Grounded Intuition of GPT-Vision's Abilities with Scientific Images

Alyssa Hwang Andrew Head Chris Callison-Burch

Department of Computer and Information Science
University of Pennsylvania

GPT-Vision has impressed us on a range of vision-language tasks, but it comes with the familiar new challenge: we have little idea of its capabilities and limitations. In our study, we formalize a process that many have instinctively been trying already to develop "grounded intuition" of this new model. Inspired by the recent movement away from benchmarking in favor of example-driven qualitative evaluation, we draw upon grounded theory and thematic analysis in social science and human-computer interaction to establish a rigorous framework for qualitative evaluation in natural language processing. We use our technique to examine alt text generation for scientific figures, finding that GPT-Vision is particularly sensitive to prompting, counterfactual text in images, and relative spatial relationships. Our method and analysis aim to help researchers ramp up their own grounded intuitions of new models while exposing how GPT-Vision can be applied to make information more accessible.

GPT-Vision 在一系列视觉语言任务上给我们留下了深刻的印象，但它也带来了一个熟悉的新挑战：我们对其能力和局限性知之甚少。在我们的研究中，我们将许多人本能地尝试开发这种新模型的"扎实直觉"的过程形式化。受最近从基准测试转向示例驱动的定性评估的趋势的启发，我们借鉴了社会科学和人机交互中的 扎根理论和 主题分析，为自然语言处理中的定性评估建立了一个严格的框架。我们使用我们的技术来检查科学图表的替代文本生成，发现 GPT-Vision 对提示、图像中的反事实文本和相对空间关系特别敏感。我们的方法和分析旨在帮助研究人员增强他们对新模型的扎实直觉，同时揭示如何应用 GPT-Vision 使信息更易于访问。

Index

1 Introduction

1.1 Motivation

The recent release of GPT-Vision (OpenAI, 2023a) has prompted widespread excitement—promising to usher in a new era of multi-modal generative AI applications (Yang et al., 2023). However, a pre-requisite for large-scale utilization of new AI technology is a comprehensive understanding of its associated limitations and failure cases. Without such an understanding, we risk deploying our models in ways that cause real harm to real people—especially in high-stakes domains.
GPT-Vision (OpenAI, 2023a) 的近期发布引起了广泛关注，有望开启多模态生成式 AI 应用的新时代 (Yang et al., 2023)。然而，大规模使用新 AI 技术的先决条件是全面了解其相关限制和失败案例。如果没有这样的理解，我们可能会以对真人造成真正伤害的方式部署我们的模型，尤其是在高风险领域。

In this paper we conduct a qualitative example-driven analysis of the various capabilities and limitations of the newly-released GPT-Vision model. Following the lead of Bubeck et al. (2023), rather than conducting our analysis in the more traditional way (e.g. collecting a large dataset and computing automatic metrics), we take a more example-driven approach—focusing intently on a small number of illustrative data points and analyzing them extremely closely to glean broader insights and trends. Such an analysis, contrary to the language used in Bubeck et al. (2023), has substantial precedent in the social science and human-computer interaction literature and is widely accepted to be scientifically rigorous.
在本文中，我们对新发布的 GPT-Vision 模型的各种功能和局限性进行了定性示例驱动分析。遵循 Bubeck et al. (2023) 的指导，我们不是采用更传统的方式（例如收集大量数据集并计算自动指标）进行分析，而是采取一种更以示例为导向的方法——专注于少数说明性数据点并对其进行极其仔细的分析，以收集更广泛的见解和趋势。与 Bubeck et al. (2023) 中使用的语言相反，这种分析在社会科学和人机交互文献中具有大量先例，并被广泛认为是科学严谨的。

Furthermore, drawing inspiration from grounded theory (Blandford et al., 2022b) and thematic analysis (Blandford et al., 2022a), we develop and standardize a rigorous method for conducting qualitative analyses of generative AI models. Our method consists of five stages: (1) data collection, (2) data review, (3) theme exploration, (4) theme development, and (5) theme application. As we demonstrate from our findings, such analysis when performed properly allows for deep and intuitive understanding of model capabilities, even when done on relatively small sample sizes.
此外，从扎根理论 (Blandford et al., 2022b) 和主题分析 (Blandford et al., 2022a) 中汲取灵感，我们开发并标准化了一种严格的方法，用于对生成式 AI 模型进行定性分析。我们的方法包括五个阶段：(1) 数据收集、(2) 数据审查、(3) 主题探索、(4) 主题开发和 (5) 主题应用。正如我们从研究结果中证明的那样，如果正确执行此类分析，即使在相对较小的样本量上进行分析，也可以深入了解模型功能。

To illustrate these claims, we focus on one particular task domain: alt text generation for pages and figures in scientific papers. This is a particularly fertile area for analysis, as properly describing the contents of a particular page or figure often requires complex reasoning capabilities that go far beyond simple object detection. Through our analysis we find that GPT-Vision, while extremely impressive, has a tendency to over-rely on textual information, is particularly sensitive to the wording of its prompts, and struggles with reasoning about spatial locality. We are also able to confirm the existence of many of the pitfalls and shortcomings quoted by OpenAI in their model card (OpenAI, 2023a). Overall, we not only provide insights into the limitations of the newly-released GPT-Vision model but also provide an example of the judicious application of qualitative analysis techniques to generative AI models.
为了说明这些说法，我们专注于一个特定的任务领域：科学论文中页面和图表的替代文本生成。这是一个特别有分析价值的领域，因为正确描述特定页面或图表的内容通常需要复杂的推理能力，远远超出简单的对象检测。通过我们的分析，我们发现 GPT-Vision 虽然非常令人印象深刻，但它倾向于过度依赖文本信息，对其提示的措辞特别敏感，并且在推理空间局部性方面存在困难。我们还能够确认 OpenAI 在其模型卡 (OpenAI, 2023a) 中引用的许多陷阱和缺点的存在。总的来说，我们不仅深入了解了新发布的 GPT-Vision 模型的局限性，而且还提供了一个将定性分析技术明智地应用于生成式 AI 模型的示例。

1.2 Background & Related Work

Trying to evaluate the performance of a given model has always been a challenging task. However, recently the rising capabilities of our best models have begun to reveal longstanding shortcomings in our existing evaluations.
评估给定模型的性能一直是一项艰巨的任务。然而，最近我们的最佳模型能力的不断提升已经开始暴露出我们现有评估中长期存在的缺陷。

Now that large language models are capable of producing such sophisticated output for a wide range of requests, "evaluating generated text is now about as hard as generating it" (Neubig, 2023). Recent work has warned us against relying on long-used automatic metrics for tasks like machine translation (Fomicheva and Specia, 2019), question answering (Chen et al., 2019), and summarization (Jain et al., 2023; Goyal et al., 2023) because they may fail to accurately assess novel and abstractive text against a gold standard. Even reference-free metrics have been shown to underestimate the quality of generated text, perhaps because those metrics were trained or evaluated on the same reference-based benchmarks (Goyal et al., 2023). Automatic metrics have long been criticized for unreliably correlating with human judgment, even before the rise of LLMs (Deutsch et al., 2021; Belz and Reiter, 2006). Reference-free metrics are disproportionately weak at evaluating alt text for blind and low-vision readers (Kreiss et al., 2022). Some work has attempted to mitigate these challenges by using an intermediary LLM to evaluate generated text (Liu et al., 2023; Ding et al., 2023), designing AI tools to aid data annotation (Gao et al., 2023), or improve metrics and datasets for new LLMs (Jain et al., 2023; Zhong et al., 2023; Sawada et al., 2023).
现在，大型语言模型能够为各种请求生成如此复杂的输出，"评估生成的文本现在与生成文本一样困难" (Neubig, 2023)。最近的研究警告我们不要依赖长期使用的自动指标来完成机器翻译 (Fomicheva and Specia, 2019)、问答 (Chen et al., 2019) 和摘要 (?Goyal et al., 2023) 等任务，因为它们可能无法根据黄金标准准确评估新颖和抽象的文本。甚至无参考指标也被证明低估了生成文本的质量，这可能是因为这些指标是在相同的基于参考的基准上训练或评估的 (Goyal et al., 2023)。早在 LLM 兴起之前，自动指标就一直因与人类判断不可靠相关而受到批评 (Deutsch et al., 2021; Belz and Reiter, 2006)。无参考指标在评估盲人和低视力读者的替代文本方面特别薄弱 (Kreiss et al., 2022)。一些研究尝试通过使用中间 LLM 来评估生成的文本 (Liu et al., 2023; Ding et al., 2023)、设计 AI 工具来帮助数据注释 (Gao et al., 2023)，或改进新 LLM 的指标和数据集 (?Zhong et al., 2023; Sawada et al., 2023) 来缓解这些挑战。

Recent example-driven qualitative analyses of GPT-4 and GPT-Vision have already stepped toward robust qualitative analysis for modern LLMs (Bubeck et al., 2023; OpenAI, 2023a; Yang et al., 2023). While these studies tend to provide brief commentary on a large number of tasks and examples, we examine a small set of results more deeply through a method based on grounded theory (Blandford et al., 2022b) and thematic analysis (Blandford et al., 2022a), which are frequently used in human-computer interaction research. Grounded theory is a data-driven or "bottom-up" perspective on data collection and analysis. In grounded theory, patterns and conclusions "emerge" from the data, much like an inductive analysis (Bingham, 2023). Grounded theory instructs analysts to make meaning solely from the data to avoid bias from preconceived notions or existing theories. It includes a method called theoretical sampling, which is based on the idea that we can carefully select data that contains characteristics we care about as opposed to sampling at random or gathering a large dataset Blandford et al. (2022c). Theoretical sampling also allows data to be gathered iteratively to address findings as they arise throughout the analysis until we hit "theoretical saturation": a subjective yet evidence-based instinct that further data collection and analysis will not reveal any more major insights.
最近对 GPT-4 和 GPT-Vision 进行的示例驱动定性分析已经朝着现代 LLM 的稳健定性分析迈进了一步 (Bubeck et al., 2023; OpenAI, 2023a; Yang et al., 2023)。虽然这些研究倾向于对大量任务和示例提供简短的评论，但我们通过基于扎根理论 (Blandford et al., 2022b) 和主题分析 (Blandford et al., 2022a) 的方法更深入地研究了一小组结果，这些方法经常用于人机交互研究。扎根理论是一种数据驱动或 "自下而上" 的数据收集和分析视角。在扎根理论中，模式和结论从数据中 "浮现"，就像归纳分析一样 (Bingham, 2023)。扎根理论指导分析师仅从数据中获取意义，以避免先入为主的观念或现有理论的偏见。它包括一种称为理

论抽样的方法，该方法基于这样一种理念：我们可以仔细选择包含我们关心的特征的数据，而不是随机抽样或收集大量数据集 Blandford et al. (2022c)。理论抽样还允许迭代收集数据以解决分析过程中出现的问题，直到我们达到“理论饱和”：一种主观但基于证据的直觉，即进一步的数据收集和分析不会揭示任何重大见解。

Thematic analysis is a flexible framework through which grounded theory can be applied. First, “themes” are gathered from the data, refined, and then applied to the entire dataset to reveal patterns within it. When adopted formally, thematic analysis is a rigorous process that can involve evaluating inter-annotator agreement and setting up infrastructure to protect reliability in qualitative research (McDonald et al., 2019). It has been used regularly in well reputed HCI studies like supporting healthcare (Bowman et al., 2023), analyzing social media posts (Gauthier et al., 2022), and conducting literature reviews (Cooper et al., 2022). Brand-new work to be published in the Findings of EMNLP 2023 even proposes an LLM-in-the-loop collaboration framework to assist with thematic analysis (Dai et al., 2023). Our work adapts thematic analysis and grounded theory specifically for evaluating LLMs in NLP research.
主题分析是一个灵活的框架，通过它可以应用扎根理论。首先，从数据中收集“主题”，对其进行提炼，然后应用于整个数据集以揭示其中的模式。正式采用主题分析是一个严格的过程，可能涉及评估注释者之间的一致性并建立基础设施以保护定性研究中的可靠性 (McDonald et al., 2019)。它已在著名的 HCI 研究中定期使用，例如支持医疗保健 (Bowman et al., 2023)、分析社交媒体帖子 (Gauthier et al., 2022) 和进行文献综述 (Cooper et al., 2022)。即将在 EMNLP 2023 的成果中发表的全新研究甚至提出了一个 LLM-in-the-loop 协作框架来协助主题分析 (Dai et al., 2023)。我们的工作专门采用主题分析和扎根理论来评估 NLP 研究中 LLM。

Our analysis focuses on GPT-Vision’s ability to describe scientific images. Past work on describing images has included automatic image captioning (Tang et al., 2023; Hsu et al., 2021; Spreafico and Carenini, 2020; Guinness et al., 2018) and alt text generation (Wu et al., 2017; Salisbury et al., 2017; Williams et al., 2022; Chintalapati et al., 2022). Alt text is a written version of an image that appears in place of it (VLE Guru, 2022). Although alt text is typically associated with screen readers and vision loss, it can also help users with information processing disorders, like issues with visual sequencing, long- or short-term visual memory, visual-spatial understanding, letter or symbol reversal, or color blindness (McCall and Chagnon, 2022). Alt text can even help in purely situational circumstances like broken image links or loading issues due to expensive data roaming or weak internet connectivity, which may disproportionately affect individuals with lower incomes (VLE Guru, 2022). Beyond reading online documents, alt text and curated image descriptions can allow audio books and screen readers to “read aloud” visual content, giving all of us even more access to news articles, textbooks, blog posts, scientific papers, and other mixed-media texts. These image descriptions, however, need to be generated carefully and, most likely, adaptively. Alt text by definition depends on the audience, content, and situation, so one approach will not work for all images or people (Eggert et al., 2022). This claim was validated in practice by a user study (Stangl et al., 2021). Blind and sighted readers diverge (Lundgard and Satyanarayan, 2021). Even placement of text has an impact (Stokes et al., 2022).
我们的分析重点是 GPT-Vision 描述科学图像的能力。过去描述图像的工作包括自动图像字幕 (?) 和替代文本生成 (?)。替代文本是代替图像出现的书面版本 (VLE Guru, 2022)。虽然替代文本通常与屏幕阅读器和视力丧失有关，但它也可以帮助信息处理障碍的用户，例如视觉排序问题、长期或短期视觉记忆、视觉空间理解、字母或符号反转或色盲 (McCall and Chagnon, 2022)。替代文本甚至可以在纯粹的情境情况下提供帮助，例如由于昂贵的数据漫游或互联网连接不佳而导致的图像链接断开或加载问题，这可能会对低收入人群产生不成比例的影响 (VLE Guru, 2022)。除了阅读在线文档之外，替代文本和精选的图像描述还可以让有声读物和屏幕阅读器“大声朗读”视觉内容，让我们所有人能够更多地访问新闻文章、教科书、博客文章、科学论文和其他混合媒体文本。然而，这些图像描述需要谨慎生成，而且很可能是自适应的。从定义上来说，替代文本取决于受众、内容和情况，因此一种方法并不适用于所有图像或人 (Eggert et al., 2022)。这一说法已在实践中得到用户研究的验证 (Stangl et al., 2021)。盲人和视力正常的读者存在差异 (Lundgard and Satyanarayan, 2021)。即使文本的位置也会有影响 (Stokes et al., 2022)。

Part of this analysis is “objective,” such as detecting objects, transcribing labels, and identifying spatial positions, but many aspects are inherently human-centered. What is the “correct” interpretation of a graph

or the "main idea" of a diagram? What is an "appropriate" description—not too long, not too short, not too detailed, not too vague? Now that LLMs are so powerful and widely used, we need to address what we as users want from the model beyond just the facts, which we can start to investigate through the human-centered design framework (Norman, 2013). We should also acknowledge that different users have different needs, which are often affected by their differing abilities. The same ability can vary in duration and context—consider a user who can needs to have a book read aloud because they are blind, are in a dark room, or had their pupils dilated—as suggested by the ability-based design framework (Wobbrock et al., 2011). Together, ability-based human-centered design can help us build inclusive tools for everyone (Hwang, 2023).
这种分析的一部分是"客观的"，例如检测物体、转录标签和识别空间位置，但许多方面本质上是以人为中心的。什么是图形的"正确"解释或图表的"主要思想"？什么是"适当"的描述——不太长，不太短，不太详细，不太模糊？既然 LLM 如此强大且被广泛使用，我们需要解决我们作为用户希望从模型中得到什么，而不仅仅是事实，我们可以通过以人为中心的设计框架开始调查这些。我们还应该承认，不同的用户有不同的需求，这往往受到他们不同能力的影响。同一种能力在持续时间和情境上可能会有所不同——设想一个用户因为失明、在黑暗的房间或瞳孔扩大而需要大声朗读一本书——正如基于能力的设计框架所建议的那样 (Wobbrock et al., 2011)。综合起来，基于能力的以人为本的设计可以帮助我们为每个人构建包容性的工具 (Hwang, 2023)。

1.3 Contributions

In this paper, we contribute:

- Deep, grounded insights on GPT-Vision describing scientific images
  对 GPT-Vision 描述科学图像的深入、扎实见解
- A qualitative analysis framework based on grounded theory and thematic analysis for evaluating LLMs
  基于扎根理论和主题分析的定性分析框架，用于评估 LLMs
- The images we used and the text we generated for future work and reproducibility
  我们使用的图像和我们为未来工作和可重复性生成的文本

Part of our goal was to formalize a process that people have already naturally taken to evaluate LLMs: trying a selection of images and prompts, skimming through generated text, and noticing patterns until we are satisfied with our "intuition." Our method provides an organized, systematic framework for intentionally developing this intuition grounded in concrete data. A practical guide on our method and theoretical background will be released soon.
我们的目标之一是将人们自然而然采用的评估 LLM 的过程正式化：尝试选择图像和提示，浏览生成的文本，并注意模式，直到我们对自己的"直觉"感到满意。我们的方法提供了一个有组织的系统框架，用于有意地基于具体数据发展这种直觉。我们将很快发布有关我们的方法和理论背景的实用指南。

2 Methods and Data

Analysis Procedure We based our approach to qualitative analysis on well established practices in grounded theory and thematic analysis (see Section 1.2). It consisted of five phases: (1) data collection, (2) data review, (3) theme exploration, (4) theme development, and (5) theme application. During the data collection phase, we prompted GPT-Vision to describe a set of scientific figures. We then lightly reviewed the data for notable patterns before carefully searching for "themes" during the theme exploration phase. Theme development was dedicated to consulting literature and refining the themes that had emerged in the exploration. Finally, we passed through the data one last time to apply the finalized themes to the entire dataset. This method allowed us to conduct a more rigorous qualitative analysis to gain evidence-grounded intuition about a brand-new

model, as we discuss in Section 3.
我们的定性分析方法基于扎根理论和主题分析中成熟的实践（参见 1.2 节）。它包括五个阶段：（1）数据收集、（2）数据审查、（3）主题探索、（4）主题开发和（5）主题应用。在数据收集阶段，我们提示 GPT-Vision 描述一组科学数字。然后，我们在主题探索阶段仔细搜索“主题”之前，对数据进行了简单的审查，以寻找值得注意的模式。主题开发致力于查阅文献并细化探索中出现的主题。最后，我们最后一次检查数据，将最终确定的主题应用于整个数据集。这种方法使我们能够进行更严格的定性分析，以获得关于全新模型的基于证据的直觉，正如我们在 3 节中讨论的那样。

Data collection During the first phase of our analysis, we collected data through a theoretical sampling approach (Blandford et al., 2022c). We were initially interested in GPT-Vision's ability to describe scientific figures and eventually expanded to images of code, math, and even full pages from research publications. Our final set of images contained two photos, three diagrams, four graphs, three tables, five screenshots of full pages, three images with computer code, and two images with mathematical notation for a total of 21 images (see Appendix Tables 3 and 4). For figures, we included the texts of the caption and a reference paragraph as context as well. We queried GPT-Vision with the following two prompts for each image, giving us a total of 42 generated passages:
在分析的第一阶段，我们通过理论抽样方法 (Blandford et al., 2022c) 收集数据。我们最初对 GPT-Vision 描述科学图形的能力感兴趣，最终扩展到代码、数学甚至研究出版物的整页图像。我们最终的图像集包含两张照片、三张图表、四张图、三张表格、五张整页截图、三张带有计算机代码的图像和两张带有数学符号的图像，总共 21 幅图像 (参见附录表 3 和 4)。对于图形，我们还包含了标题文本和参考段落作为上下文。我们对每幅图像使用以下两个提示查询 GPT-Vision，总共生成了 42 个段落：

> “alt”: Write alt text for this <input> .
>
> “desc”: Describe this <input> as though you are speaking with someone who cannot see it.

We replaced <input> with “figure” for photos, diagrams, and graphs; “table” for tables; “page” for screenshots of full pages; and “image” for images of special text (code or math).[1]
我们将照片、图表和图形的 <input> 替换为“figure”；将表格的“table”；将整页截图的“page”；将特殊文本（代码或数学）的图像的“image”。[2]

Data review After settling on a preliminary set of scientific images, we generated passages with GPT-Vision and skimmed them for prominent patterns and surprises. We recorded these initial observations in “memos,” a flexible form of taking notes (Blandford et al., 2022a). The goal of this process was to gain familiarity with our data as a whole in preparation for theme exploration. We periodically noticed some trends that we wished to investigate further during this phase. Following the theoretical sampling methodology, we prompted GPT-Vision for more data as insights surfaced from our initial image set (Blandford et al., 2022c). Additional images for “one-off experiments” are included in Appendix Table 3 as P1.1 and T1.1.
在确定了一组初步的科学图像后，我们使用 GPT-Vision 生成了段落，并浏览了其中的突出模式和惊喜。我们将这些初步观察结果记录在“备忘录”中，这是一种灵活的记笔记形式 (Blandford et al., 2022a)。此过程的目的是熟悉我们的整体数据，为主题探索做好准备。我们定期注意到一些趋势，希望在此阶段进一步调查。按照理论抽样方法，随着从初始图像集 (Blandford et al., 2022c) 中浮现出的见解，我们促使 GPT-Vision 获取更多数据。附录表 3 中包含了“一次性实验”的其他图像，作为 P1.1 和 T1.1。

Theme exploration Usually called “open coding” or “open pass” in grounded theory methodology, this phase focused on discovering patterns—typically termed “themes” or “codes”—within the data (Corbin and Strauss, 1990). We carefully read each generated passage, recording themes and evidence (e.g., quotes) in a structured

[1]Images, context, and generated passages can be found at https://github.com/ahwang16/grounded-intuition-gpt-vision.
[2]图像、上下文和生成的段落可在 https://github.com/ahwang16/grounded-intuition-gpt-vision 找到。

Theme	Definition
Linguistic characteristics	General features of text generated by GPT-Vision
• Persona	GPT-Vision's "personality," attitude, or tone of voice
▪ Stoic authority	Matter-of-fact, assertive, straightforward (aloof and certain)
▪ Customer service rep	Conversational, polite, easygoing (engaging and uncertain)
• First-person language	Instances of first-person language (I/me/my/mine, we/us/our/ours)
Figure descriptions	Characteristics of how standalone elements with a caption are described
• Main idea	The purpose or critical message of a figure, if described

Table 1: Examples of finalized themes after theme development. Indentations represent sub-themes that were categorized under a larger parent theme (e.g., "Stoic Authority" is a sub-theme of "Persona," which is a sub-theme of "Linguistic characteristics").

document. Diverging from original methodology, we consulted relevant literature to inform the final themes. We also conducted "aggregate analyses," a new step we established specifically for evaluating generative AI models. In traditional approaches, data is inspected one at a time, with insights from the pool of previous data guiding the next analysis. In an aggregate analysis, we directly compared groups of related passages (e.g., all graphs). At the end of our exploration, we had a hierarchy of 51 preliminary themes like hallucination, numerical reasoning, writing style, and contextual influence.
在扎根理论方法论中，此阶段通常称为"开放式编码"或"开放式传递"，侧重于发现数据中的模式（通常称为"主题"或"代码"（Corbin and Strauss, 1990））。我们仔细阅读了生成的每个段落，在结构化文档中记录了主题和证据（例如引语）。与原始方法不同，我们查阅了相关文献以确定最终主题。我们还进行了"聚合分析"，这是我们专门为评估生成式 AI 模型而建立的新步骤。在传统方法中，一次检查一个数据，使用来自先前数据池的见解指导下一次分析。在聚合分析中，我们直接比较相关段落组（例如所有图表）。在探索结束时，我们得到了 51 个初步主题的层次结构，例如幻觉、数字推理、写作风格和上下文影响。

Theme development We finalized our themes during the theme development phase by renaming, redefining, removing, creating, merging, or splitting themes from the exploration phase as needed. This phase was based on "axial coding" from the original grounded theory methodology, in which themes are grouped together if they share a connection of "axis" of similarity Corbin and Strauss (1990). At the end of the development phase, our finalized hierarchy consisted of 94 themes. We present a sample of these themes in Table 1.[3]
我们在主题开发阶段通过根据需要重命名、重新定义、删除、创建、合并或拆分探索阶段的主题来确定主题。此阶段基于原始扎根理论方法中的"轴向编码"，其中如果主题共享相似性"轴"连接，则将其分组在一起 Corbin and Strauss (1990)。在开发阶段结束时，我们最终确定的层次结构由 94 个主题组成。我们在表 1 中展示了这些主题的一个示例。[4]

Theme application During the final phase of our analysis, we passed through the data another time to apply our finalized themes. For each generated passage, we recorded any overlooked evidence that fit into a theme. The outcome of this phase was a detailed document of themes and evidence. These themes were the building blocks for our findings (see Section 3), analogous to "latent representations" for our ultimate conclusions.
在分析的最后阶段，我们再次检查数据以应用我们最终确定的主题。对于生成的每个段落，我们记录了任何符合主题的被忽略的证据。此阶段的成果是一份详细的主题和证据文件。这些主题是我们研究结果的基石（参见 3 部分），类似于我们最终结论的"潜在表征"。

[3]The full set of themes can be found at https://github.com/ahwang16/grounded-intuition-gpt-vision.
[4]完整的主题集可在 https://github.com/ahwang16/grounded-intuition-gpt-vision 中找到。

3 Findings

In this section, we discuss our findings across all images and prompts. We refer to images with a letter signifying the image type (Photo, Diagram, Graph, Table, Code, Math, or Full page) and a number. A list of all images can be found in Tables 3 (photos, diagrams, graphs, and tables) and 4 (full pages, code, and math). We used two prompts for each image (see Section 2). The first prompt, which we call "alt," is a straightforward request for alt text. The second prompt is identified as "desc" and instructs GPT-Vision to describe the image as though it were speaking with someone who could not see it. The "alt" prompt often resulted in a paragraph with a matter-of-fact tone, while most generated passages for the "desc" prompt were about a page long with varying levels of cheerfulness. All images and generated passages can be found at https://github.com/ahwang16/grounded-intuition-gpt-vision.
在本节中，我们将讨论所有图像和提示中的发现。我们用字母表示图像类型（Photo、Diagram、Graph、Table、Code、Math 或 Full page）和一个数字来表示图像。所有图像的列表可在表格 3（照片、图表、图形和表格）和 4（完整页面、代码和数学）中找到。我们对每幅图像使用了两个提示（参见 2 节）。第一个提示，我们称之为 "alt"，是直接请求 alt 文本。第二个提示被标识为 "desc"，指示 GPT-Vision 描述图像，就像它正在与看不到图像的人交谈一样。"alt" 提示通常会产生一段语气平淡的段落，而 "desc" 提示生成的大多数段落大约有一页长，语气欢快程度各不相同。所有图像和生成的段落都可以在 https://github.com/ahwang16/grounded-intuition-gpt-vision 找到。

3.1 Margin for Error

One of the most apparent patterns of GPT-Vision's writing style was how much margin for error it included in the generated passage. As mentioned in its system card, GPT-Vision can sometimes speak in a matter-of-fact tone (OpenAI, 2023a). Other times, it implies imprecision—this or that, it seems like, and so on.
GPT-Vision 写作风格最明显的模式之一是它在生成的段落中包含了多少错误余地。正如其系统卡中提到的那样，GPT-Vision 有时会用一种平淡无奇的语气说话 (OpenAI, 2023a)。其他时候，它暗示着不精确——这 或那样，它 看起来如此，等等。

Sometimes counterbalancing an error Sometimes, a wide margin for error compensated for a mistake GPT-Vision made, like describing elements in a complicated diagram. D3, the front-page figure representing symbolic knowledge distillation, is particularly complex. Symbolic knowledge distillation involves training a language model to generate commonsense knowledge graphs, which are then used to train other commonsense models (West et al., 2022). GPT-Vision described two parts of D3 with notable margin for error: the standard Apple of a robot's face and a connected graph of nodes and edges, shown below (Figure 1).
有时，较大的误差幅度可以弥补 GPT-Vision 所犯的错误，例如描述复杂图表中的元素。D3 是代表符号知识提炼的头版图表，它特别复杂。符号知识提炼涉及训练语言模型以生成常识知识图谱，然后将其用于训练其他常识模型 (West et al., 2022)。GPT-Vision 描述了 D3 的两个部分，具有明显的误差幅度：机器人脸部的标准 Apple 和节点和边的连接图，如下图所示（图 1）。

Figure 1: Robot emoji and knowledge graph from D3 (West et al., 2022).

In D3, the robot emoji is used multiple times to represent different language models. One of the robots shows only half of its face, revealing one large eye and a red "ear." GPT-Vision described it as
在 D3 中，机器人表情符号被多次使用来代表不同的语言模型。其中一个机器人只露出了半张脸，露出一只大眼睛和一只红色的 "耳朵"。GPT-Vision 将其描述为

> a cute character with a square-shaped body, has a single large eye, and a small red part on its side that I assume represents an arm or [sic] sorts. (D3 desc)

GPT-Vision made a couple errors: it described the half-emoji as having a "square-shaped body" even though it has just a face. It exhibited some margin for error by saying "I assume" a small red part is an arm, which is incorrect. Appropriate margin for error can help the user develop trust in the model, as long as the implied uncertainty matches the actual accuracy of the claim.
GPT-Vision 犯了几个错误：它将半个表情符号描述为具有 "方形身体"，尽管它只有一张脸。它通过说 "我假设" 表现出一定的错误余地，其中一小块红色部分是手臂，这是不正确的。只要隐含的不确定性与声明的实际准确性相匹配，适当的错误余地可以帮助用户对模型产生信任。

P1 desc, T1 desc, T2 desc, F4 desc, D1 desc, G1 alt and desc, G2 desc, and C2 desc contain similar behavior.

Occasionally distracting or excessive While useful for indicating uncertainty, margin for error occasionally cluttered GPT-Vision's output with too much bloat. For example, when describing the connected graph of nodes and edges, it wrote
虽然误差幅度对于表示不确定性很有用，但偶尔会使 GPT-Vision 的输出变得过于臃肿。例如，在描述节点和边的连通图时，它写道:

> there is an illustration of what appears to be some form of inter-connected network, which may be meant to visually represent the structure of a knowledge graph. (D3 desc)

Ironically, GPT-Vision sometimes sounded confident when it was wrong and unsure when it was right. This level of hedging may especially confuse readers without access to the image because they cannot interpret it for themselves.
讽刺的是，GPT-Vision 有时会在错误时显得自信满满，而在正确时则显得犹豫不决。这种程度的含糊其辞可能会让无法看到图片的读者感到困惑，因为他们无法自己解读图片。

C3 desc and M2 desc contain similar behavior.

Often necessary Including some margin for error was often necessary because GPT-Vision's claim could not be verified by the input alone. Even a detail as seemingly obvious as
留出一些误差空间通常是必要的，因为 GPT-Vision 的说法不能仅通过输入来验证。即使是像

> The image is a photo of a section of an academic paper or textbook, focused on a specific topic titled "3.1 Decoder: General Description." (M1 desc)

cannot be confirmed with the input GPT-Vision has been given. It correctly guessed the origin of M1—a section from Bahdanau et al. (2016)—but the image itself does not state that it is from an academic paper. Depending on how confident we are in GPT-Vision's internal knowledge, tuning the margin for error empowers users to make informed decisions with LLM assistance.
无法通过 GPT-Vision 给出的输入进行确认。它正确猜测了 M1 的来源——来自 Bahdanau et al. (2016) 的一个部分——但图像本身并未表明它来自学术论文。根据我们对 GPT-Vision 内部知识的信心程度，调整误差幅度可让用户在 LLM 的帮助下做出明智的决策。

C1 desc, C3 desc, P1 alt and desc, F1 desc, F2 desc, F3 desc, F4 desc, T3 desc, M1 desc, M2 desc, and D1 desc contain similar behavior.

3.2 Hallucination

Hallucination was one of the main vulnerabilities listed in GPT-Vision's system card, but we argue that not all hallucination needs to be avoided. In fact, some forms of hallucination are highly desired.
幻觉是 GPT-Vision 系统卡中列出的主要漏洞之一，但我们认为并非所有幻觉都需要避免。事实上，某些形式的幻觉是人们所期望的。

Hallucination as general knowledge and inference When defining it as information in the output that does not appear in the input, then general knowledge and inference can be considered helpful forms of hallucination.
当将其定义为输出中未出现在输入中的信息时，常识和推理可以被视为有用的幻觉形式。

GPT-Vision displayed several signs of "internal knowledge":
GPT-Vision 表现出了几种 "内部知识" 的迹象：

- "Egg Biryani is an Indian dish" (P1 desc).
- "The page has mathematical symbols and technical terms commonly found in computer science literature" (F5 alt).
- "[The Python code] uses comments (text preceded by a '#' symbol")" (C3 desc).

many of which are accurate. P1 desc, D2 desc, C2 alt and desc, and M2 desc contain similar behavior.
其中许多都是准确的。P1 desc、D2 desc、C2 alt and desc 和 M2 desc 包含类似的行为。

GPT-Vision made reasonable inferences even more often. These claims seemed reasonable given the input but are not stated directly within it, like
GPT-Vision 做出合理推断的次数更多。这些说法在输入的情况下似乎是合理的，但并没有直接说明，比如

- "...another gray dashed horizontal line near the top, labeled 'Human', [indicates] the human-level performance benchmark" (G3 alt).
- "[α_{xy}] probably refers to a certain value that depends on x and y" (C1 desc).

One perspective on natural language inference relates to the model's ability to reason. In our case, we see it as a hallucination that happens to be correct.
自然语言推理的一个观点与模型的推理能力有关。在我们的案例中，我们将其视为恰好正确的幻觉。

C1 desc, C2 desc, D1 desc, D2 desc, D3 desc, F2 desc, G3 desc, M1 desc, and M2 desc contain similar behavior.

Beware of possibly naive assumptions A handful of "interpretations" seem like valid and impressive inferences, but we cannot know for sure without additional studies on internal model mechanisms. In a particularly subtle but impactful instance, GPT-Vision described the trend of the table of errors in T2 as
少数 "解释" 似乎是有效且令人印象深刻的推论，但如果没有对内部模型机制进行额外的研究，我们就无法确定。在一个特别微妙但影响深远的例子中，GPT-Vision 将 T2 中的错误表趋势描述为

	Intersection			Union			Refined method		
Corpus size	Precision	Recall	AER	Precision	Recall	AER	Precision	Recall	AER
0.5K	91.5	71.3	18.7	63.4	91.6	29.0	75.5	84.9	21.1
8K	95.6	82.8	10.6	68.2	94.4	24.2	83.3	90.0	14.2
128K	96.7	90.0	6.3	77.8	96.9	16.1	89.4	94.4	8.7
1470K	96.8	92.3	5.2	84.2	97.6	11.3	91.5	95.5	7.0

Figure 2: A table of performance metrics from Bahdanau et al. (2016) (T2).

> The values in these columns generally decrease as the size of the training corpus increases, indicating improved performance with more data. (T2 alt)

At first glance, this claim seems reasonable—impressive, even. We may be surprised that decreasing values indicate "improved performance," but even this makes sense knowing that the values are error rates. However, we should be careful before assuming that GPT-Vision can "read" tables. This assertion may have been a lucky coincidence because model performance often improves in general as the training corpus increases. Our analysis of "artificial behavior" focuses on capturing these external patterns, which should not be conflated for the underlying processes of "artificial cognition" or the mechanical structures of "artificial neuroscience."
乍一看，这种说法似乎很合理，甚至令人印象深刻。我们可能会惊讶于数值的下降表示"性能提高"，但即使这样，知道这些值是错误率也是有道理的。然而，在假设 GPT-Vision 可以"读取"表格之前，我们应该小心谨慎。这种断言可能是一个幸运的巧合，因为随着训练语料库的增加，模型性能通常会提高。我们对"人工行为"的分析侧重于捕捉这些外部模式，这些模式不应与"人工认知"的底层过程或"人工神经科学"的机械结构混为一谈。

F3 alt, F1 alt, T3 alt, F5 alt, and C3 desc contain similar behavior.

3.3 Incorporation of Source Material

Conversely from hallucinating, GPT-Vision also incorporated source material in a few ways.
与幻觉相反，GPT-Vision 还通过几种方式整合了源材料。

Direct quotes GPT-Vision commonly provided exact section headers, text in diagrams, and publication metadata as direct quotes. Some of these quotes helped describe the structure of the image:
GPT-Vision 通常提供精确的章节标题、图表中的文本和出版元数据作为直接引用。其中一些引文有助于描述图像的结构：

> The left column then lists "CSS CONCEPTS", which look like categories that the article might belong to, and is followed by one entry that reads " • Human-centered computing → Interactive systems and tools." (F1 desc)

GPT-Vision replicated section exactly, down to the bullet point. Future models in human-centered applications can consider elaborating special formatting even more, especially if the content will played by audio books and screen readers.
GPT-Vision 准确复制了该部分，直至要点。以人为本的应用的未来模型可以考虑进一步完善特殊格式，尤其是如果内容将通过有声读物和屏幕阅读器播放的话。

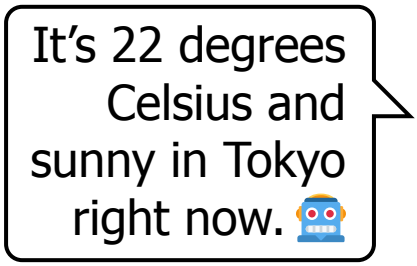


Figure 3: The chatbot's response from D1 (Zhu et al., 2023).

Exact direct quotes are also used to indicate specific words from the image:
精确的直接引用也用于表示图像中的特定词语：

> [The chatbot] says, "It's 22 degrees Celsius and sunny in Tokyo right now." (D1 desc)

Overall, GPT-Vision displayed impressive ability to recognize text in images, which will be a great strength for describing images in general. As noted in (OpenAI, 2023a), it sometimes makes errors, especially when two text elements are close to each other. It understandably made more mistakes on smaller or blurrier text, which it could indicate to the user to help them judge the quality of GPT-Vision's descriptions.
总体而言，GPT-Vision 表现出了令人印象深刻的图像文本识别能力，这对于描述图像来说是一个很大的优势。如 (OpenAI, 2023a) 中所述，它有时会出错，尤其是当两个文本元素彼此靠近时。可以理解的是，它在较小或较模糊的文本上犯了更多错误，它可以向用户显示这些错误，以帮助他们判断 GPT-Vision 描述的质量。

C1 desc, P1 desc, T1 desc, F2 alt, T3 alt, F4 alt, D1 desc, D2 desc, G2 desc, G3 alt and desc, and G4 desc demonstrate similar behavior.

Slightly altered quotes Some of the direct quotes were slightly different from the original text, which may misrepresent the intent of the original author. These altered quotes were occasionally benign, like removing the hyphens in "Herb-Roasted Salmon with Tomato-Avocado Salsa" (F1 alt) or capitalizing "method" in "The columns from left to right are titled 'Corpus size', 'Intersection', 'Union', and 'Refined Method'" (T2 desc).
一些直接引用与原文略有不同，可能歪曲了原作者的意图。这些修改后的引用有时是无害的，例如删除"香草烤鲑鱼配番茄鳄梨莎莎酱"（F1 alt）中的连字符，或将"从左到右的列标题为"语料库大小"、"交集"、"并集"和"精炼方法""（T2 desc）中的"方法"大写。

Other changes were more severe, however. The last name of one of the authors of F5, Frohlich, was misprinted as "Fritsche" even though the other three names were spelled correctly (alt).
但其他改动则更为严重。F5 的作者之一 Frohlich 的姓氏被错误地打印为"Fritsche"，尽管其他三个名字的拼写均正确（alt）。

In addition, GPT-Vision sometimes omitted parts of the text that affected its meaning, such as removing "search" from a figure title:
此外，GPT-Vision 有时会省略影响其含义的部分文本，例如从图形标题中删除"搜索"：

> ...there is a figure titled "Fig. 1. Generators for binary ~~search~~ trees." (F5 alt)

The lead author of this paper confirmed that this modification misrepresents the caption because not all binary trees are binary search trees and binary search trees in particular were important for that figure.
本文的主要作者确认，这种修改歪曲了标题，因为并非所有二叉树都是二叉 search 树，并且二叉搜索树对于该图尤其重要。

We also witnessed one instance of GPT-Vision merging nearby text elements in a figure, which was mentioned in OpenAI (2023a) (D2 alt) When quoting the original source, LLMs should represent the source accurately or indicate where changes were made with brackets, ellipses, or other devices.
我们还看到了 GPT-Vision 在图中合并附近文本元素的一个实例，这在 OpenAI (2023a) (D2 alt) 中提到。引用原始来源时，LLM 应准确表示来源或使用括号、省略号或其他设备指示更改的位置。

"Plagiarism" GPT-Vision often generated text that was very similar to the source. In the following example, the boldface text from the generated passages appears verbatim in the source:
GPT-Vision 生成的文本经常与源文本非常相似。在以下示例中，生成的段落中的 boldface 文本与源文本一模一样：

> This text explains that the goal of the study is to understand how voice assistants can effectively guide people through complex tasks, like following recipes. (F2 alt)

Comparing this passage to the original text makes it sound eerily familiar:
将这段话与原文进行比较，会发现它听起来异常熟悉：

> We designed an observational study to understand how voice assistants can effectively guide people through complex tasks, using recipes as an example. Hwang et al. (2023)

Some examples seem "paraphrased" (emphasis ours)
有些例子似乎是“释义”（emphasis 我们的）

> The context vector is computed by an RNN and relies on a sequence of annotations, with each annotation containing information about the whole input sequence with a focus on surrounding parts of a specific word. M1 alt

but still too similar to the original text to be acceptable (bracketed ellipsis [...] ours).
但仍然与原文太相似，无法接受（括号内的省略号 [...] 我们的）。

> The context vector c_i depends on a sequence of annotations $(h_1, \ldots, h_{T_z})$ [...] Each annotation h_j contains information about the whole input sequence with a strong focus on the parts surrounding the i-th word of the input sequence. [...] The context vector c_i is, then, computed as a weighted sum [...] (Bahdanau et al., 2016)

In the worst case scenario, these kinds of reproduction could be flagged as plagiarism or copyright violation. LLMs have strong potential to help with complex writing tasks from crafting emails to redrafting reports, so they should be carefully tuned to quote significant amounts of reproduced text and paraphrase properly.
在最坏的情况下，这些复制行为可能会被标记为抄袭或侵犯版权。LLM 有很强的潜力帮助完成从撰写电子邮件到重新起草报告等复杂的写作任务，因此应仔细调整以引用大量复制文本并正确解释。

3.4 Sensitivity to Typographical Influence

Sometimes, leaning too much on text in an image for context is risky. GPT-Vision was particularly prone to typographical attacks, reminiscent of its predecessor CLIP (Goh et al., 2021) and related to the vulnerability to the order of images mentioned in GPT-Vision's system card (OpenAI, 2023a).
有时，过于依赖图像中的文本来获取背景信息是有风险的。GPT-Vision 特别容易受到排版攻击，这让人想起了它的前身 CLIP (Goh et al., 2021)，并且与 GPT-Vision 的系统卡 (OpenAI, 2023a) 中提到的图像顺序漏洞有关。

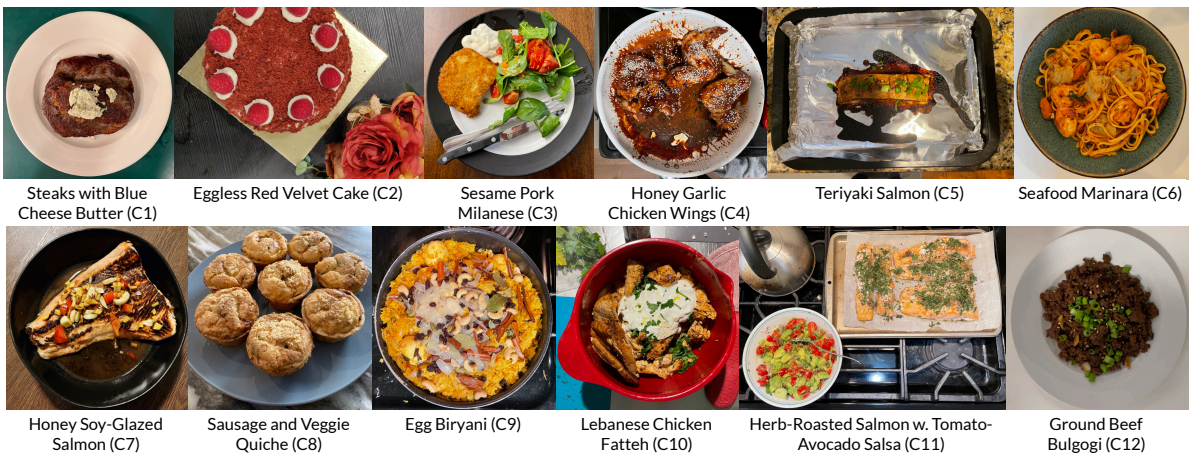


Figure 4: Dishes prepared by participants in a recent study (P1) (Hwang et al., 2023).

Successfully incorporating original labels One of our images, P1, consists of a 2x6 grid of photos showing twelve dishes prepared by participants in a recent study (Hwang et al., 2023). The photos are also labeled with the participant's identification number and the name of the dish underneath each one.
我们的其中一张图片 P1 由 2x6 的照片网格组成，展示了最近一项研究 (Hwang et al., 2023) 中的参与者准备的十二道菜。每张照片下方还标有参与者的身份证号码和菜品名称。

The first photo shows "(C1) Steaks with Blue Cheese Butter," which GPT-Vision aptly described as
第一张照片显示 "(C1) 蓝纹奶酪黄油牛排"，GPT-Vision 恰当地将其描述为

> (C1) A perfectly cooked steak topped with blue cheese butter on a white plate. (P1 alt)

All of the dishes in this passage incorporate the corresponding label in some way, and nearly all of them are excellent, suggesting that text in images can be a helpful source of context.
本文中的所有菜肴都以某种方式包含了相应的标签，并且几乎所有菜肴都非常棒，这表明图片中的文字可以成为有用的背景来源。

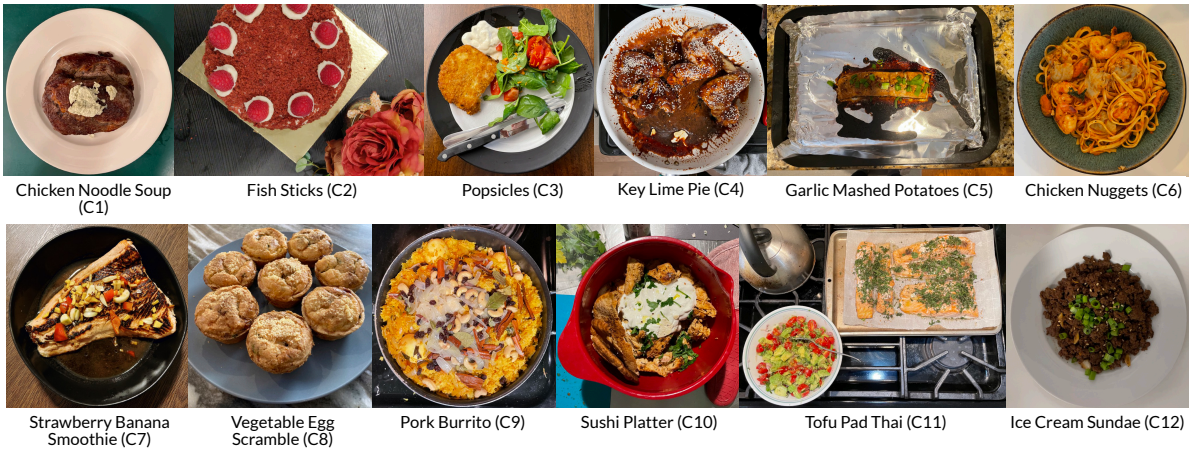


Figure 5: The same figure as P1, but with adversarial labels (P1.1) (Hwang et al., 2023).

However, in a one-off experiment with adversarially modified labels, this blessing turned into a curse. We labeled the same dish as "Chicken Noodle Soup," which GPT-Vision continued to incorporate:
然而，在一次使用对抗性修改标签的一次性实验中，这种祝福变成了诅咒。我们将同一道菜标记为 "鸡肉面条汤"，GPT-Vision 继续将其纳入其中：

> (C1) Chicken Noodle Soup, where a bowl is presented with a dark broth and a dollop of cream... (P1.1 alt)

The photo clearly shows dark, cooked meat topped with a scoop of butter-like dressing, but GPT-Vision still tried to incorporate the new label. All twelve photos were similarly affected in both the "alt" and "desc" generated passages (P1.1).
照片清楚地显示了深色熟肉上浇着一勺黄油状调料，但 GPT-Vision 仍试图融入新标签。所有 12 张照片在 "alt" 和 "desc" 生成的段落中都受到了类似的影响（P1.1）。

Reality check When asked to verify if the given labels were correct and provide alternatives otherwise, GPT-Vision's descriptions improved for both the original
当被要求验证给定的标签是否正确并提供替代方案时，GPT-Vision 的描述对原始的

> ...The label is correct. The photo shows a steak with a pat of blue cheese butter on top.

and adversarial labels.

> The label is incorrect. The photo shows what appears to be a steak with butter on top. The correct label could be “Steak with Butter”.

These corrections sometimes sound very certain (see Section 3.1), leading GPT-Vision to provide some inaccurate descriptions in an authoritative tone. This is a known limitation specified in its system card (OpenAI, 2023a).
这些修正有时听起来非常肯定（参见 3.1 节），导致 GPT-Vision 以权威的语气提供一些不准确的描述。这是其系统卡 (OpenAI, 2023a) 中指定的已知限制。

Hazarding a guess GPT-Vision also performed moderately well when presented the photos without labels (P1.2).
当呈现没有标签的照片（P1.2）时，GPT-Vision 的表现也相当不错。

> ...a seared steak with butter... (P1.2 alt)

GPT-Vision is clearly a powerful vision model, and it can become even more powerful by learning to mitigate the extent to which text in an input image can change the way GPT-Vision talks about it.
GPT-Vision 显然是一个强大的视觉模型，并且通过学习减轻输入图像中的文本改变 GPT-Vision 谈论它的方式的程度，它可以变得更加强大。

3.5 Lossy Expansion

During our investigation, we conducted a one-off experiment to evaluate GPT-Vision’s performance on figures when they are contained within larger images. When describing the two-by-six grid of food photos in P1, the “alt” passage correctly stated there were 12 photos (although it incorrectly characterized the figure as a “4x3 grid”). The “desc” did not specify a total number of photos, but it correctly stated that the figure was “in a two-row rectangular format with six dishes displayed on each row.”
在我们的调查过程中，我们进行了一次一次性实验，以评估 GPT-Vision 在处理包含在较大图像中的图形时的性能。在描述 P1 中的 2x6 食物照片网格时，“alt” 段落正确地指出了有 12 张照片（尽管它错误地将图形描述为 “4x3 网格”）。“desc” 没有指定照片总数，但它正确地指出了图形 “采用两行矩形格式，每行显示六道菜”。

We were expecting to see similar behavior for F2, which is a screenshot of the page that contains P1, but GPT-Vision instead claimed that there were 10 (alt) and 8 (desc) photos in the grid. We attempted to investigate further, but asking GPT-Vision to start analyzing F2 by focusing on the figure first did not improve the results. Asking it directly for the number of photos did not help either. This could cause problems down the line when users are asking for descriptions of arbitrary images. GPT-Vision may err on subcomponents of an image and users may not think to provide the subcomponent on its own and try again, especially if GPT-Vision is providing an accessibility service for an image they cannot see.
我们原本希望看到 F2 的类似行为，F2 是包含 P1 的页面的屏幕截图，但 GPT-Vision 却声称网格中有 10 张照片（alt）和 8 张照片（desc）。我们试图进一步调查，但要求 GPT-Vision 首先关注图形来开始分析 F2 并没有改善结果。直接询问照片数量也无济于事。当用户要求提供任意图像的描述时，这可能会导致后续问题。GPT-Vision 可能会在图像的子组件上出错，用户可能不会想到单独提供子组件并重试，尤其是当 GPT-Vision 为他们看不到的图像提供可访问性服务时。

We noticed similar behavior with the table in T1 and the full page it appears in (F2). T1 is a table from the same study that contains a list of recipe titles matching the names of the dishes in P1. GPT-Vision showed no issues reproducing the recipe titles in T1, but it suddenly started making errors when listing the same recipe titles from the same table in F2. “Eggless Red Velvet Cake” turned into “Eggs Red Velvet Cake” (alt) or “Egg Fried Rice” (desc) and “Sesame Pork Milanese” became “Sesame Pork Medallions,” among other new errors.

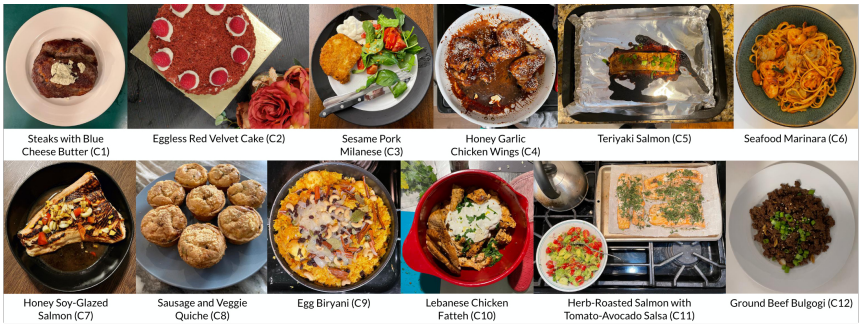


Figure 3: Completed dishes. Cooks prepared a variety of dishes of their choice following the guidance of a voice assistant. These dishes varied in complexity: some required interaction with the voice assistant for many steps (i.e., C2's egglessred velvet cake), while others involved just a few (i.e., C12's ground beef bulgogi).

reviewed the annotated recipe and their thought process with them. Finally, we debriefed the participant and concluded the session. Participants were compensated with a gift card amounting to the cost of ingredients and an additional $100 USD.

3.4 Analysis

Our study yielded four kinds of data: audio and video recordings, researcher notes, questionnaire data, and annotated paper recipes. We used Rev.com to transcribe audio recordings. Notes, transcripts

with pseudonyms C1–12.[2] Quotes from participants are sometimes lightly edited for brevity and clarity.

4.1 Overview

Cooks followed recipes ranging in familiarity, complexity, length, and cultural origin, adding to the richness of their experiences beyond self-reported cooking skill and frequency of using voice assistants. Most recipes were entrées, with two being baked goods (C2, C8) (see Figure 3). Of the 12 cooks, 6 reported being unfamiliar

Figure 6: The beginning of F2, a page from Hwang et al. (2023).

我们注意到 T1 中的表格和它出现的整页 (F2) 有类似的行为。T1 是来自同一项研究的表格，其中包含与 P1 中的菜肴名称相匹配的食谱标题列表。GPT-Vision 在重现 T1 中的食谱标题时没有出现任何问题，但在 F2 中列出同一张表中的相同食谱标题时突然开始出错。除了其他新错误外，“无蛋红丝绒蛋糕”变成了“鸡蛋红丝绒蛋糕”（alt）或“蛋炒饭”（desc），“芝麻猪肉米兰”变成了“芝麻猪肉奖章”。

Our analysis of “artificial behavior” focuses on exposing these patterns rather than inferring why they occur, so we are unsure why GPT-Vision sometimes describes the same elements drastically differently—if this is a frequent problem at all. Our small sample size of 21 images may not have much statistical power, but a phenomenon occurring more than once is bound to be compelling in such few cases.
我们对“人工行为”的分析侧重于揭示这些模式，而不是推断它们发生的原因，因此我们不确定为什么 GPT-Vision 有时会对相同元素做出截然不同的描述——如果这是一个常见问题的话。我们的 21 张图像的小样本可能没有太多的统计能力，但在如此少数的情况下，发生不止一次的现象必然引人注目。

In fact, P1 and T1 were two of only three images that also appeared in full-page screenshots, constraining our sample even more. The third image, P2 could be an outlier: unlike P1 and T1, it does not contain any text. We speculate that the way images are “tokenized” may lead to these errors, and further investigation into the “artificial cognition” and “neuroscience” of these behaviors will hopefully reveal the answer.
事实上，P1 和 T1 是仅有的三张出现在整页截图中的图像中的两张，这进一步限制了我们的样本。第三张图像 P2 可能是一个异常值：与 P1 和 T1 不同，它不包含任何文本。我们推测图像的“标记化”方式可能会导致这些错误，进一步研究这些行为的“人工智能”和“神经科学”有望揭示答案。

3.6 Taking Context into Consideration

P2 is a photo of a female study participant cooking a dish in a kitchen with the guidance of a voice assistant. The voice assistant, a blue spherical Alexa Echo Dot, is displayed on the right side of the image with a circle around it.
P2 是一张女性研究参与者在厨房里在语音助手的指导下烹饪菜肴的照片。语音助手是一个蓝色球形 Alexa Echo Dot，显示在图片右侧，周围有一个圆圈。

Although the type of device circled in image is not immediately clear, the caption that was given as additional context stated that “[participants] followed recipes of their own choice with the help of Amazon

Figure 1: Study setting. Participants followed recipes of their choice with the help of Amazon Alexa (Echo Dot, circled on the right). Participants were observed at home and encouraged to cook however felt natural as we observed, only occasionally asking clarifying questions. We filmed the session with a camera on a tripod out of the way of the kitchen.

ID	Selected Recipe	Self-Rated Skill	Prior Use
C1	Steaks with Blue Cheese Butter	■■■■·	daily
C2	Eggless Red Velvet Cake	■■■■·	weekly
C3	Sesame Pork Milanese	■■····	<monthly
C4	Honey Garlic Chicken Wings	■■■···	<monthly
C5	Teriyaki Salmon	■·····	monthly
C6	Seafood Marinara	■■····	never
C7	Honey Soy-Glazed Salmon	■■■■·	never
C8	Sausage and Veggie Quiche	■■····	daily
C9	Egg Biryani	■■■■■	weekly
C10	Herb-Roasted Salmon with Tomato-Avocado Salsa	■■■■·	weekly
C11	Lebanese Chicken Fatteh	■■····	never
C12	Ground Beef Bulgogi	■■■···	weekly

Table 1: Participants. Participants were mostly graduate students and chose a wide variety of recipes to prepare. They represented a range of cooking skill ("Self-Rated Skill" on a 5-point Likert scale) and frequency of voice assistant usage ("Prior Use").

research has also indicated the nuance involved in helping users navigate sets of instructions with a voice interface. Abdolrahmani et al. [1] propose that voice assistants in complex environments like an airport provide support through short transactions. Other work has suggested that interfaces should support multiple kinds of pauses and jumps [14], handle implicit conversation cues [53], and support jumps according to both conventional navigation instructions and content-based anchors [64]. Our paper contributes a deep, validated, actionable design inspiration while being possible to arrange in a way that a full contextual inquiry would not be.

3.1 Technology Probe

Figure 7: An excerpt from F3, the third page of Hwang et al. (2023).

Alexa (Echo Dot, dircled on the right)" (Hwang et al., 2023).
虽然图片中圈出的设备类型并不明显，但作为附加背景的标题指出，"[参与者] 在亚马逊 Alexa（Echo Dot，右侧带圆圈）的帮助下按照自己选择的食谱烹饪"（Hwang et al., 2023）。

Taking a hint GPT-Vision incorporated this information well for the "desc" prompt,
GPT-Vision 很好地将这些信息融入了 "desc" 提示中，

> What stands out is an Amazon Alexa Echo Dot, which is circled for emphasis. It is placed to the far right on the countertop near some other kitchen tools. (P2 desc)

Missing the point but not for the "alt" prompt.
但不适用于 "alt" 提示。

Figure 8: A photo of a participant cooking in a kitchen with an Amazon Alexa Echo Dot circled on the right (P2) (Hwang et al., 2023).

There is a small circular clock with a white frame hanging on the wall, indicated by a circle. (P2 alt)

In most cases, information from the context did not appear in the generated passages. This suggests that GPT-Vision "ignored" it, but we cannot know for sure based on our behavioral evaluation. We do, however, have evidence that GPT-Vision has the capacity to leverage text and image inputs at the same time.
在大多数情况下，上下文中的信息不会出现在生成的段落中。这表明 GPT-Vision "忽略" 了它，但根据我们的行为评估，我们无法确定。然而，我们确实有证据表明 GPT-Vision 有能力同时利用文本和图像输入。

D3 alt, P1 desc, P2 desc, and G2 alt show similar examples of incorporating context. D3 alt and desc, T1 alt and desc, P2 alt, T2 alt, T3 alt and desc, D1 alt and desc, G1 alt and desc, G2 alt and desc, and G4 alt and desc contain similar behavior of lacking context.
D3 alt、P1 desc、P2 desc 和 G2 alt 显示了类似的包含上下文的示例。D3 alt and desc、T1 alt and desc、P2 alt、T2 alt、T3 alt and desc、D1 alt and desc、G1 alt and desc、G2 alt and desc 和 G4 alt and desc 包含类似的缺乏上下文的行为。

3.7 Code-to-English Translation

In general, GPT-Vision's descriptions of code demonstrated some internalized knowledge of programming languages at a high-level. The specificity of the Python description compared to Haskell suggests a deeper knowledge of Python, while its fluent "translation" of pseudocode to natural language indicates good potential in the programming space.
总体而言，GPT-Vision 的代码描述展示了一些高级编程语言的内部知识。与 Haskell 相比，Python 描述的特殊性表明了对 Python 的更深层次的了解，而其将伪代码流畅地 "翻译" 为自然语言的能力表明了编程领域的巨大潜力。

```python
def build_prompt(self, messages: list[ChatMessage], functions: list[AIFunction] | None = None):
  tokens = []
  prompt_buf = []  # parts of the user-assistant pair
  for message in messages:
    if message.role == ChatRole.USER:
      prompt_buf.append(f"{B_INST} {message.content} {E_INST}")
    elif message.role == ChatRole.ASSISTANT:
      prompt_buf.append(f" {message.content} ")
      # turn the current round into tokens
      prompt_round = "".join(prompt_buf)
      # if we see a " {E_INST}{B_INST} " we should replace it with empty string
      # (it happens immediately after a system + user message)
      prompt_round.replace(f" {E_INST}{B_INST} ", "")
      tokens.extend(self.tokenizer(prompt_round))
      # tokenizer adds the BOS token but not the EOS token
      tokens.append(eos_token_id)
      prompt_buf.clear()
    else:
      prompt_buf.append(f"{B_INST} {B_SYS}{message.content}{E_SYS} {E_INST}")
  # flush rest of prompt buffer (probably a user message) into tokens
  if prompt_buf:
    tokens.extend(self.tokenizer("".join(prompt_buf)))
  return torch.tensor([tokens], device=self.device)
```

Figure 9: The build_prompt() method from C3 (Zhu et al., 2023).

Python GPT-Vision correctly indicates that C3 "contains a screenshot of Python code which defines a class named 'LlamaEngine' that inherits from 'HuggingEngine'" (C3 alt, and that the class "has three methods:

'___init___', 'build_prompt', and 'message_len'" (C3 alt)
GPT-Vision 正确地指出 C3 "包含一个 Python 代码截图，该代码定义了一个名为 "LlamaEngine" 的类，该类继承自 "HuggingEngine" "（C3 alt），并且该类 "有三种方法：'___init___'、'build_prompt' 和 'message_len'"（C3 alt）

It even elaborates on the methods, such as correctly stating that "the 'build_prompt' is meant for building and tokenizing a prompt from a user-assistant conversation, using incoming messages and functions. It accepts messages and functions as parameters and appends tokens to build a prompt'" (C3 alt)
它甚至详细阐述了这些方法，例如正确地指出 "'build_prompt' 用于使用传入的消息和函数从用户助理对话中构建和标记提示。它接受消息和函数作为参数并附加标记以构建提示'"（C3 alt）

When queried with the "desc" prompt, GPT-Vision provides even more specific details:
当使用 "desc" 提示进行查询时，GPT-Vision 提供了更具体的详细信息：

> the `build_prompt` method... accepts two parameters: 'self', which is standard for class methods, and 'messages', which is expected to be a list of ChatMessage. (C3 desc)

These details, while correct, may not be the most helpful overview of a piece of such a sophisticated code. It incorporates special tokens depending on the chat role and specifies particular Python type annotations. GPT-Vision even misprints one of the types as "SomeFunction" rather than "AIFunction" (C3 desc). Generative AI models describing code should look for the most critical structures within it, which may not be the most obvious pieces.
这些细节虽然正确，但可能不是对如此复杂代码片段最有用的概述。它根据聊天角色包含特殊标记，并指定特定的 Python 类型注释。GPT-Vision 甚至将其中一种类型错误地打印为 "SomeFunction" 而不是 "AIFunction"（C3 desc）。描述代码的生成式 AI 模型应该在其中寻找最关 键的结构，而这些结构可能不是最明显的部分。

ALGORITHM 1: Algorithm OnlineMatch

require: Scaling parameter $\gamma \in (0, 1]$
input : Online arrivals of agents

```
(α_xy) := Solution to LP-UB;
for each agent i arriving at time t, say of type y ∈ X do
    for each type x ∈ X in a uniformly random order do
        if there is at least one unmatched agent j of type x in the market then
            match i and j with probability γ · α_xy · max(1, μ_x / λ_x)
    end
end
```

Figure 10: A pseudocode algorithm for a dynamic matching (C1) (Collina et al., 2021).

Pseudocode GPT-Vision also "translates" pseudocode to natural language quite well, besides some errors in reproducing mathematical text (see Section 3.12 for more details). C1 contains an if statement in two nested for loops to demonstrate an dynamic matching algorithm. Instead of reproducing the pseudocode verbatim, it describes it more generally:
GPT-Vision 还可以很好地将伪代码 "翻译" 成自然语言，除了在重现数学文本时出现一些错误（有关详细信息，请参阅第 3.12 节）。C1 在两个嵌套的 for 循环中包含一个 if 语句，以演示动态匹配算法。它没有逐字重现伪代码，而是更笼统地描述了它：

> The algorithm starts by calculating the solution to 'LP-UB' and stores it in 'αxy'. Then, for each agent 'i' arriving at time 't' of a certain type 'y', and for each type 'x' in a random order, it check if there's at least one unmatched agent 'j' of type 'x'. If so, agent 'i' is matched with 'j' with a calculated probability dependent on 'γ' and 'αxy'. The algorithm terminates after processing all agents. (C1 alt)

When responding to the "desc" prompt, it breaks the pseudocode down into four steps:
当响应 "desc" 提示时，它将伪代码分解为四个步骤：

1. First, it calculates "(a_xy)"...
2. The algorithm then enters a loop where for each agent "i" ..., it further loops through each type "x"...
3. Inside this nested loop, there is a conditional statement...
4. If this condition is true, the algorithm matches agent "i" and "j"... (C1 desc)

Haskell Descriptions of Haskell, however, tend to be much more superficial, like stating that C2 starts with "the definition for a data type called `Freer`, followed by definitions for `Return` and `Bind`" (desc, or "[the] type alias `Reflective` [is] defined as `Freer (R b)`" (alt) The difference between GPT-Vision's behavior with Python and Haskell may suggest that it is "more familiar" with the former.
然而，对 Haskell 的描述往往更加肤浅，例如，C2 以 "名为 `Freer` 的数据类型的定义" 开头，后跟 `Return` 和 `Bind`" 的定义（desc，或 "类型别名 `Reflective` [被] 定义为 `Freer (R b)`" （alt）GPT-Vision 与 Python 和 Haskell 的行为之间的差异可能表明它 "更熟悉" 前者。

C PROOFS OF LEMMA 4.1 (LAWS)

This appendix proves the equations from Lemma 4.1.

(*M*1) `return a >>= f = f a`
(*M*3) `(x >>= f) >>= g = x >>= (\ a -> f a >>= g)`
(*PMP*3) `(lmap f . prune) (return y) = return y`
(*PMP*4) `(lmap f . prune) (x >>= g) = (lmap f . prune) x >>= lmap f . prune . g`

Using the following relevant definitions:

```
data Freer f a where
  Return :: a -> Freer f a
  Bind :: f a -> (a -> Freer f c) -> Freer f c

data R b a where
  Pick :: [(Weight, Choice, Reflective b a)] -> R b a
  Lmap :: (c -> d) -> R d a -> R c a
  Prune :: R b a -> R (Maybe b) a
```

Figure 11: The beginning of a Haskell proof (C2) (Goldstein et al., 2023).

3.8 Visions of Summarization

To our surprise, when given screenshots of full images, GPT-Vision often showed signs of summarizing paragraphs within them. The ability for vision models to handle sophisticated language tasks like summarization opens many opportunities for them to handle dense, text-dominant documents as well as the images within them.
令我们惊讶的是，当给出完整图片的截图时，GPT-Vision 经常显示出在其中总结段落的迹象。视觉模型能够处理诸如总结之类的复杂语言任务，这为它们处理密集的、以文本为主的文档以及其中的图像提供了许多机会。

One full-page screenshot (F3) that GPT-Vision started to summarize was the third page of Hwang et al. (2023), which discusses a human-computer interaction study on how voice assistants tend to deliver complex instructions (see Figure 12). It looked like a coherent summary at first glance, but a deeper look showed us that most of it was composed of paraphrased sentences from throughout the document. After detailing the layout of the two-column paper, GPT-Vision described portions of the page starting with the left column:

GPT-Vision 开始总结的一张整页截图（F3）是 Hwang et al. (2023) 的第三页，该页讨论了一项关于语音助手如何传递复杂指令的人机交互研究（见图 12）。乍一看，它看起来像是一个连贯的摘要，但仔细观察后我们发现，其中大部分内容都是由整篇文档中的释义句子组成的。在详细介绍了两栏论文的布局后，GPT-Vision 从左栏开始描述了页面的部分内容：

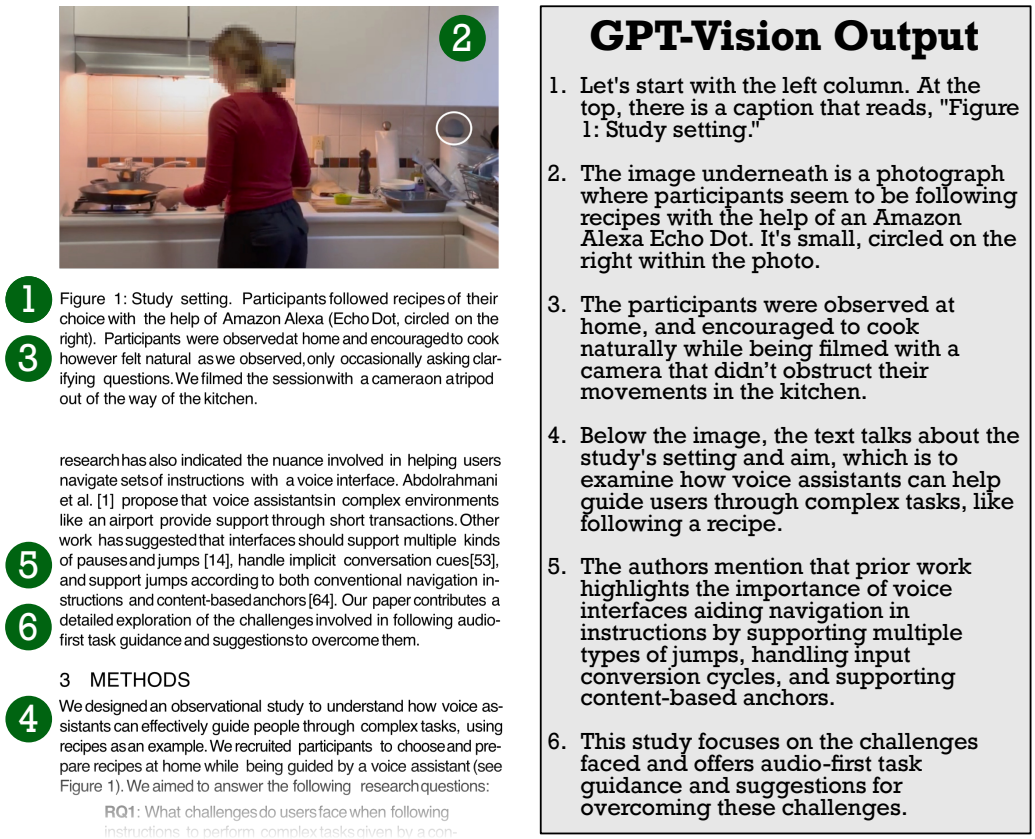


Figure 12: Part of the left column of F3 and a portion of GPT-Vision's description of it (Hwang et al., 2023).

> Let's start with the left column. At the top, there is a caption that reads, "Figure 1: Study setting." The image underneath is a photograph where participants seem to be following recipes with the help of an Amazon Alexa Echo Dot. It's small, circled on the right within the photo.

This part of the summary quoted and paraphrased the caption of Figure 1, which is at the top of the left column. GPT-Vision displayed a strong tendency of moving from left to right and top to bottom, but this path may not be ideal for reading a scientific paper. Readers may prefer to read the figure after it has been referenced in the main body of the paper. Ideally, a human-centered tool would be able to adapt to individual preferences, which seems within the realm of possibility for the current state of generative AI.
摘要的这一部分引用并解释了图 1 的标题，该标题位于左栏顶部。GPT-Vision 表现出强烈的从左到右、从上到下的移动趋势，但这种路径可能不适合阅读科学论文。读者可能更喜欢在论文正文中引用该图后再阅读。理想情况下，以人为本的工具能够适应个人偏好，这似乎是生成式 AI 当前状态下可能实现的。

Next, GPT-Vision writes,

> The participants were observed at home, and encouraged to cook naturally while being filmed with a camera that didn't obstruct their movements in the kitchen.

This sentence is very similar to the back half of the caption, which will become important as we read more of GPT-Vision's summary. It suddenly jumps to the beginning Section 3 (Methods), which is at the bottom of

the left column:
这句话与标题的后半部分非常相似，当我们阅读更多 GPT-Vision 的摘要时，这将变得非常重要。它突然跳转到开头的第 3 节（方法），该节位于左栏的底部：

> Below the image, the text talks about the study's setting and aim, which is to examine how voice assistants can help guide users through complex tasks, like following a recipe.

However, Section 3 is not immediately below figure in the page, contrary to what it might imply by describing it immediately following the figure. It then jumps backward to the paragraph between the figure and Section 3, which is the end of a section continued from the previous page.
但是，第 3 节并不在页面中的图的正下方，这与紧跟在图之后描述它所暗示的相反。然后它跳转到图和第 3 节之间的段落，这是上一页的延续部分的结尾。

> The authors mention that prior work highlights the importance of voice interfaces aiding navigation in instructions by supporting multiple types of jumps, handling input conversion cycles, and supporting content-based anchors.

This excerpt once again closely resembles the original text, but "handling input conversion cycles" has no meaning in this context. It seems like a misreading or misinterpretation of "handle implicit conversation cues."
这段摘录再次与原文非常相似，但“处理输入转换周期”在此上下文中毫无意义。这似乎是对“处理隐含对话线索”的误读或误解。

The closing sentence about the left column features a close paraphrase of the last sentence of section 2.4.
关于左栏的结束语与第 2.4 节最后一句话密切相关。

> This study focuses on the challenges faced and offers audio-first task guidance and suggestions for overcoming these challenges.

The "summary" of the left column covers very little of it: it paraphrases a the figure caption, a sentence from Section 3, and a couple of sentences from Section 2.4 while omitting key information like the research questions and study design choices. Furthermore, GPT-Vision may not be performing a language task after all—it may be picking visual details to relay, much like it picks elements of diagrams to describe. It also collapsed the entire left column into one section even though scientific papers are not meant to be read that way. Vision models that present syntheses of text material should make sure to represent the full scope well, at the risk of readers unknowingly missing crucial points.
左栏的“摘要”涵盖的内容很少：它解释了图表标题、第 3 节中的一句话和第 2.4 节中的几句话，同时省略了关键信息，例如研究问题和研究设计选择。此外，GPT-Vision 可能根本不是在执行语言任务——它可能只是挑选视觉细节来传达，就像它挑选图表元素来描述一样。它还将整个左栏折叠成一个部分，尽管科学论文并不是以这种方式阅读的。呈现文本材料综合的视觉模型应确保很好地代表整个范围，否则读者可能会在不知不觉中错过关键点。

3.9 Respecting Boundaries

GPT-Vision adeptly described individual elements in many generated passages, but it diminished when speaking of overlapping elements in D1. D1 is an overview of Kani, a framework for building chat-based applications (Zhu et al., 2023). The diagram shows a cartoon avatar of a user talking with a chatbot that is powered by Kani. Kani contains three components, which are represented as three rectangles. GPT-Vision described the Kani square as follows:
GPT-Vision 熟练地描述了许多生成段落中的单个元素，但在谈到 D1 中的重叠元素时，它的表现有所减弱。D1 是 Kani 的概述，Kani 是一个用于构建基于聊天的应用程序 (Zhu et al., 2023) 的框架。该图显示了一个卡通头像，其中的用户正在与由 Kani 提供支持的聊天机器人交谈。Kani 包含三个组件，它们表示为三个矩形。GPT-Vision 对 Kani 方块的描述如下：

> The top section of this [the Kani square] shows a chat history window. The window has the name 'Kani' at the top and next to the name there is a red crab icon... Just below the chat history window, there is a section called 'Function Context.' ... It's in a box with rounded edges and a light yellow background. (D1 desc)

GPT-Vision blurs a few details here. The label "Kani," and the crab, does exist at top of the Kani square, but it is outside the rectangle labeled "Chat History." GPT-Vision correctly states that chat history and function context are the top two rectangles in that order, but it stated the wrong color. Function context is actually pink. This was not the only time GPT-Vision seemed to mix up nearby elements (see P1). Identifying absolute positions is a great start, but vision models need to interpret structural relationships well to represent the full range of images properly.
GPT-Vision 在这里模糊了一些细节。标签 "Kani" 和螃蟹确实存在于 Kani 方块的顶部，但它位于标有 "聊天记录" 的矩形的外面。GPT-Vision 正确地指出聊天记录和功能上下文是按该顺序排列的前两个矩形，但它指出了错误的颜色。功能上下文实际上是粉红色的。这并不是 GPT-Vision 唯一一次混淆附近的元素（参见 P1）。识别绝对位置是一个很好的开始，但视觉模型需要很好地解释结构关系，才能正确地表示整个图像范围。

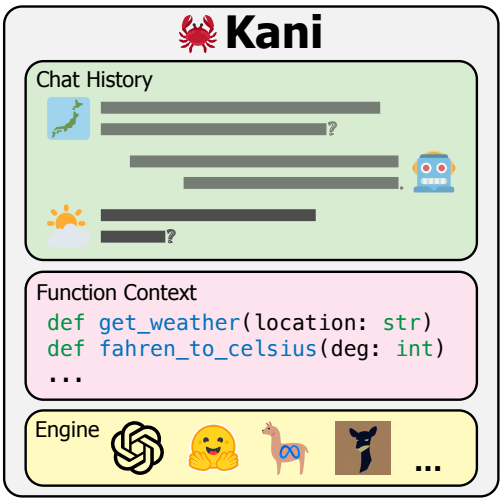


Figure 13: Excerpt from D1, an overview of "Kani" a framework for building chat-based LLM applications (Zhu et al., 2023).

3.10 Spatial Relationships

One of GPT-Vision's most consistent successes was in describing the positions of elements in an image. When describing a complex diagram about symbolic knowledge distillation by West et al. (2022), GPT-Vision accurately stated where each piece of the diagram was located:
GPT-Vision 最持续的成功之一是描述图像中元素的位置。在描述 West et al. (2022) 的关于符号知识提炼的复杂图表时，GPT-Vision 准确地说明了图表中每个部分的位置：

> [In] the top left corner, there's a cartoonish depiction of a robot [...] Next to this robot character, in the top center of the image, there's some text that reads "GPT-3" with three bullet points below it saying ""175B Parameters"", ""General Model"" (D3 desc)

Even though the "three bullet points" do not exist in the image, GPT-Vision described the positions of the elements, and the elements themselves, very well.
即使图像中不存在 "三个要点"，GPT-Vision 也很好地描述了元素的位置以及元素本身。

One of GPT-Vision's frequent weaknesses, however, was in describing the relationships between these elements. "GPT-3," "175B Parameters," and "General Model" are not arbitrary floating pieces of text; they are labels that describe what the robot represents. GPT-Vision did manage to present this detail in the "alt" prompt:
然而，GPT-Vision 经常出现的弱点之一是描述这些元素之间的关系。"GPT-3"、"175B 参数" 和 "通用模型" 不是任意浮动的文本；它们是描述机器人所代表内容的标签。GPT-Vision 确实设法在 "alt" 提示中呈现了这个细节：

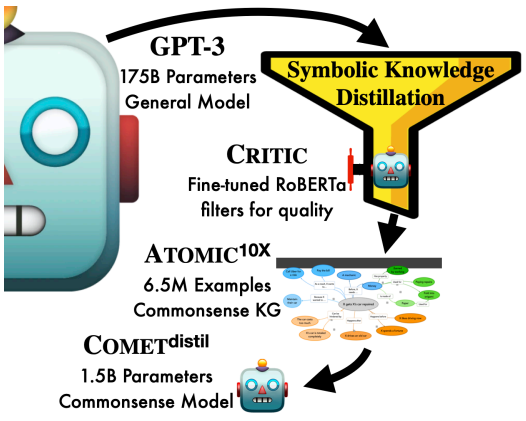


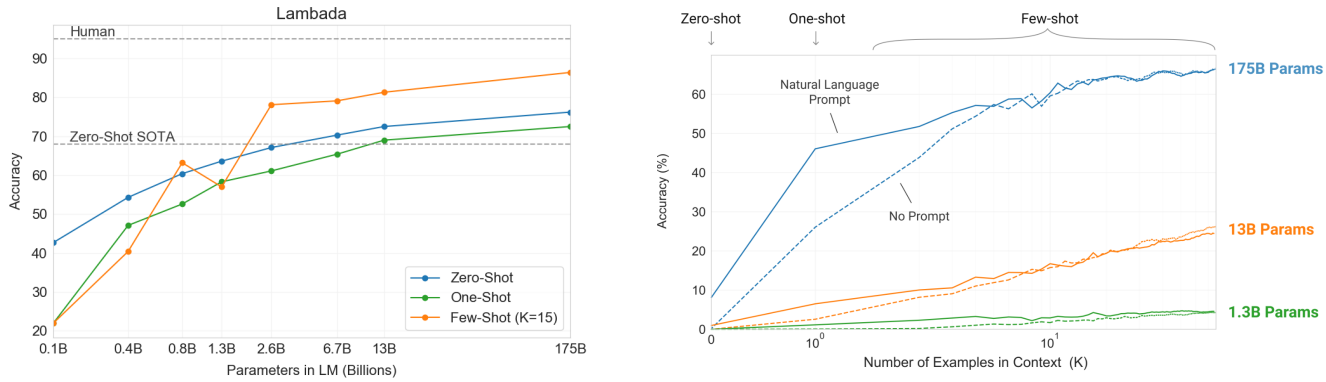
Figure 14: Symbolic knowledge distillation (D3) (West et al., 2022).

On the upper left corner, there is an illustration of a robot, representing GPT-3 which has 175 billion parameters and is labeled as a General Model. (G4 alt)

GPT-Vision often described the same images differently when responding to different prompts, amplifying the challenge of finding coherent patterns in its behavior. We should investigate GPT-Vision with a larger number of samples, experiment with controlled changes in prompts, and unveil the cognitive and neural structures beneath the behavior to learn more.
GPT-Vision 在响应不同的提示时通常会以不同的方式描述同一幅图像，这增加了在其行为中寻找连贯模式的难度。我们应该用更多的样本来研究 GPT-Vision，尝试对提示进行受控更改，并揭示行为背后的认知和神经结构，以了解更多信息。

P1 alt and desc, P2 desc, D2 alt and desc, D3 alt and desc, G2 desc, and G3 desc, F2 alt and desc, F1 desc, F3 desc, F4 alt and desc, F5 desc, C2 desc, and M1 desc contain mentions of spatial relationships.
P1 alt and desc、P2 desc、D2 alt and desc、D3 alt and desc、G2 desc 和 G3 desc、F2 alt and desc、F1 desc、F3 desc、F4 alt and desc、F5 desc、C2 desc 和 M1 desc 包含空间关系的提及。

3.11 Graphic Misinterpretations



(a) G3, a line graph portraying the accuracy of zero-shot, one-shot, and few-shot prompting on the LAMBADA dataset as language model size increases.

(b) G4, a line graph suggesting that increased model size leads to improved in-context learning abilities.

Figure 15: Line graphs from Brown et al. (2020).

GPT-Vision struggled with graphs, like G3 and G4. Both are line graphs sourced from the publication introduced GPT-3, a text-only predecessor to GPT-Vision (Brown et al., 2020). G3 (Figure 15a shows the accuracy of zero-shot, one-shot, and few-shot prompting as GPT-3 increases from 0.1 to 175 billion parameters.
GPT-Vision 在处理 G3 和 G4 等图表时遇到了困难。这两张图都是线图，来源于介绍 GPT-3 的出版物，GPT-3 是 GPT-Vision (Brown et al., 2020) 的纯文本前身。G3（图 15a 显示了 GPT-3 从 0.1 增加到 1750 亿个参数时零样本、一次样本和少量样本提示的准确率。

G4 (Figure 15b) is more complex and shows the accuracy of 1.3 billion–, 13 billion–, and 175 billion–parameter versions of GPT-3 as the number of in-context examples grows from 0 to 32. Each model size is represented by two lines: a solid line for a “Natural Language Prompt” and a dashed line for “No Prompt.” The graph contains free-floating labels for prompt style and model size as opposed to a legend, like in G3.
G4（图 15b）更为复杂，随着上下文示例的数量从 0 增长到 32，GPT-3 的 13 亿、130 亿和 1750 亿个参数版本的准确率分别达到了惊人的 96

Axes GPT-Vision described the x- and y-axes of each line graph moderately well, except that it consistently underestimated the bounds of the axes depending on the labels. For example, the y-axis in G4 is labeled from 0 to 60 in increments of 10, but the line itself extends to 70 without a tick label for $y = 70$:
GPT-Vision 对每条线图的 x 轴和 y 轴的描述相当好，但它始终低估了轴的边界（具体取决于标签）。例如，G4 中的 y 轴以 10 为增量从 0 到 60 进行标记，但线本身延伸到 70，而 y=70 时没有刻度标签：

> The y-axis, or vertical axis, is labeled "Accuracy (%)" and has a linear scale ranging from 0 to 60. (G4 desc)

This seems to be a stylistic trend because the bar graph (G1) and both line graphs (G3, G4) omit the tick label for the greatest value on the y-axis. GPT-Vision mistook the bounds of an axis when describing all three of these graphs, which was particularly precarious when the data went beyond the printed bounds (G3, G4). GPT-Vision appeared to have a bias toward text in an image when it incorporated adversarial labels into its output (see Section 3.4). Future work in "artificial cognition" to expose what GPT-Vision pays attention to can help mitigate this weakness.
这似乎是一种风格趋势，因为条形图 (G1) 和两个折线图 (G3、G4) 都省略了 y 轴上最大值的刻度标签。GPT-Vision 在描述这三个图表时都误解了轴的边界，当数据超出打印边界 (G3、G4) 时，这种情况尤其危险。当 GPT-Vision 将对抗性标签纳入其输出时，它似乎对图像中的文本有偏见（参见 3.4 节）。未来在 "人工智能" 领域的工作将揭示 GPT-Vision 关注的重点，这将有助于弥补这一弱点。

GPT-Vision made a subtler text-based error when describing the x-axis in G4:
GPT-Vision 在描述 G4 中的 x 轴时犯了一个更微妙的基于文本的错误：

> The x-axis... has a logarithmic scale, starting at 10^0 and increasing to 10^1. (G4 desc)

The x-axis is labeled with "0," "10^0," and "10^1," reminiscent of a logarithmic scale, but "10^1" is further from "10^0" than "10^0" is from "0." These values would be equally spaced on a true logarithmic scale. Successfully reading axes partially requires the ability to judge visual distance because we estimate values on a graph by examining how close a point is to a value on a number line. GPT-Vision has already shown a good start in describing positions of elements in an image (see Section 3.10), inspecting how well GPT-Vision describes the amount of space between two points is a natural next step.
x 轴上标有 "0"、"10^0" 和 "10^1"，让人联想到对数刻度，但 "10^1" 与 "10^0" 之间的距离比 "10^0" 与 "0" 之间的距离要远。这些值在真正的对数刻度上是等距的。成功读取轴部分需要判断视觉距离的能力，因为我们通过检查一个点与数轴上的值的接近程度来估计图表上的值。GPT-Vision 在描述图像中元素的位置方面已经取得了良好的开端（参见 3.10 节），检查 GPT-Vision 对两点之间空间量的描述能力是自然而然的下一步。

Data trends GPT-Vision imprecisely represented data trends in both line graphs. G3, for example, displays three solid lines with circle markers for "Zero-Shot" (blue), "One-Shot" (green), and "Few-Shot (K=15)" (orange) prompting. It described the lines qualitatively with some clarity:
GPT-Vision 在两个折线图中都无法精确地表示数据趋势。例如，G3 显示三条带有圆圈标记的实线，分别表示 "零次测试"（蓝色）、"一次测试"（绿色）和 "少量测试（K=15）"（橙色）。它对这些线条进行了定性的描述，并有一定的清晰度：

> The third line, depicted in blue and labeled "Zero-Shot", appears to be an upward leaning curve... The fourth line, represented in green and labeled "One-Shot", is similar to the third but starts at a slightly higher accuracy... Lastly, an orange line labeled "Few-Shot (K=15)"... increases quite sharply... (G3 desc)

The Zero-Shot and One-Shot curves do look similar to each other, starting lower and rising gently. The One-Shot line, however, starts at a substantially lower accuracy (about 20%) than Zero-Shot (about 40%).
零次和单次曲线看起来确实很相似，起始点较低，然后缓慢上升。然而，单次曲线的起始准确度（约 20%）比零次（约 40%）低得多。

Numerical estimates GPT-Vision also imprecisely estimated the starting and ending values of each curve: it noted Zero-Shot as ranging from 30% to 60% (closer to 40%–75%), One-Shot from 40% to 70% (closer to 20%–70%), and Few-Shot (K=15) from 35% to 90% (closer to 20%–85%). It stated that Few-Shot (K=15) "surpass[ed] the One-Shot accuracy at around 2.6 billion parameters," which is inaccurate as well—Few-Shot surpassed One-Shot much earlier, between 0.4 and 0.8 billion parameters.
GPT-Vision 还不精确地估计了每条曲线的起始值和结束值：它指出 Zero-Shot 的范围为 30% 到 60%（接近 40%-75%），One-Shot 的范围为 40% 到 70%（接近 20%-70%），Few-Shot (K=15) 的范围为 35% 到 90%（接近 20%-85%）。它指出 Few-Shot (K=15) "在约 26 亿个参数上超过了 One-Shot 的准确率"，这也是不准确的——Few-Shot 更早地就超过了 One-Shot，在 4 亿到 8 亿个参数之间。

Similar behavior occurred in both generated passages for G4 as well. The line representing GPT-3 1.3B starts at 0% accuracy and remains nearly flat, but GPT-Vision described it as
G4 生成的两个段落也出现了类似的行为。代表 GPT-3 1.3B 的线从 0

> barely rising above 10% accuracy as the number of examples increases (G4 desc)

which implies that the model achieved at least 10% accuracy. However, G4 shows GPT-3 1.3B remaining well under the 10% grid line, which the response to the "alt" prompt actually describes appropriately.
这意味着该模型至少实现了 10

> 1.3 billion parameter model has the least accuracy, remaining below 10%... (G4 alt)

These mixed insights hint at GPT-Vision's emerging graph-reading abilities, especially when it described the shapes of the lines in G3 and G4. Teaching GPT-Vision to read axes properly would allow it to make deeper insights about complex data, and maybe even uncover some unnoticed trends.
这些混合见解暗示了 GPT-Vision 新兴的图形阅读能力，尤其是当它描述 G3 和 G4 中的线条形状时。教 GPT-Vision 正确读取轴将使其能够更深入地了解复杂数据，甚至可能发现一些未被注意到的趋势。

3.12 Writing the Math Out

GPT-Vision generated numerous errors when reproducing mathematical text, from misprinting "1^5" as "1^5^2" (T3 desc), to misrepresenting "$\left(\alpha - \frac{(1-\alpha)2mw(\lceil m \rceil)}{k-1}\right)$" as "$(\alpha - 1/(2k-2))$" (C1 alt). These errors can have drastic consequences if they are not corrected or verified. In addition, GPT-Vision often produced LaTeX-style, some of which would not have compiled (T3 desc, M1 and desc, and C1 alt and desc).
GPT-Vision 在重现数学文本时会产生大量错误，从将"1^5"误印为"1^5^2"（T3 desc），到将"$\left(\alpha - \frac{(1-\alpha)2mw(\lceil m \rceil)}{k-1}\right)$"错误表示为 "$(\alpha - 1/(2k-2))$"（C1 alt）。如果不进行纠正或验证，这些错误可能会产生严重后果。此外，GPT-Vision 经常生成 LaTeX 样式，其中一些无法编译（T3 desc、M1 和 desc 以及 C1 alt 和 desc）。

Besides the downstream challenge that the end-user's system may not render LaTeX, many of the LaTeX-style reproductions were wrong. GPT-Vision sometimes omitted subscripts or misprinted them (which could be attributed to low resolution). It was particularly prone to error when a subscript was longer than one character. In LaTeX, a subscript starts with an underscore followed by the characters to be subscripted. A single character can be written alone, like "x_i = x_i," but multiple characters need to wrapped in curly braces. In all cases except one, GPT-Vision missed this distinction. For example, it reproduced "α_{xy}" as "a_xy," which would have compiled to "a_xy" (C1 alt). These errors were internally consistent, so GPT-Vision referred to the same values in the same way within each passage.
除了最终用户的系统可能无法呈现 LaTeX 这一下游挑战之外，许多 LaTeX 风格的复制品都是错误的。GPT-Vision 有时会省略下标或错误打印下标（这可能归因于低分辨率）。当下标长于一个字符时，它特别容易出错。在 LaTeX 中，下标以下划线开头，后跟要加下标的字符。单个字符可以单独书写，如 "x_i = x_i"，但多个字符需要用花括号括起来。除了一种情况外，GPT-Vision 在所有情况下都忽略了这一区别。例如，它

将 “α_{xy}” 复制为 “a_xy”，而这将编译为 “a_xy” (C1 alt)。这些错误在内部是一致的，因此 GPT-Vision 在每个段落中以相同的方式引用相同的值。

The one multi-character subscript GPT-Vision reproduced correctly was from this equation:
GPT-Vision 正确再现的一个多字符下标来自以下等式：

$$p(y_i|y1, ..., y_{i-1}, \mathrm{x}) = g(y_{i-1}, s_i, c_i),$$

which it wrote as

p(y_i|y_1, ..., y_{i-1}, x) = g(y_{i-1}, s_i, c_i) (M1 desc)

Given the frequency of its LaTeX errors and the ubiquity of this expression as a conditional probability for a recurrent neural network, we should avoid assuming that GPT-Vision “understood” how to print it correctly. Assessing the underlying abilities of closed-source models is a common challenge since we cannot verify the training data, but our behavioral analysis seems to suggest that this success is coincidental.
考虑到其 LaTeX 错误的频率以及这种表达式作为循环神经网络条件概率的普遍性，我们应避免假设 GPT-Vision “理解” 如何正确打印它。评估闭源模型的底层能力是一项常见的挑战，因为我们无法验证训练数据，但我们的行为分析似乎表明这种成功是巧合。

Powerful generative AI models like GPT-Vision could go even further by describing mathematical text in natural language instead of solely reproducing it as is. GPT-Vision already showed good performance when describing a pseudocode algorithm (see Section 3.7). With adjustments, GPT-Vision has great potential to advance learning, accessibility, and inclusion.
像 GPT-Vision 这样的强大生成式 AI 模型可以更进一步，用自然语言描述数学文本，而不是仅仅按原样再现它。GPT-Vision 在描述伪代码算法时已经表现出良好的性能（参见 3.7 节）。经过调整后，GPT-Vision 具有巨大的潜力来促进学习、可访问性和包容性。

3.13 Counting Errors

Counting objects was another frequent source of error, with GPT-Vision miscounting on 10 of 21 images. For example, T1 shows a table of participants from a cooking study. One column, labeled “Self-Rated Skill,” lists how the participants rated their cooking abilities on a five-point Likert scale. Instead of listing the number, the table presents the skill levels with a sequence of five boxes. The number of large “filled” boxes represents the participant’s skill level out of five.
计数物体是另一个常见的错误来源，GPT-Vision 在 21 张图像中有 10 张计数错误。例如，T1 显示了一张烹饪研究的参与者表格。其中一列标记为 “自我评价技能”，列出了参与者如何根据五点李克特量表评价他们的烹饪能力。表格没有列出数字，而是用五个方框的序列来显示技能水平。大 “填充” 框的数量代表参与者的五分制技能水平。

For example, this reproduction ■■■▪▪ shows three large boxes followed by two small boxes, so it represents a skill level of three out of five. GPT-Vision counted ten out of twelve of these boxes incorrectly when responding to the “desc” prompt (it did not count at all for the “alt” prompt) (see Table 16).

Original	#	GPT-V
■■■■▪	4	3
■■■■▪	4	3
■■▪▪▪	2	4
■■■▪▪	3	4
■▪▪▪▪	1	4
■■▪▪▪	2	3
■■■■▪	4	3
■■▪▪▪	2	5
■■■■■	5	3
■■■■▪	4	4
■■▪▪▪	2	5
■■■▪▪	3	3

Figure 16: The skill levels from T1 (Original) with their true counts (#) and GPT-Vision’s interpretation (Hwang et al., 2023).

例如，此复刻版 ■■■▪▪ 显示三个大框后面跟着两个小框，因此它代表五分之三的技能水平。GPT-Vision 在响应“desc”提示时错误地计算了十二个框中的十个（对于“alt”提示，它根本没有计算）（参见表格 16）。

The responsibility for LLMs to handle numbers well is unclear. Some have argued that LLMs should be given a calculator rather than be trained to calculate (Andor et al., 2019). The mechanism for numerical reasoning may vary, but number sense will remain an important capability for describing images.
法学硕士是否应该很好地处理数字还不清楚。有人认为，应该给法学硕士配备一个计算器，而不是训练他们计算 (Andor et al., 2019)。数字推理的机制可能有所不同，但数字感仍将是描述图像的重要能力。

P1 alt and desc, G1 alt, G2 alt and desc, G3 desc, T1 desc, T2 desc, T3 alt and desc, F2 desc, F4 desc, and C2 desc contain counting errors.

3.14 (Lack of) Logo Recognition

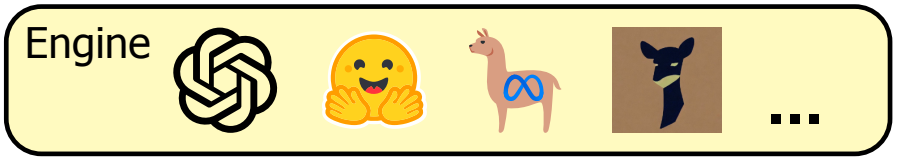


Figure 17: The Engine section of D1 (Zhu et al., 2023).

GPT-Vision did not recognize the three logos displayed in the “Engine” section at the bottom of D1. These logos were supposed to represent the language models that Kani supports:
GPT-Vision 无法识别 D1 底部“引擎”部分显示的三个徽标。这些徽标应该代表 Kani 支持的语言模型：

1. OpenAI (a circular logo resembling three intertwined chain links) (OpenAI, 2023b),
2. Hugging Face (a yellow emoji-like happy face with open hands) (Hugging Face, 2023),
3. LLaMA (not an official logo, a brown cartoon llama with the Meta logo) (Touvron et al., 2023),
4. Vicuna (head of a cartoon vicuna, which has a tall neck and pointy ears) (The Vicuna Team, 2023).

which GPT-Vision described in the desc passage as
GPT-Vision 在描述段落中将其描述为

1. “a caduceus [two serpents twisted around a staff] with only one snake” (Wikipedia, 2023),
2. “a yellow smiley face,”
3. “a flamingo,” and
4. “a letter ‘Y’ with what looks like animal ears on top.”

Besides mistaking the llama for a flamingo, GPT-Vision’s descriptions of the engine icons are not far off, but not recognizing the logos themselves obscured the point of this part of the diagram.
除了将骆驼误认为火烈鸟之外，GPT-Vision 对引擎图标的描述也差不多，但无法识别徽标本身掩盖了该部分图表的重点。

Original Colors		GPT-Vision's Interpreted Colors	
Color	Category	Color	Category
green	grammar	blue	grammar errors
orange	repetition	green	repetitions
blue	irrelevant	purple	irrelevance
pink	contradicts_sentence	yellow	contradictions with sentence context or knowledge
light green	contradicts_knowledge		
yellow	common_sense	light purple	commonsense and coherence errors
tan	coreference	orange	coreference errors
gray	generic	red	generic
green	other	gray	other errors

Table 2: The legend from G2 with GPT-Vision's interpretation (desc) (Dugan et al., 2023).

3.15 Color Blindness

As mentioned in OpenAI (2023a), GPT-Vision consistently failed to recognize colors. This was especially apparent when it described G2, a plot of pie charts with nine color-coded categories (see Table 2). The legend in G2 shows a vertical list of categories with their associated colors:
如 OpenAI (2023a) 中所述，GPT-Vision 始终无法识别颜色。这一点在描述 G2 时尤其明显，G2 是一个包含九个颜色编码类别的饼图（见表 2）。G2 中的图例显示了类别及其相关颜色的垂直列表：

> (1) "grammar" (green), (2) "repetition" (orange), (3) "irrelevant" (blue), (4) "contradicts_sentence" (pink), (5) "contradicts_knowledge" (light green), (6) "common_sense" (yellow), (7) "coreference" (tan), (8) "generic" (gray), and (9) "other" (green, repeated),

but GPT-Vision reported them slightly differently (emphasis ours).
但 GPT-Vision 的报告略有不同（emphasis 我们的）。

> (1) "grammar errors," (2) "repetitions," (3) "irrelevance," (4+5) "contradictions with sentence context or knowledge," (6+) "commonsense and coherence errors," (7) "coreference errors," and (8) "other errors." (G2 desc)

GPT-Vision also mislabeled most of the colors. It mistook green for blue, orange for green, blue for purple, yellow for light purple, tan for orange, gray for red, and green for gray (G2 desc). It seems to have merged "contradicts_sentence" (pink) and "contradicts_knowledge" (light green) into one category of the color yellow. GPT-Vision displayed similar behavior with the legend in G1 as well, suggesting that color recognition is a serious weakness. It may define colors differently than we expect or suffer from one-off errors from misaligning the color swatches with their labels. Further investigation into the source of this behavior may help us fix it.
GPT-Vision 还错误地标记了大多数颜色。它把绿色误认为蓝色，把橙色误认为绿色，把蓝色误认为紫色，把黄色误认为浅紫色，把棕褐色误认为橙色，把灰色误认为红色，把绿色误认为灰色（G2 desc）。它似乎将"矛盾句子"（粉红色）和"矛盾知识"（浅绿色）合并为黄色的一个类别。GPT-Vision 也表现出与 G1 中的图例类似的行为，这表明颜色识别是一个严重的弱点。它可能对颜色的定义与我们预期的不同，或者由于颜色样本与其标签未对齐而导致一次性错误。进一步调查此行为的根源可能会帮助我们修复它。

3.16 Quality of Alt Text

Length Most of the alt text generated by GPT-Vision was about a paragraph in length, the exception being alt text for full pages that was typically much longer. This clashes with standard guidelines for alt text,

which recommend a brief sentence because screen readers may impose character limits (Eggert et al., 2022; VLE Guru, 2022). One work in generating image descriptions for accessibility, however, found that blind/low-vision participants actually preferred longer descriptions (while sighted participants showed no clear pattern), in contrast with typical guidelines (Kreiss et al., 2022). With the right adjustments, GPT-Vision's ability to generate long text can lead to detailed, valuable image descriptions.
GPT-Vision 生成的大多数替代文本长度约为一段，但整页的替代文本除外，其长度通常要长得多。这与替代文本的标准指南相冲突，后者建议使用简短的句子，因为屏幕阅读器可能会施加字符限制 (Eggert et al., 2022; VLE Guru, 2022)。然而，一项为无障碍生成图像描述的研究发现，盲人/视力低下的参与者实际上更喜欢较长的描述（而视力正常的参与者没有表现出明显的模式），这与典型的指南 (Kreiss et al., 2022) 形成了鲜明对比。通过适当的调整，GPT-Vision 生成长文本的能力可以产生详细、有价值的图像描述。

Audience, content, purpose Good alt text depends on the audience, content, and purpose of the image, so one alt text cannot necessarily fit all situations for the same image (VLE Guru, 2022). Our analysis found that GPT-Vision tended to focus too much on visual details and too little on the main ideas. For example, when describing a full-page screenshot of Hwang et al. (2023), GPT-Vision wrote,
好的替代文本取决于图像的受众、内容和用途，因此一个替代文本不一定适合同一图像的所有情况 (VLE Guru, 2022)。我们的分析发现，GPT-Vision 倾向于过多关注视觉细节，而很少关注主要思想。例如，在描述 Hwang et al. (2023) 的整页截图时，GPT-Vision 写道，

> This is an image of a research paper page titled "Rewriting the Script: Adapting Text Instructions for Voice Interaction." The page contains a figure and two sections of text with bullet points. (F3 alt)

The details GPT-Vision chose to highlight misrepresent the likely audience and purpose of this image. The audience of such a paper is likely to be researchers in computer science or user experience design. The purpose of the page is to convey information about the study, namely the methods, technology probe, and participants. GPT-Vision instead surfaced the figure and "two sections of text" to the reader, giving them very little idea of the content itself. A more useful alt text may have been
GPT-Vision 选择强调的细节歪曲了这张图片的可能受众和目的。这类论文的受众可能是计算机科学或用户体验设计的研究人员。该页面的目的是传达有关研究的信息，即方法、技术探索和参与者。GPT-Vision 反而向读者展示了图表和 "两段文字"，让他们对内容本身知之甚少。更有用的替代文本可能是

> Page from a research paper titled "Rewriting the Script: Adapting Text Instructions for Voice Interaction" discussing part of section 3, "METHODS," with a figure of the "study setting" and a table of "participants."

Not all readers are the same, of course. Lundgard and Satyanarayan (2021) found a stark divide between blind/low-vision (BLV) and sighted readers for graphs: sighted readers appreciated a "story" about the data but BLV readers strongly disliked "subjective interpretations, contextual information, or editorializing." BLV readers wanted a more literal description of the graph so they could interpret the data for themselves.
当然，并非所有读者都是一样的。Lundgard and Satyanarayan (2021) 发现盲人/低视力 (BLV) 读者和视力正常的读者在图表方面存在明显差异：视力正常的读者欣赏有关数据的 "故事"，但 BLV 读者非常不喜欢 "主观解释、背景信息或社论"。BLV 读者希望对图表进行更为文字化的描述，以便他们可以自己解释数据。

For BLV readers, the emphasis on visual details that GPT-Vision tended to provide may be very useful if it selects the most important details to describe. Sighted readers will need a different kind of alt text while readers with non-visual disorders like issues with long- or short-term visual memory may need another set of standards altogether. GPT-Vision showed impressive performance on a diverse set of images with just two simple prompts. It shows great promise to generate high-quality alt text for more than just scientific images.
对于 BLV 阅读器来说，如果 GPT-Vision 选择最重要的细节进行描述，它倾向于强调视觉细节，这可能非

常有用。视力正常的读者需要不同类型的替代文本，而患有非视觉障碍（如长期或短期视觉记忆问题）的读者可能需要完全不同的一套标准。GPT-Vision 仅通过两个简单的提示就对一组不同的图像表现出色。它显示出为科学图像以外的领域生成高质量替代文本的巨大潜力。

4 Conclusion

In this paper we have presented a framework for a more rigorous and structured application of qualitative analysis to generative AI models. Our proposed framework not only alleviates the concerns of previous large-scale qualitative analysis work being "unscientific", but also allows us the opportunity to develop an alternative approach to evaluation separate from traditional benchmarks. Through our analysis we are able to identify a number of general trends in the capabilities of the newly-released GPT-Vision model such as its heavy reliance on textual information and its sensitivity to prompts. Such insights will no doubt be useful in future applications and can serve as guidelines for future areas of research.
在本文中，我们提出了一个框架，用于更严格、更结构化地将定性分析应用于生成式 AI 模型。我们提出的框架不仅缓解了人们对以前大规模定性分析工作"不科学"的担忧，还让我们有机会开发一种不同于传统基准的替代评估方法。通过我们的分析，我们能够识别出新发布的 GPT-Vision 模型功能中的一些总体趋势，例如它严重依赖文本信息和对提示的敏感性。这些见解无疑将在未来应用中发挥作用，并可作为未来研究领域的指导方针。

One important caveat is that, while our analysis offers key insight on GPT-Vision's behavior with scientific images, such insights should not be conflated with a statistical understanding of the relative frequency of these issues or a scientific explanation of why such issues occur. Even a suitable description of an image does not necessarily mean that GPT-Vision "explained the image" if the same information could have been hallucinated from its internal knowledge or training data. Much like in psychology, behavioral studies cannot fully supplant cognitive research or neuroscience. Further investigation on the "cognitive" processes of LLMs, like attention and memory, and the mathematical basis of neural networks is crucial for understanding LLMs holistically.
一个重要的警告是，虽然我们的分析提供了有关 GPT-Vision 处理科学图像的行为的关键见解，但这些见解不应与对这些问题的相对频率的统计理解或对这些问题发生原因的科学解释相混淆。即使是对图像的适当描述也不一定意味着 GPT-Vision "解释了图像"，如果相同的信息可以从其内部知识或训练数据中产生幻觉。就像心理学一样，行为研究不能完全取代认知研究或神经科学。进一步研究 LLM 的"认知"过程，如注意力和记忆力，以及神经网络的数学基础，对于全面理解 LLM 至关重要。

Acknowledgments

First and foremost, we would like to thank Liam Dugan for his tremendous support and feedback. We were also inspired by early talks with Jonathan Bragg and Doug Downey. We are also grateful for feedback from Harry Goldstein, Andrew Zhu, and Natalie Collina on GPT-Vision's descriptions of their work. Finally, we are thankful for the community at Penn NLP and Penn HCI that could make this work possible.

References

Daniel Andor, Luheng He, Kenton Lee, and Emily Pitler. 2019. Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension. ArXiv:1909.00109 [cs].

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv:1409.0473 [cs, stat].

Anja Belz and Ehud Reiter. 2006. Comparing Automatic and Human Evaluation of NLG Systems. In 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Andrea Bingham. 2023. Qualitative Analysis: Deductive and Inductive Approaches.

Ann Blandford, Dominic Furniss, and Stephann Makri. 2022a. Analysing Data. In Qualitative HCI Research: Going Behind the Scenes, 1 edition, Synthesis Lectures on Human-Centered Informatics, pages 51–60. Springer Cham. Citekey: thematic-analysis.

Ann Blandford, Dominic Furniss, and Stephann Makri. 2022b. Paradigms and Strategies. In Qualitative HCI Research: Going Behind the Scenes, 1 edition, Synthesis Lectures on Human-Centered Informatics, pages 61–78. Springer Cham. Citekey: grounded-theory.

Ann Blandford, Dominic Furniss, and Stephann Makri. 2022c. Sampling and Recruitment. In Qualitative HCI Research: Going Behind the Scenes, 1 edition, Synthesis Lectures on Human-Centered Informatics, pages 23–31. Springer Cham. Citekey: sampling.

Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using Thematic Analysis in Healthcare HCI at CHI: A Scoping Review. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv:2303.12712 [cs].

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating Question Answering Evaluation. In Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. 2022. A Dataset of Alt Texts from HCI Publications: Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations in Scientific Papers. The 24th International ACM SIGACCESS Conference on Computers and Accessibility, pages 1–12. Conference Name: ASSETS '22: The 24th International ACM SIGACCESS Conference on Computers and Accessibility ISBN: 9781450392587 Place: Athens Greece Publisher: ACM.

Natalie Collina, Nicole Immorlica, Kevin Leyton-Brown, Brendan Lucier, and Neil Newman. 2021. Dynamic Weighted Matching with Heterogeneous Arrival and Departure Rates. ArXiv:2012.00689 [cs].

Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Juliet M. Corbin and Anselm Strauss. 1990. Grounded Theory Research: Procedures, Canons, and Evaluative Criteria. Qualitative Sociology, 13(1):3–21.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. ArXiv:2310.15100 [cs].

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods. Transactions of the Association for Computational Linguistics, 9:1132–1146.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or Fake Text?: Investigating Human Ability to Detect Boundaries between Human-Written and Machine-Generated Text. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11):12763–12771. Number: 11 citekey: roft-analysis.

Eric Eggert, Shadi Abou-Zahra, and Brian Elton. 2022. Images Tutorial.

Marina Fomicheva and Lucia Specia. 2019. Taking MT Evaluation Metrics to Extremes: Beyond Correlation with Human Judgments. Computational Linguistics, 45(3):515–558.

Jie Gao, Yuchen Guo, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. CollabCoder: A GPT-Powered WorkFlow for Collaborative Qualitative Analysis. In Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing, CSCW '23 Companion, pages 354–357, New York, NY, USA. Association for Computing Machinery.

Robert P Gauthier, Mary Jean Costello, and James R Wallace. 2022. "I Will Not Drink With You Today" : A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, pages 1–17, New York, NY, USA. Association for Computing Machinery.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. Distill, 6(3).

Harrison Goldstein, Samantha Frohlich, Meng Wang, and Benjamin C. Pierce. 2023. Reflecting on Random Generation. Proceedings of the ACM on Programming Languages, 7(ICFP):200:322–200:355. Citekey: pl-reflecting-random.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News Summarization and Evaluation in the Era of GPT-3. ArXiv:2209.12356 [cs].

Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pages 1–11, New York, NY, USA. Association for Computing Machinery.

Ting-Yao Hsu, C. Lee Giles, and Ting-Hao 'Kenneth' Huang. 2021. SciCap: Generating Captions for Scientific Figures. ArXiv:2110.11624 [cs] version: 2.

Hugging Face. 2023. Hugging Face.

Alyssa Hwang. 2023. Part 3: Critiquing Our Design. In Build Your Own ChatGPT.

Alyssa Hwang, Natasha Oza, Chris Callison-Burch, and Andrew Head. 2023. Rewriting the Script: Adapting Text Instructions for Voice Interaction. In Proceedings of the 2023 ACM Designing Interactive Systems Conference, DIS '23, pages 2233–2248, New York, NY, USA. Association for Computing Machinery. Citekey: rewriting.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-Dimensional Evaluation of Text Summarization with In-Context Learning. ArXiv:2306.01200 [cs].

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4685–4697, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. ArXiv:2303.16634 [cs].

Alan Lundgard and Arvind Satyanarayan. 2021. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. ArXiv:2110.04406 [cs].

Karen McCall and Beverly Chagnon. 2022. Rethinking Alt Text to Improve Its Effectiveness. In Computers Helping People with Special Needs, Lecture Notes in Computer Science, pages 26–33, Cham. Springer International Publishing.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):72:1–72:23.

Graham Neubig. 2023. Is My NLP Model Working? The Answer is Harder Than You Think.

Donald A. Norman. 2013. The Design of Everyday Things, revised and expanded edition edition. Basic Books, New York, New York.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19–51.

OpenAI. 2023a. GPT-4V(ision) System Card. Citekey: gptvision.

OpenAI. 2023b. OpenAI.

Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 5:147–156.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. ARB: Advanced Reasoning Benchmark for Large Language Models. ArXiv:2307.13692 [cs].

Andrea Spreafico and Giuseppe Carenini. 2020. Neural Data-Driven Captioning of Time-Series Line Charts. In Proceedings of the International Conference on Advanced Visual Interfaces, pages 1–5, Salerno Italy. ACM.

Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '21, pages 1–15, New York, NY, USA. Association for Computing Machinery. Citekey: going-beyond-one-size.

Chase Stokes, Vidya Setlur, Bridget Cogley, Arvind Satyanarayan, and Marti Hearst. 2022. Striking a Balance: Reader Takeaways and Preferences when Integrating Text and Charts. IEEE Transactions on Visualization and Computer Graphics, pages 1–11. ArXiv:2208.01780 [cs].

Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. Citekey: vistext.

The Vicuna Team. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

VLE Guru. 2022. What is Alternative Text? How Do I Write It for Images, Charts, and Graphs? Citekey: alt-text-video.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. ArXiv:2110.07178 [cs].

Wikipedia. 2023. Caduceus. Page Version ID: 1178441181.

Candace Williams, Lilian de Greef, Ed Harris, Leah Findlater, Amy Pavel, and Cynthia Bennett. 2022. Toward Supporting Quality Alt Text in Computing Publications. In Proceedings of the 19th International Web for All Conference, W4A '22, pages 1–12, New York, NY, USA. Association for Computing Machinery. Citekey: quality-alt-text.

Jacob O. Wobbrock, Shaun K. Kane, Krzysztof Z. Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-Based Design: Concept, Principles and Examples. ACM Transactions on Accessible Computing, 3(3):1–27.

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, pages 1180–1192, New York, NY, USA. Association for Computing Machinery.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). ArXiv:2309.17421 [cs].

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. ArXiv:2304.06364 [cs].

Andrew Zhu, Liam Dugan, Alyssa Hwang, and Chris Callison-Burch. 2023. Kani: A Lightweight and Highly Hackable Framework for Building Language Model Applications. In Proceedings of the Third Workshop for NLP Oper Source Software (NLP-OSS), Singapore. Association for Computational Linguistics. Citekey: kani.

Type	ID	Description
Photo	P1	A 2x6 photo collage of various dishes labeled with their names prepared in (Hwang et al., 2023).
	P1.1	The same image as P1 but with modified labels (Hwang et al., 2023).
	P2	A study participant cooking in a kitchen with an Alexa Echo Dot, which is circled on the right (Hwang et al., 2023).
Diagram	D1	An illustration of Kani, a framework for building applications with large language models, from the first page of (Zhu et al., 2023)
	D2	The transformation of a written recipe for audio delivery by editing the original text (Hwang et al., 2023).
	D3	A complex, abstract representation of symbolic knowledge distillation with emojis, arrows, text labels, and other visual details (West et al., 2022).
Graph	G1	A plot of three bar graphs, each displaying two groups of four differently colored columns with error bars (Dugan et al., 2023).
	G2	A 3x3 plot of 9 pie charts representing 9 color-coded categories (Dugan et al., 2023).
	G3	A line graph displaying three color-coded data lines and two dashed horizontal benchmark lines (Brown et al., 2020).
	G4	A line graph similar to G3, but with three pairs lines (dashed and solid) and additional text labels on the plot (Brown et al., 2020).
Table	T1	A 13x4 table (including header) containing text and sequences of boxes graphically representing Likert scales (Hwang et al., 2023).
	T1.1	The same table as T1, but with the caption included beneath it.
	T2	A 6x10 (including 2 rows for the header) table of performance metrics; some columns are merged (Och and Ney, 2003, Table 18).
	T3	A 16x6 table (including 2 rows for the header) of model training schemes for varying corpus sizes; some rows are merged (Och and Ney, 2003, Table 4).

Table 3: Images of figures used in our analysis.

Type	ID	Description
Full page	F1	The first page of a research publication with the title, authors, two columns of text, and metadata (Hwang et al., 2023).
	F2	The full page of a research publication that includes P1 and its caption spanning the top half followed by two columns of text (Hwang et al., 2023).
	F3	The full page of a research publication displaying P2 in the top of the left column of text and T1 in the top of the right (Hwang et al., 2023).
	F4	A full page of a research publication with a large, text-based table covering the top two-thirds and some text in two columns beneath it (Hwang et al., 2023).
	F5	A full page of a research publication in one-column format beginning with two brief side-by-side snippets of Haskell code (Goldstein et al., 2023).
Code	C1	A brief pseudocode algorithm featuring a nested for loop, an if statement, and some mathematical text (Collina et al., 2021).
	C2	A page-long proof from a research publication on programming languages that includes Haskell code (Goldstein et al., 2023).
	C3	A page-long excerpt of Python code defining a class and a few instance methods for a chat-based application (Zhu et al., 2023).
Math	M1	An excerpt from a machine learning research publication introducing a new model architecture with mathematical equations (Bahdanau et al., 2016).
	M2	The definition of an algorithmic theorem followed by its proof, featuring bullet points and mathematical representations of abstract concepts (Collina et al., 2021).

Table 4: Images of full pages and special text (code and math) used in our analysis.