacts on the workforce and academic integrity. Fueled by the model's unprecedented popularity, accessibility, and power, a niche industry of papers attempting to rigorously investigate the abilities of ChatGPT — and the GPT3/GPT3.5 family underlying it more broadly — have materialized in short order. However, efforts thus far have almost exclusively focused on language-centric tasks (Liu et al., 2023). Filling this gap, we explore GPT3.5's abilities to "see" and "draw" — critically, doing so without first summarizing the inputs into a verbal description for the model. Our vehicle in order to conduct this analysis is ASCII-art (AArt) (O'Riordan, 2014). Ultimately, GPT3.5 demonstrates noticeable visual acumen. We uncover that GPT3.5 has subtly more vision-related acumen than has been appreciated.

2 Related Work

Most work on ChatGPT has considered canonical NLP problems (Zhang et al., 2023; Liu et al., 2023; Zhong et al., 2023). As pointed out in (Liu et al., 2023), ChatGPT's diverse capabilities and accessibility have fueled a deluge of papers exploring its potential and limitations. The model has proven performant in areas ranging from poetry (Cushman, 2022) to programming (Sadik et al., 2023) to verbally-enabled room navigation (Joublin et al., 2023). Within this space, most relevant to us are efforts treating ChatGPT's spatial reasoning, as well as those exploring its integration into multi-component pipelines geared toward text-based image recognition, manipulation, or generation.

Both (Deshpande and Szefer, 2023) and (Zhang et al., 2023) — examining, respectively, the network's performance in an introductory engineering course and from surveying across the literature — observed limitations in GPT3.5's abilities to handle "diagrams or figures" and to "perform spatial, temporal, or physical inferences". Muddying their conclusions, however, are a subset of reported instances where the network produced AArt— but with major qualifiers of being rare and generic enough to likely be rote memorization.

There have been attempts to integrate recent GPT-family models into VQA (Yang et al., 2022; Bongini et al., 2022; Si et al., 2023; Yang et al., 2022; Chalvatzaki et al., 2023; Tiong et al., 2022; Li et al., 2023; Huang et al., 2023; Mu et al., 2023; Srivastava et al., 2023b), image generation (Yang

[1] An extended version of this write-up is available at: https://arxiv.org/abs/2307.16806.

Findings of the Association for Computational Linguistics: EACL 2024, pages 2063–2077
March 17-22, 2024 ©2024 Association for Computational Linguistics

et al., 2023; Maddigan and Susnjak, 2023; Nanwani et al., 2023; Qin et al., 2023; Todd et al., 2023), graph analysis (ex., layout descriptions, scene graphs, etc.) (Zhang, 2023; Wang et al., 2023a; Guo et al., 2023; Shi et al., 2023; Zhu et al., 2023; Bartolomeo et al., 2023), and in other problem settings where visual-content could play either input or output roles (Shen et al., 2023; Wu et al., 2023). The diversity of implementation-specifics notwithstanding, the takeaways are largely the same: these works either (1) prior to querying the LLM, summarize context verbally or in a human-readable data structure via different foundation model specifically engineered for image-related tasks or (2) modify the language models in question to explicitly include visual knowledge, often coupling this with additional training of parts that are woven intimately into the LLMs. For our purposes, adopting either strategy disqualifies a work from bearing fully on our main question. That is, many existing works simply "let GPT3.5 see" by either modifying it to the point of being a fundamentally different model, or giving it a seeing-eye dog (i.e., another foundation model that addresses all the seeing and manipulation). Each of (Ye et al., 2023), (Chen et al., 2023), and (Joublin et al., 2023) examine GPT3.5's spatial reasoning, navigation, and interaction tasks, but yet again all exchanges were mediated through verbal descriptions of the world state and action space, though sparing the use of a separate foundation model to produce the words. We comment further on the nature of this distinction in Appendix A.

The aforementioned aside, some substantive works exist that somewhat resonate with our work.

Under the impetus of differentiating content generated by ChatGPT versus humans, (Wang et al., 2023b) curated questions that emphasized the areas where LLMs' aptitude most differed — for better or worse — from that of a human. Among the eight tests considered, identification of AArt was one, exposing a patent gap between human and ChatGPT performance — 94% and 8% accuracy respectively on 50 cataloged drawings. In addition to the limited show-verbatim-and-describe nature of these trials, we highlight that all samples came from a public website existent for years before ChatGPT's release, the ASCII Art Archive,[2] risking membership in GPT3.5's training data; moreover, the images' online popularity may predate their inclusion in the catalog. While 8% accuracy is not astounding, it is not nothing; questions remain as to how much is from memorization, actual recognition ability, and random chance.

Under similar inspiration, the massive, collaborative effort of the "BIG Benchmark" (Srivastava et al., 2023a) showcases 204 diverse tasks examining language model's capabilities. Three such tasks nominally featured AArt, but all concerned the recognition of text that was "rendered" in that fashion. The only other germane task we saw was the "text navigation game",[3] which featured a small input grid containing an AArt "maze", requiring the models to verbally specify moves from the start to the goal; no instances of "success" were observed by any model for board sizes above 5-by-5, and moreover the authors made reference to success rates on smaller boards being on par with random movement. Overall, we find a lack of sufficient subtlety in the benchmark's pertinent tasks, them failing to be sensitive enough — at least as explored — to detect all but the most obvious performance. Furthermore, probing specific to evaluating vision systems — such as robustness to rotation, noise or translation — were not carried out, leaving insights only at the high-level outcomes of the raw tests. Both of these aspects help distinguish our work from theirs, not to mention the fact that we examine generation of visual content in addition to its recognition.

Like us, (Dabkowski and Begus, 2023) study capabilities of OpenAI's GPT model family, version 3.5 and 4 in their case,[4] using a series of prompts without additional training or system modifications. Their endeavor partially examined rudimentary AArt produced to explore the models' recursive generation abilities — however, whether this is a "visual" task or essentially an algebraic computation is debatable. The authors note that certain examples displayed are likely memorized from training data, but also point out (rightfully) that the more exotic figures produced are less subject to this concern. Either way, the concern underscores the fact that their prompts for AArt did not (obviously) impose novelty-constraints on the output, thus failing to rule out preprepared responses as "correct"

[2] https://web.archive.org/web/20180305160309/https://www.asciiart.eu/.

[3] https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/text_navigation_game

[4] Note that GPT4 is not relevant to our focus since that model was explicitly designed to include visual processing.

outputs. In contrast, out experiments require responses to correspond with unique, freshly generated structures provided in our prompts, reducing the feasibility of context-independent, pre-canned responses passing scrutiny.

Finally, we remark on a certain degree of "folk knowledge" about ChatGPT's drawing abilities — for instance (Wetrorave, 2022; Arora, 2022; Blocks, 2022). However, exchanges in this category directly featuring AArt (i.e., not code for diagrams, etc.) were mostly sporadic acts, not systematic or deep explorations. A theme throughout is the appearance of AArt of reasonable quality, but occurring at inappropriate times in respect to the prompts — hallmarks of shallow memorization, repeating training examples without deeper, semantically-meaningful interpretation or modification. As a result, the casual consensus judges ChatGPT's abilities in this regard as poor. We endeavor to perform more rigorous analysis than the loose folk-perceptions.

3 The ASCII-Art Used in Experiments

We use AArt of box diagrams (AADs) to nontrivially probe GPT3.5's vision-related capabilities. We briefly share the inspiration for this particular choice, since we believe the observations valuable:

First, we realized that AADs are used as illustrations in many settings — e.g., electrical-circuit diagrams, placement charts, and flowcharts online. Indeed, mini-languages like PIC (Kernighan, 1982) exist to aid their creation, though manual drawing is rarely difficult. GPT3.5 may therefore have a substantial amount of varied training data available for these drawings, e.g., as part of Common-Crawl.[5] Additionally, owing to their common role as a visual aids accompanying verbal descriptions, these depictions likely have appreciable amounts of granular visio-lingual coupled data.

Second, we encountered quite promising results during early investigations into ChatGPT's germane abilities. In a trial, we requested drawings of several town layouts, each with certain buildings and accompanying labels. Illustrations were generated that matched our specification, a feat not easily dismissed as mere memorization. Reasonable success continued during additional requests (e.g., for roads) that followed.

Following these leads, we have run experiments featuring randomly generated AADs to gauge ChatGPT's aptitude in typical vision-related tasks: content recognition despite changes due to rotation, scale, "pixel" noise, and translation. If GPT3.5 can handle these tasks, then it suffices to say it is not *entirely* incapable of "doing well at AArt", despite impressions held in folk knowledge.

3.1 Generation of AADs

Our AADs start with a blank 24-by-24 character canvas to which boxes are progressively added. Per box, five values are needed: two values per lower-left and top-right vertex — all constrained to stay on canvas — and a name comprised of a single ASCII alphanumerical character which is optionally displayed. A box is added after two-phases: proposal then, as needed, rejection.

During proposal, a start position and length are chosen for each axis independently, the former uniformly over the canvas, the latter via draw from a Poisson distribution. Using $\lambda = 8$ for the Poisson made reasonable illustrations with an appealing variation in layout and complexity — for instance, results can range from well-aligned rows of roughly uniform boxes, to nested complexes arranged in a scattered fashion. Lengths are required to be at least 3 — the minimum to fit a name and boundary lines — and are resampled until then.

In the rejection phase, we throw out boxes that run off the canvas or overlap existing boxes. Additionally, to reserve space for potential names, we reject boxes that are tightly nested in the corner of another box. Upon rejection, we sample a new box until either 1000 tries have failed or 14 boxes are established, after which results go forward to the next phase.

Having abstractly determined box placement, we place characters to reflect it. We attempt reasonable diligence in ensuring the network cannot cheat through trivial illustration artifacts. As one precaution, for experiments that require comparing multiple AADs, each option has the same number of each type of character present.[6] Proposed AADs that fail to have character counts matching earlier drawings are rejected; we then make another generation attempt. To improve acceptance rates, we clip box lengths post-draw to ensure that the number of proposed characters never exceeds constraints.

Box boundaries are drawn using dashes ("-") for the horizontal (x) length and pipe symbols ("|") for the vertical (y) length. We considered adding "+" at

[5] https://commoncrawl.org/big-picture/

[6] An exception in some trials being added noise characters.

vertices, but character-matching constraints would then require all drawings to have an equal number of boxes — a needless restriction on the possible outcomes. Instead, corners are left unfilled.

By default, we pad the right-margin of the AArt with spaces so that all lines are the same length, the alternative having been to leave the right-edge ragged. We choose this default since, on balance, the added uniformity boosts our confidence that any positive outcomes are not the result of leveraging non-visible structure, e.g. a unique right-edge. Also, we suspect that this provides the best chance for the model to demonstrate any ability it truly has, it not having to contend with additional environmental instability.[7]

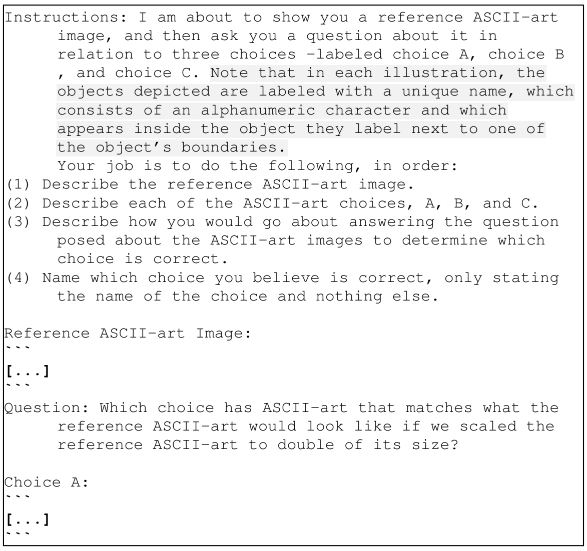
Names are drawn inside boxes in one corner selected at random. Within an experiment trial, if we show multiple AADs (e.g., in Section 4) each drawing must use the same set of names, a fact also requiring that the number of boxes in each picture match. The assignment of names to boxes is randomized. By default, names are not in AADs, since lack of such identifiers should increase difficulty; while we do not want to set the model up for failure, we deemed this a reasonable difficulty-threshold to start with for the inherently easier tasks (looking ahead: image recognition versus generation) which we can relax should the barrier prove too high to detect anything non-trivial.

We overview our experiments next. In addition to the below, we ran trials to verify that GPT3.5 was performant at recognizing and generating provided AArt ***verbatim***; this sanity-check was of interest since the LLM was not trained to handle large sections of such non-lingual content. Results for those trials were near perfection and largely as hoped — thus, to respect space, we limit their discussion to this note.

4 Recognition Experiments

4.1 Setup

We ran experiments to gauge GPT3.5's native image recognition abilities. The model was given a prompt displaying a reference AArt, followed by a request to select from among three randomly-ordered choices one depiction that corresponds to the reference in a way matching the prompt. While one can imagine trials where multiple options are based on the reference but only one corresponds to the correct transform — for example, each being a different rotation, with the goal to find the 90°turn — it is imprudent to start with such added difficulty. Overall, we are interesting in judging GPT3.5's ability to identify an image after it has undergone typical vision-related changes — e.g., translation, enlargement, rotation, etc. If it is unable to succeed when only one option is derived from the reference art, then it seems reasonable to suppose having more derived choices would cause performance to degrade even further.



~~~
Instructions: I am about to show you a reference ASCII-art
      image, and then ask you a question about it in
      relation to three choices -labeled choice A, choice B
      , and choice C. Note that in each illustration, the
      objects depicted are labeled with a unique name, which
      consists of an alphanumeric character and which
      appears inside the object they label next to one of
      the object's boundaries.
      Your job is to do the following, in order:
(1) Describe the reference ASCII-art image.
(2) Describe each of the ASCII-art choices, A, B, and C.
(3) Describe how you would go about answering the question
      posed about the ASCII-art images to determine which
      choice is correct.
(4) Name which choice you believe is correct, only stating
      the name of the choice and nothing else.

Reference ASCII-art Image:
```
[...]
```
Question: Which choice has ASCII-art that matches what the
      reference ASCII-art would look like if we scaled the
      reference ASCII-art to double of its size?

Choice A:
```
[...]
```
~~~

Figure 1: The prompt we used for recognition experiments that featured scaling. AArt would be placed where the bolded, bracketed ellipsis (**[...]**) are shown. In the limits of space, we display only Choice A; Choice B and Choice C follow the same pattern, going to the end of the prompt. The highlighted text is only present for experiments that label AADs with names.

Taking the cue from Chain-of-Thought (CoT) Prompting (Wei et al., 2022), we asked the model warm-up questions to facilitate examination of the AArt provided, build up focus towards facets of the depiction pertinent to the main query. See Figure 1.

Queries are issued once for each prompt using OpenAI's API for `gpt-3.5-turbo` with no additional context maintained between calls. Responses are drawn with a temperature of zero, since the space of correct answers is small. Despite this temperature, preliminary trials showed that responses were meaningfully diverse, including differences in response to the main question. We query once per prompt since that suffices to produce the statistics of interest, and also avoids de-

[7] I.e., if performance is good, we may have more trust cheating did not occur, and if it is poor, we may have greater confidence that the model categorically lacks those abilities.

pendencies that would muddy interpretation.

Responses we received reliably had answers located next to their corresponding subquestion number, for instance, *"(1) The reference looks like[...](2)[...](3) To determine which, I would[...](4) The answer is Choice A because [...]"*. Basic string parsing (e.g. regular expressions) was able to consistently extract the primary response (i.e., which option corresponds to the reference); see Appendix C for more comments in this regard.

In most cases, our prompts did not give any information about the AArt's content, either in terms of the objects shown (boxes) or the meaning of characters. For instance, in trials involving (geometric) translation, we only ask which option matches the target if it was shifted horizontally or vertically — we do not indicate the amounts shifted. Additional details are at Appendix B.

4.1.1 Matching After Translation

To test the model's ability to match images after translation, we embed our AArt into a larger canvas and pick a random position for the inner-canvas's bottom-left corner. Specifically, the larger canvas is 48-by-48 and the offset is drawn from Uniform($\mathbb{Z} \cap [0, 23]$) for each dimension.[8] We force the offsets for the reference image and the correct choice to be different, ensuring all queries are nontrivial. We place no such constraints on the other choices.

4.1.2 Matching After Rotation

For rotation, we have the reference image undergo a 90°clockwise turn. Early trials suggested that this task is difficult, which is unsurprising since the transform changes character locations in a fashion atypical for prose. Attempting due diligence in detecting any aptitude GPT3.5 has for this task, we tried several settings of the drawings' side-length (s), maximum number of boxes (B) and Poisson parameter (λ), specifically $(s, B, \lambda) \in \{(24, 14, 8), (15, 9, 5), (8, 5, 3)\}$. These settings reflect scaling the values to 1.0 (the default), roughly 0.6, and roughly 0.3; Table 1 refers to them as such. Under the same motivation, trials were carried out with box names present.

4.1.3 Matching Despite Noise

Images commonly have pixel noise — small-scale, random alterations that are neither attributed to obvious geometric transforms nor are semantically impactful. Investigating GPT3.5's robustness to this ubiquitous phenomenon, we inject randomly drawn characters into the AArt— both the reference and, sampled independently, each choice — then ask the LLM to find the match. We use a small set of otherwise unused ASCII special characters as noise elements,[9] and place them where spaces initially were. By only replacing whitespace, we ensure that a drawing's main structures are unambiguously visible, preventing critical information loss that could otherwise set the model up to fail.[10]

We use two noise levels: 0.04 — that for each space, there is a 4% chance that it will be replaced by a noise character — and 0.32. We repeat the injection process until at least one noise character is added. In combination with this, we experiment with either the default padding (i.e., guaranteed 24 characters per line) and maximum number of boxes (14), or with a ragged right-edge and at most six boxes; this explores the performance impacts of additional variation in token structure combined with "less signal" due to fewer boxes.

4.1.4 Matching After Rescaling

Image recognition requires detecting a pattern despite changes in its scale. To study this, we generate AArt at half its typical size then decide to display either the reference or the choices, but not both, at double their initial size; the choice of which is a parameter. The initial art generated has a 12-by-12 canvas, at most 7 boxes, and λ of 4; when enlarged, the canvas is the standard 24-by-24 size. In addition to choosing the target of scaling, we examine the impact of naming boxes, resulting in a total of four different experiment settings.

4.2 Results

In Table 1, we list the observed accuracy for each setting and $\alpha = 5\%$ Clopper-Pearson confidence intervals (CIs) on them. Random guessing would have an expected performance of 33.3%.[11] We see that all raw observations exceed this measure save one, and the majority of CIs are strictly above it.

While we did not made family-wise significance corrections to the individual intervals, given the

[8] GPT3.5's tokenizer captures whitespace verbatim — e.g., newlines and multi-spaces are not substituted out.

[9] Specifically, chars in the set: $\{", @, *, ., ,\}$.

[10] Though a somewhat fanciful comparison, an analogous requirement is that adversarial injections to modern CV systems do not, to humans, add overt changes (Eykholt et al., 2018; Khalid et al., 2021).

[11] A one-sided hypothesis test based on our CIs would have a significance-level of $\alpha/2$, which is *more* conservative (rejects the null less often) than a $\alpha = 5\%$ test.

12 independent CIs of $\alpha \leq 0.05$, the probability that three or more fail to contain the parameter is less than a threshold of 5% (in fact $< 2\%$); this and the fact that 7 CIs are strictly above $\frac{1}{3}$ — the performance if purely guessing — support the idea that the figures are not purely the outcome of guessing, aiding the notion that GPT3.5 does have some acumen for distinguishing between AADs.

We observe an appreciable performance boost for translation, which we speculate results from prose often being indented, thus making it likely that the training set had many pertinent examples. Also, for English, whitespace rarely carries semantic value, thus making it more obviously ignorable.

Our results also do suggest that, all else equal, recognition is aided by the presence of names and more boxes with uniform padding to the right margin — however, this should be taken with reservation, since the CIs overlap in the comparisons. With a similar caveat, performance degrades with higher noise levels, as one would expect, while (less reservedly) AAD size does not obviously impact accuracy on rotation. Additionally, we notice that when the choices in the rescaling-trials are enlarged, the raw performance drops, though comparable CIs continue to intersect.

Exp.	Params	GPT3.5 Acc. (%)		Sample Size
		Obs.	CI, $\alpha = 0.05$	
Rotat.	scaling: 0.3	34.0	[29.4, 38.9]	397
	scaling: 0.6	35.2	[30.5, 40.1]	395
	scaling: 1.0	34.5	[29.8, 39.4]	397
Tr.	—	90.5	[87.2, 93.2]	399
Scale	ref., -name	39.6	[34.8, 44.7]	396
	ref., +name	42.4	[37.5, 47.4]	401
	cho., -name	31.5	[27.0, 36.3]	400
	cho., +name	38.0	[33.2, 43.0]	400
Noise	0.04, +pad.	44.0	[39.0, 49.0]	398
	0.04, -pad.	42.1	[37.2, 47.1]	399
	0.32, +pad.	40.5	[35.6, 45.5]	398
	0.32, -pad.	39.9	[35.0, 44.9]	396

Table 1: Results for recognizing AADs. $+$ or $-$ indicate, respectively, presence or absence; "pad." stands for padding and "name" for names. In the parameters, "ref." indicates the reference was shown at 24-by-24 scale and the options where 12-by-12, while "cho." means the reverse assignment of sizes. For noise trials, 0.04 and 0.32 indicate the noise level.

5 Generation Experiments

We examine GPT3.5's ability to generate AArt, tasking it to transform input images as specified.

5.1 AArt Used and Queries Issued

To access the model's AArt generation abilities while anchoring to something we can access, we follow a modification of the prompt-with-image-reference scheme detailed in Section 3.1 and 4.1, using the same process to form the references. Again leveraging CoT reasoning, we issue warm-up questions leading to the ultimate request. We tried to avoid revealing excessive, step-by-step instructions in order to better gauge the degree to which GPT3.5 already had a notion of what our queries involved; nonetheless, some transforms required more details than others to be specified unambiguously and in reasonably pithy ways. See Appendix D for the prompts used in this section.

Before proceeding, we detail the parameters used in generating experiments. In contrast to most earlier probing (e.g., Section 4), *all* AADs in this section contain name labels. This was motivated by the belief that (1) the generation task is inherently harder than the recognition task, and (2) providing names to anchor and minimally queue GPT3.5 as to structure would reduce the chance of "missing interesting behavior" by setting the LLM up for failure (i.e., starting with unnecessary difficulty). For translation we asked the model to return the image without the extras spaces (we explicitly stated it this way), and for the rescaling-trials, we displayed a half-size image and tasked the model to scale it up by two. Noise trials were conducted at the 0.04 level with padding retained, and rotations were done at size 1.0; see Table 1. Unlike in the recognition experiments, we informed GPT3.5 of what characters were non-noise, as can be seen in Figure 3c.

As before, the network's output was consistently structured well enough to extract content automatically with simple string parsing and lightweight heuristics. More details are in Appendix C.

In order to get a sense of GPT3.5's behavior on these tasks, we manually examined outcomes from randomly generated queries for each of the transforms under analysis. While we considered judging "correctness" with more ridged and mechanical approaches,[12] we observed that GPT3.5 did not simply fail or succeed at tasks, but appreciably often generated content along an orthogonal axis, where the outputs were not wrong per se, but also were not quite what we envisioned. Notwithstanding

[12] Ex: AuROC of a simple model's distance measure between generated content, expected results, and alternatives.

refinements to the prompts we attempted to narrow conceivable ambiguity after observing this behavior during preliminary investigations, the potential for meaningful nuances warrants the examination by a reasonably context-informed human.

In the rest of this section, we summarize the outcomes on 30 randomly selected queries per transform, and attempt to give a sense of successes, difficulties, and curiosities. As with the recognition experiments, our focus will be on the final outcome, which here is the AArt returned by the network, not the verbal responses provided in reply to our CoT prompting preceding it. Figure 2 shows examples of middle-grade outcomes from each of the experiments we run; they are neither the best nor the worst instances observed, but are in the representative "middle", illustrative of general trends.

5.2 An Overall Trend: No Hallucinations

Across our experiments, we observed that GPT3.5 did not invent nonexistent box names; for some experiments, while names may be *lost*, there did not appear to be "hallucinations" (Ji et al., 2023) where names not present in the reference were newly added. In respect to entire boxes, while some trials showed duplication or templating from the reference content (ex, Figure 2a), boxes by and large did not seem invented whole cloth. Given general concerns of LLMs concocting answers, this "honesty" in respect to the reference is worth noting.

5.3 Translation Trials

On the whole, translation results showed a mixed success, instances spanning from near perfection, to irrelevant output, and everywhere between. Only 8 cases had seemingly random code or prose mixed with the art, of which only 3 had images failing to clearly reflect the reference. Most commonly, excess whitespace on the periphery was trimmed, as desired. This success was tempered by certain "failure modes," namely loss of boxes, distortion of inner-distances, or muddling of box boundary alignments. In all such cases, remnants of the reference image remained clearly visible, with a minimum of one to two boxes intact. Finally, we noted 3 results very close to perfect, preserving the boxes almost exactly (a few boundaries were mildly misaligned) and performing close to the full translation desired (all having $\leq$ 2 extra left-aligned spaces), while 2 others retained the image structure, but kept excess left-padding. Overall, while the network was not spot-on completing this task, some nontrivial achievement of the visual manipulation requested was witnessed.

5.4 Noise Trials

Over the 30 noise trials studied, results tended to be reasonable but incomplete or mildly flawed. As to reasonableness, unlike the "squashing" or loss of boxes that occurred in a number of translation trials, the result boxes aligned with the reference, save a minority of rows that on occasion were visibly shifted, more often left than right; this shifting is predominately responsible for the "mild flaws" we saw. Another type of mistake was the removal of box names in addition to noise characters: only 1 occasion had all names removed, but 20 instances had at least one name missing.

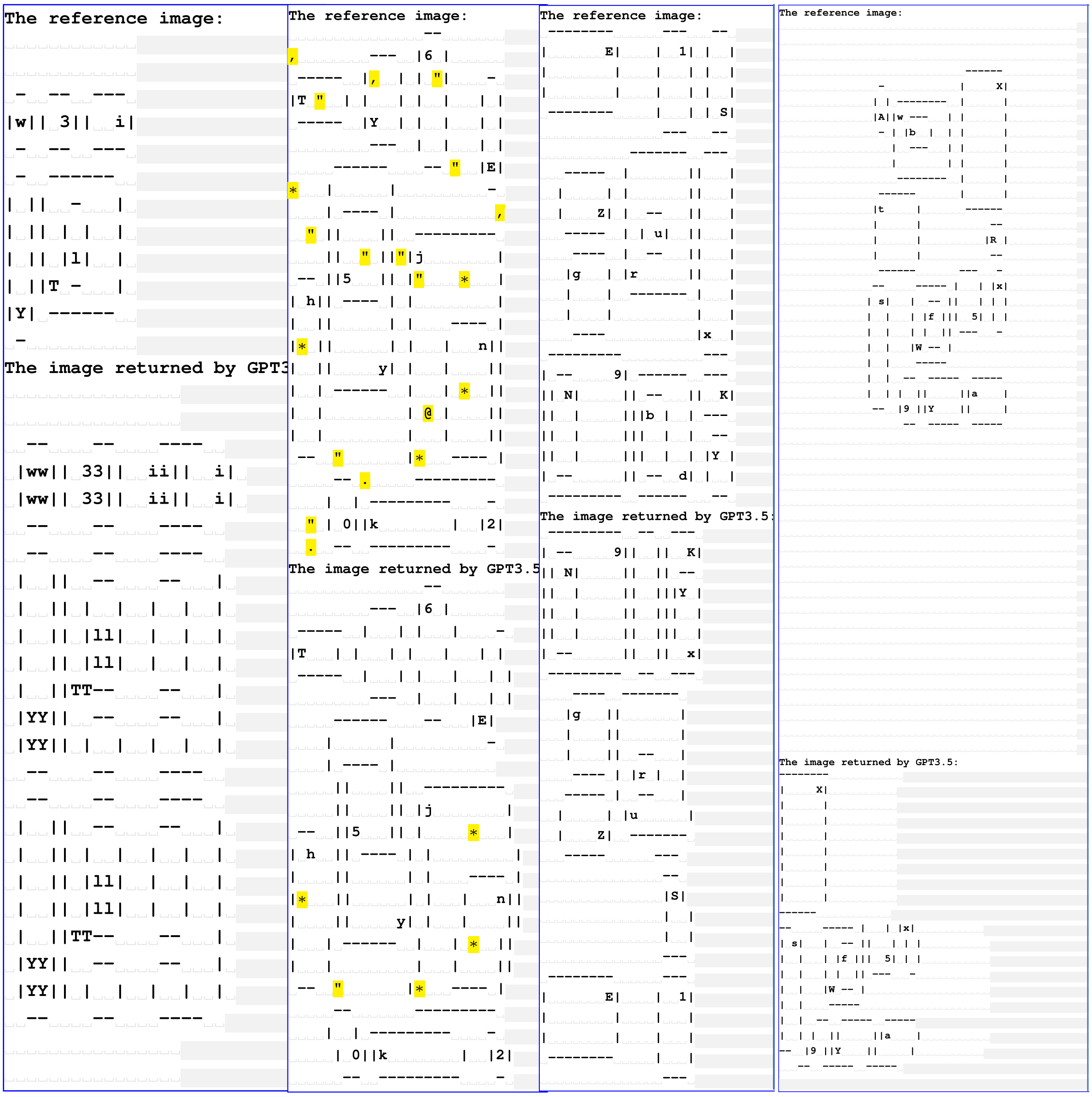
As to the removal of noise characters, we observed the following: We did not see any example where all noise characters were removed, though there was at least one case where the input was cleaned of all such marking (originally 16 characters) and retained only one. Every observed instance removed at least some of the undesired characters. No case that we saw added more noise than was originally present, and moreover the strict subset of noise remaining was located in the same position in the result as the original, save a handful of cases where the entire row was shifted one space left or right. The treatment of undesired characters did not obviously correlate with the type of noise character, location in the image, or whether it shared a row or column with other noise characters.

Taken together, this consistent decrease of noise in an image while failing to totally remove it causes us to label the outcomes as "incomplete." In light of the amount of structure retained while noise is reduced, however, a reasonable interpretation suggests GPT3.5 does not lack all prowess here.

5.5 Rescaling Trials

The 30 rescaling-trials we scrutinized were diverse and, of our experiments, most subject to the moniker "not wrong per se, but not what was initially envisioned." Indeed, it was these experiments that initially lead us to more fully appreciate the modalities of pronounced, arguably-correct behavior that would otherwise be underappreciated by more rigid, narrowly focused analysis.

We rarely saw instances where images were scaled by *exactly* double. GPT3.5 did display, however, a consistent ability to enlarge images along at


The reference image:
The image returned by GPT3.5:
The reference image:
The image returned by GPT3.5
The reference image:
The image returned by GPT3.5:
The reference image:
The image returned by GPT3.5:

(a) Scaling (b) Noise (c) Rotation (d) Translation

Figure 2: Representative, middle-grade examples of results generated by GPT3.5. The subcaptions indicate the trial from which an example is drawn. To make visible any patterns on the right-edge, we add gray blocks at the line endings. Individual spaces are distinguished with gray under-brackets. For 2b, we highlight the noise characters in yellow to ease interpretation.

least one axis or "enlarge by doubling" the picture in reasonable but unexpected ways. In the case of the first, we note that precise arithmetic is appreciated as difficult for NLP LLMs. Exactness notwithstanding, within an image, one axis generally grew while the other was kept the same size. [13] The fact scaling occurred along either axis, sometimes vertically and sometimes horizontally (and certainly at times both), is of some interest since GPT3.5 was trained mostly on languages that are read horizontally; that said, horizontal expansion appeared more frequent. Results appreciably often had a mix of boxes that were enlarged and those that retained their original size. Reductions in size were rare. This mix of behaviors across (and at times within) the instances leaves us uncomfortable commenting on the prevalence of each mode, beyond noting that each occurred appreciably often, except for the rarity of shrinking.

Consecutive repetition of names was common,

[13] In fairness to the model, we did accidentally use the singular form of "axis" in our prompt (see Figure 3d), whereas we meant the plural "axes" — the rest of the query's language hopefully conveyed what we intended despite this oversight.

either horizontally (most prominent), vertically, or, at times, in a rectangular patch within their box. 19 cases exhibited this phenomena. Name repetition tended to coincide with growth by the corresponding box. Speaking on an opposite phenomena, 15 instances lacked at least one name from the input — though not always lacking the corresponding box (which would be displayed without its label). Of these, only 7 were missing more than one name, with an observable skew towards lower counts of names missing.

As alluded to, a common modality of expansion was to repeat reference boxes, most frequently doing so in some structure-informed way (e.g. same inner-distances to copied landmarks, not thrown in haphazardly). For instance, content could be copied and translated straight down or across. Relatedly, 5 instances of the 30 featured repetition of characters until the context window end, either repeating boundaries of boxes that extended indefinitely downward or as a subset of the boxes tessellated. Of these, all but one was missing a box label; that is, they contributed 4 to the aforementioned 15 where the outputs had certain names absent.

Only a handful of times ($\approx$ 3) did the output seem largely divorced from the structure and naming of the input. Name placement in the outputs roughly matched the reference in respect to relative positioning; similar can be said of the boxes, though it appeared their size and *absolute* location varied more. All unforeseen nuances weighed, it is fair to say that certain substantive visually information was retained in the typical case, as visible in Figure 2a.

5.6 Rotation Trials

In this setting, we found two undesired modes to comprised virtually all instances, and the remaining handful not being more successful: 1. repetition of boundary marks until the end of the context window, at times preceded by a few boxes that appeared to be copied from the reference image, 2. some shuffling of content — primarily names among structure that otherwise was a copy of the reference. Case 2 had subcases which seemingly contained content flipped over an axis (ex: Figure 2c), though it is unclear what extent that holds for most instances, and may be apophenia. Another fairly common subcase, accounting for 8 instances, was that names were moved, though the boxes present (shapes and positions) matched the input. 11 instances fell into case 1, displaying a large quantity of repeated vertical or horizontal box side-markers.

In an appreciable chunk of cases (perhaps a non-simple majority) box naming underwent some changes that might constitute a partially successful flip, or two such flips along perpendicular axes; while we believe there is enough evidence to not dismiss the idea, future work is necessary to move it outside of speculation. Names did not appear to be consistently moved to destination boxes whose distance from the image boundary was qualitatively similar to the origin box's boundary distance in the reference; e.g., names from toward the center sometimes were moved to boxes touching the borders and vice versa.

Ultimately, we did not deem a single result of the 30 to be totally or largely a correct rotation. This is not surprising: neither the poor performance observed in the recognition experiments for rotations nor preliminary analysis we conducted during development provided fuel for optimism. This all said, a comfortable majority of the time we observed that substantial visual substructures were preserved, and moreover that the model made some attempt to shuffle or alter the image while preserving its rough scale and origin.

6 Conclusion

Drawing inspiration from the comprehension we'd expect an intelligent agent to possess across multiple signal modalities, in this work we examined GPT3.5's aptitude for visual tasks, where the inputs featured diagrams rendered as ASCII-art. In sharp contrast to the large majority of prior works, we made no attempt to overtly distill the image content into a lingual summary. We conducted experiments analyzing the model's performance on image matching tasks after various transforms typical in visual settings, as well as tasks requesting such transforms be generated. In each of these categories of experiment, we found that while GPT3.5 had room for notable improvement, results suggested it was not totally lacking in regard to visual and pictorial aptitude. Given that GPT3.5 is a model nominally trained on text-only input, we were pleasantly intrigued by these outcomes.

7 Limitations

We have not investigated the mechanisms by which ChatGPT achieves any visual performance. While we considered ways the LLM could "cheat" when we were constructing the experiments, that was as an attempt to diligently weed out artifacts and confounding factors. In respect to how GPT3.5 actually operates, we provide few insights into what it actually does to "compare between images", what it "pays most attention to" while "deciding", or "looks at" while "drawing boxes". These are all interesting avenues of future work, for which ideally we would conduct additional controlled experiments and, OpenAI's API then permitting, apply some of the latest methods of XAI (Explainable AI, (Gunning, 2019)) available. Considering the initial motivation of this work, resources available, and space to discuss, establishing that this is even a direction of potential interest is progress over previous perceptions.

As to our tests, more are possible and could provide additional insights. For instance, one could study whether GPT3.5 can identify subset relationships between boxes, or identify matches despite perturbing internals positions slightly (distortions, etc.); while we believe that our selection of experiments hit on the primary axes of consideration, certainly there exist additional minor axes over which experiments can be considered to ensure GPT3.5 behaves as expected.

Additional types of trials aside, those already in existence could be extended to probe further into the landscape of the network's performance. For instance, in the recognition trials, only one answer is based on the reference image, all others are freshly generated; one could consider circumstances where multiple choices are based on the reference and the network must select which corresponds to a particular transform — e.g., rotated a half turn left instead of a half turn right. As we discussed, part of our aim was to undertake experiments that were sensitive to any visual acumen GPT3.5 did possess, so the modifications would be worthwhile, but risked missing the phenomena of interest had we undertaken them *instead* of the arrangement used. Now, having established that —in contrast to general perceptions — there may be something of interest to study in this space, these additional experiments of added difficulty may provide additional insights as to the extent of GPT3.5's visual understanding.

In regard to our examination of AADs the model generated, we took strides to provide numeric descriptions as frequently as possible, while also providing what we believe is worthwhile, level-handed qualitative analysis. As we remarked in the text, we had weighed using a more cut-and-dry approach, such as training a classifier to distinguish between the generated results, the expected results, and some other, "negative" class. Such results could perhaps be an interesting complement to what we present, but would not be superior to them. Of particular concern is that much of the nuance we wished to expose may have been too easily missed by generic automated evaluation. That said, we recognize that such material could provide benefits in respect to exactness, digestibility (for readers), and quantifiable summarization.

In potential contrast to automated means, it may be possible that more nuance could be had with human trials, particularly by leveraging a comprehensive series of survey questions (in contrast to just manually performed image matching tasks, say). In particular, gathering detailed impressions from neutral arbiters as to the qualitative properties of outcomes would help further gauge GPT3.5 successfulness (rare as such surveys may be for accessing image generation systems). Outside of that, arrangements similar to blind A/B-testing could be performed where, given a pair of AADs prepared by some variety of means — separate random draws, input and results from GPT3.5, and perhaps other near-alternatives — must select how they relate (rotation, scaling, unrelated, etc.); this however runs into the issue of missing subtly, hence the suggestion for more detailed and expansive surveying.

Finally, while the data we generated to perform analysis has many merits, certainly there are limitations. Most obviously, the data is ultimately patterned after the shared, fundamental structure of AADs. In the same spirit of exploring the space over which GPT3.5 is visually performant, more varied datasets could be used, which would also boost confidence that outcomes are not special to our setting. Risk of the model secretly exploiting artifacts and biases specific to the generative process should be borne in mind, particularly since we mechanically generate our data.[14] This all said, however, works like (Ribeiro et al., 2016; Khalid et al., 2021) and (Eykholt et al., 2018) show that even "more respectable" CV systems and datasets are subject to similar categories of concern, if not comparable degrees of it.

[14] It is possible that the degree of hesitation one has should also correlate with the size of the AADs used, however we are not yet promoting that stance.

Acknowledgements

We would like to thank the anonymous reviewers for their efforts and feedback.

References

Amit Arora. 2022. Stop Asking ChatGPT to Create ASCII. *Medium Corporation*. Last accessed 17 June 2023.

Sara Di Bartolomeo, Giorgio Severi, Victor Schetinger, and Cody Dunne. 2023. Ask and You Shall Receive (a Graph Drawing): Testing ChatGPT's Potential to Apply Graph Layout Algorithms. *CoRR*, arXiv:2303.08819.

Building Blocks. 2022. 7 Interesting Experiments with ChatGPT. Last accessed 17 June 2023.

Pietro Bongini, Federico Becattini, and Alberto Del Bimbo. 2022. Is GPT-3 All You Need for Visual Question Answering in Cultural Heritage? In *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, volume 13801 of *Lecture Notes in Computer Science*, pages 268–281. Springer.

Rodney A. Brooks. 1991. Intelligence without Representation. *Artif. Intell.*, 47(1-3):139–159.

Heather Brown. 2023. What exactly is ChatGPT? *CBS News Minnesota*. `https://web.archive.org/web/20230209055245/https://www.cbsnews.com/minnesota/news/what-exactly-is-chatgpt/`. Last accessed 14 June 2023.

Georgia Chalvatzaki, Ali Younes, Daljeet Nandha, An Le, Leonardo F. R. Ribeiro, and Iryna Gurevych. 2023. Learning to Reason over Scene Graphs: A Case Study of Finetuning GPT-2 Into a Robot Language Model for Grounded Task Planning. *CoRR*, arXiv:2305.07716.

Liting Chen, Lu Wang, Hang Dong, Yali Du, Jie Yan, Fangkai Yang, Shuang Li, Pu Zhao, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. Introspective Tips: Large Language Model for In-Context Decision Making. *CoRR*, arXiv:2305.11598.

Jack Cushman. 2022. ChatGPT: Poems and Secrets. *The Library Innovation Lab at the Reginald F. Lewis Law Center, Harvard University*. Last accessed 12 May 2023.

Maksymilian Dabkowski and Gasper Begus. 2023. Large Language Models and (Non-)Linguistic Recursion. *CoRR*, abs/2306.07195.

Sanjay Deshpande and Jakub Szefer. 2023. Analyzing ChatGPT's Aptitude in an Introductory Computer Engineering Course. *CoRR*, arXiv:2304.06122.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634. Computer Vision Foundation / IEEE Computer Society.

David Gunning. 2019. Darpa's explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*. ACM.

Jiayan Guo, Lun Du, and Hengyu Liu. 2023. GPT4Graph: Can Large Language Models Understand Graph Structured Data ? An Empirical Evaluation and Benchmarking. *CoRR*, arXiv:2305.15066.

Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. 2023. Instruct2Act: Mapping Multi-modality Instructions to Robotic Actions With Large Language Model. *CoRR*, arXiv:2305.11176.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. 2023. A Glimpse in ChatGPT Capabilities and its impact for AI research. *CoRR*, arXiv:2305.06087.

Brian W. Kernighan. 1982. PIC-A Language for Typesetting Graphics. *Softw. Pract. Exp.*, 12(1):1–21.

Faiq Khalid, Muhammad Abdullah Hanif, and Muhammad Shafique. 2021. Exploiting Vulnerabilities in Deep Neural Networks: Adversarial and Fault-Injection Attacks. *CoRR*, arXiv:2105.03251.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. 2023. LMEye: An Interactive Perception Network for Large Language Models. *CoRR*, arXiv:2305.03701.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *CoRR*, arXiv:2304.01852.

Paula Maddigan and Teo Susnjak. 2023. Chat2VIS: Generating Data Visualisations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models. *CoRR*, arXiv:2302.02094.

Bernard Marr. 2023. 10 Amazing Real-World Examples Of How Companies Are Using ChatGPT In 2023. *Forbes*. Last accessed 14 June 2023.

Hans Moravec. 1993. The universal robot. *NASA. Lewis Research Center, Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace*.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. EmbodiedGPT: Vision-Language Pre-Training via Embodied Chain of Thought. *CoRR*, arXiv:2305.15021.

Laksh Nanwani, Anmol Agarwal, Kanishk Jain, Raghav Prabhakar, Aaron Monis, Aditya Mathur, Krishna Murthy, Abdul Hafez, Vineet Gandhi, and K. Madhava Krishna. 2023. Instance-Level Semantic Maps for Vision Language Navigation. *CoRR*, arXiv:2305.12363.

Kate O'Riordan. 2014. *ASCII art*. Encyclopædia Britannica, Inc. `https://www.britannica.com/topic/ASCII-art`. Last accessed 16 June 2023.

Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. 2023. Instructvid2vid: Controllable video editing with natural language instructions. *CoRR*, arXiv:2305.12328.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Ahmed R. Sadik, Antonello Ceravola, Frank Joublin, and Jibesh Patra. 2023. Analysis of ChatGPT on Source Code. *CoRR*, arXiv:2306.00597.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks With ChatGPT and its Friends in Hugging Face. *CoRR*, arXiv:2303.17580.

Yucheng Shi, Hehuan Ma, Wenliang Zhong, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. 2023. ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs. *CoRR*, arXiv:2305.03513.

Qingyi Si, Yuchen Mo, Zheng Lin, Huishan Ji, and Weiping Wang. 2023. Combo of Thinking and Observing for Outside-Knowledge VQA. *CoRR*, arXiv:2305.06407.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2023a. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Megha Srivastava, Noah Goodman, and Dorsa Sadigh. 2023b. Generating Language Corrections for Teaching Physical Control Tasks. *CoRR*, arXiv:2306.07012.

Richard Sutton. 2019. The Bitter Lesson.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C. H. Hoi. 2022. Plug-and-play VQA: zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 951–967. Association for Computational Linguistics.

Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. 2023. Level Generation Through Large Language Models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*. ACM.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023a. Can Language Models Solve Graph Problems in Natural Language? *CoRR*, arXiv:2305.10037.

Hong Wang, Xuan Luo, Weizhi Wang, and Xifeng Yan. 2023b. Bot or Human? Detecting ChatGPT Imposters with A Single Question. *CoRR*, arXiv:2305.06424.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, volume 35, pages 24824–24837.

Wetrorave. 2022. ChatGPT Can Draw, but it Started Drawing Other Things. `https://web.archive.org/web/20230617055325/https://www.reddit.com/r/artificial/comments/zc0og6/chatgpt_can_draw_but_it_started_drawing_other/`. Last accessed 17 June 2023.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing With Visual Foundation Models. *CoRR*, arXiv:2303.04671.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3081–3089. AAAI Press.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action. *CoRR*, arXiv:2303.11381.

Yang Ye, Hengxu You, and Jing Du. 2023. Improved Trust in Human-Robot Collaboration with ChatGPT. *CoRR*, arXiv:2304.12529.

Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, Gyeong-Moon Park, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon, and Choong Seon Hong. 2023. One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. *CoRR*, arXiv:2304.06488.

Jiawei Zhang. 2023. Graph-ToolFormer: To Empower LLMs With Graph Reasoning Ability via Prompt Augmented by ChatGPT. *CoRR*, arXiv:2304.11116.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *CoRR*, arXiv:2304.06364.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. *CoRR*, arXiv:2305.13168.

A Additional Comments In Regards to the Gap in Ability Between Utilizing Verbal Summaries of Images and Being Able to Directly Process Images

Handling symbolic structures that happen to be derived from spatial data may be more akin to an "algebraic computation" than "visual understanding" (using those phrases connotatively if not a firm distinction). For instance, from group theory alone, one knows that applying a transformation T followed by -T results in the identity. It may well be that T is translating a triangle 10 meters left and -T moves the same distance right; an LLM could conclude that T then -T results in no change completely divorced from whatever T is meant to represent. In that process, though, the model wouldn't necessarily know how vertices of the triangle move over the course of the transformation — and moreover, it doesn't mean that the model could derive the vertices from a bitmap image, or even be able to recognize a triangle in the picture. T and -T could just as well be depositing then withdrawing money from a bank account. The LLM may be able to handle the high-level summary of what an image contains, but by the time such a description is produced, much of what makes it a visual problem is already treated. As a historical footnote, popular perception about the difficulties symbolic AI had for processing raw visual input (e.g., Moravec's Paradox) bolster the position that this gap is not to be taken for granted; see, for instance, (Brooks, 1991; Moravec, 1993; Sutton, 2019) for a couple critical takes.

B More Details About Information We Provide in Prompts for Recognition Experiments

As noted, in general we keep the details of what we inform the model of in the AADs to a minimum. In the following circumstances, we provide a few more words which may reveal additional — albeit minimal — aspects of the AAD: 1. *When names are used:* We indicate names are alphanumeric and occur on the inside boundary of objects. 2. *Noise trials:* We explicitly refer to "boxes" being present. We do not indicate what characters comprise them or the noise. 3. *Size trials:* We specify whether the choices are scaled up or scaled down in respect to the reference.

C Regarding String Parsing to Extract Content

For recognition experiments: Basic string parsing (e.g. using regular expressions) was able to consistently extract the primary response (i.e., which option corresponds to the reference); our code flagged instances of unexpected content and separated them for manual review, but ultimately that only triggered seven times out of several thousand cases; in light of their minimal impact, we ultimately disregarded them, finding that the benefit of their use was outweighed by the added methodological cleanliness.

For generation experiments: In preliminary trials, we found that a fraction of the time GPT3.5

would reply to our prompt solely with text[15] or other non-AArt content. In order to ease planned downstream analysis, we opted to add a lightweight mechanism for detecting such cases and reissuing the query. The heuristic deployed checked that the response was at least of minimal feasible length to contain an image and at least one arrangement of characters that looked like a potential box corner (i.e., “-” on one line, “|” adjacent on a line above or below).

The illustrations we share were extracted with a two-step, heuristic process: (1) return the content in the last pair of triple back-ticks (“` ` `”) present in the output, (2) if the first option does not extract content seemingly containing a box[16] return everything after the last line holding at least two consecutive alphanumeric characters. The second step, when invoked, aims to cut out anything that may loosely look like text/words. We consider the lack of human effort in the extraction process to be both convenient and reassuring, the latter as it mitigates concern over human biases impacting output characteristics like tabbing or presence of excess whitespace margins.

D Prompts Used for AArt-Generation Trials

[15] A paraphrased example of those occasional replies: “I'm sorry, but as an LLM, I can't process ASCII-art.”

[16] Using the method of the prior paragraph.

```
Instructions: I am about to show you a reference ASCII-art image, and then ask you questions about it and a task you must
      complete. The questions are numbered 1, 2, and
      3, and the task is indicated separately. The ASCII-art depicts a collection of boxes, some of which may be nested inside
      of other boxes. Note that in the ASCII-art, each box depicted is labeled with a unique name, which consists of an
      alphanumeric character and which appears in one of the box's corners.

Reference ASCII-art Image:
```
[...]
```
```

(a) Preamble text with overview of the tasks GPT3.5 is requested to complete, followed by the placement of where ASCII-art would be, as indicated by the bolded, bracketed ellipsis (**[...]**). The bolded, italicized text in the preamble is substituted with "3 and 4," whenever the experiment involves four such questions.

```
[...Preamble from Figure 3a...]

     Your job is to do the following, in order:
(1) Describe the reference ASCII-art image.
(2) What would you do in order to form a piece of ASCII-art that matches what the reference ASCII-art would look like if it
      had no blank areas at the top of it and no empty left margin? That is, how would you change the reference ASCII-art to
      look like it was translated so that there was not unneeded empty space around it (while preserving all internal spacing
      and structured)?
(3) What would the reference ASCII-art look like if it had no blank areas at the top of it and no empty left margin? That is,
      what would the reference ASCII-art look like after it has been translated so that there was not unneeded empty space
      around it?

Task: Provide ASCII-art that matches what the reference ASCII-art would look like if it was translated to have no blank areas
      at the top of it and no empty left margin. That is, show a modified version of the reference ASCII-art that has been
      translated so that there is no unneeded empty space around it (while preserving internal spacing and structure).
```

(b) Prompt used for trials of generating image translations.

```
[...Preamble from Figure 3a...]

     Your job is to do the following, in order:
(1) Describe the reference ASCII-art image.
(2) In the reference ASCII-art, the only characters that should be present are ``|'', ``-'', alphanumeric characters, or
      whitespace. All other characters are noise that should not be present. List what characters are present in the reference
       ASCII-art that are noise.
(3) How would you remove noise from the reference ASCII-art so that only the characters that should be there are present?
(4) What would the ASCII-art look like if each character that is noise was replaced with a single space character?

Task: Provide what the reference ASCII-art would look like if you remove the noise and only leave the characters that should
      be present. Any single character you remove should be replace by a single space character.
```

(c) Prompt used for trials of generating de-noised versions of reference images.

```
[...Preamble from Figure 3a...]

     Your job is to do the following, in order:
(1) Describe the reference ASCII-art image.
(2) What would you do in order to form a piece of ASCII-art that matches what the reference ASCII-art would look like if it
      was scaled up to double the size?
(3) What would the reference ASCII-art look like if it was enlarge by a factor of two? That is, what would the reference ASCII
      -art look like if it was made twice as large?

Task you must complete after answering the questions: Provide ASCII-art that matches what the reference ASCII-art would look
      like if we scaled the reference ASCII-art to double its size. That is, produce ASCII-art that has axis which are double
      the length of the reference, and which the images shown are enlarged respectively.
```

(d) Prompt used for trials of generating enlarged copies of images.

```
[...Preamble from Figure 3a...]

     Your job is to do the following, in order:
(1) Describe the reference ASCII-art image.
(2) What would you do in order to form a piece of ASCII-art that matches what the reference ASCII-art would look like if it
      was rotated 90 degrees clockwise? That is, what you you do in order to depict the reference image after a quarter-turn
      clockwise?
(3) What would the reference ASCII-art look like if it was rotated 90 degrees clockwise? That is, what would the reference
      image look like after a quarter-turn clockwise?

Task: Provide ASCII-art that matches what the reference ASCII-art would look like if it was rotated 90 degrees clockwise. That
       is, show the reference ASCII-art after it has been rotated a quarter-turn clockwise.
```

(e) Prompts used for trials of generating image rotations.

Figure 3: Prompts Used for AArt-Generation Trials. By the nature of the generation task compared to recognition, some trials required more information be specified in the prompt to more narrowly specify the set of assemble outcomes. Compare, for instance, to the overview provided in Section 4.1 and Appendix B.