

## Trabalho Prático – Época Especial

# Sumário



1. Introdução e Objetivos
2. Ferramentas Usadas
3. Preparação e Execução do Benchmark LogicVista
4. Avaliação de Resultados com o GPT-4o
5. Testes com Imagens Externas
6. AiVista
7. Conclusão

# 1 - Introdução e Objetivos



O projeto apresentado foca-se na análise do artigo **LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts**, que propõe um novo benchmark para avaliar a capacidade de raciocínio lógico de modelos multimodais em contextos visuais. Esta abordagem combina linguagem natural e imagens, com o objetivo de testar a verdadeira compreensão e inferência lógica de LLMs.

Este trabalho tem como objetivo executar e testar o benchmark **LogicVista**, aplicando-o a diferentes conjuntos de imagens, com o intuito de avaliar a capacidade dos modelos multimodais em realizar raciocínios lógicos e generalizar para além dos dados originais.

Repositório GitHub: [MLLM Final Project](#)

## 2 - Ferramentas Usadas



- **LogicVista:**

LogicVista é uma ferramenta de avaliação desenvolvida para medir a capacidade de raciocínio lógico de Modelos Multimodais de Linguagem (MLLMs) em contextos visuais. Apresenta um benchmark abrangente composto por 448 questões de escolha múltipla, organizadas em cinco categorias de raciocínio — indutivo, dedutivo, numérico, espacial e mecânico — e nove tipos distintos de capacidades multimodais, como diagramas, OCR, tabelas e sequências. Cada questão é acompanhada de justificações escritas por humanos, permitindo avaliações abertas ou estruturadas. A ferramenta inclui ainda um sistema automatizado de extração e verificação de respostas via LLM, promovendo uma análise estatística rigorosa da performance dos modelos em tarefas cognitivamente exigentes.

### 3 - Preparação e Execução do Benchmark LogicVista



- **Clonagem do Repositório:** O código-fonte foi obtido a partir do repositório oficial disponível no GitHub: [Yijia-Xiao/LogicVista](https://github.com/Yijia-Xiao/LogicVista). O repositório inclui scripts prontos para avaliação, dataset de respostas detalhado em JSON e exemplos de imagens.
- **Ambiente e Requisitos:** Configuração do ambiente local com os requisitos especificados (langchain, langchain\_community, openai, numpy, tqdm).
- **Modelos Pré-configurados:** O repositório já inclui suporte para diversos MLLMs, como LLaVA, MiniGPT-4 e Otter, que podem ser utilizados diretamente para testes com o dataset fornecido.
- **Dataset LogicVista:** O dataset de avaliação contém 447 questões de escolha múltipla, com imagens e anotações humanas. As imagens estão organizadas na pasta data, e os metadados correspondentes incluem perguntas, alternativas, respostas corretas e justificações.

## 4 - Avaliação de Resultados do GPT-4o e comparação com o modelo LLAVA7B e GPT4

- **Integração com GPT-4o:** O modelo GPT-4o foi utilizado externamente ao pipeline original do LogicVista, uma vez que o repositório não inclui suporte nativo para este modelo. As imagens do dataset foram processadas por script adaptado.
- **Formato de Resposta:** As perguntas foram apresentadas ao GPT-4o com instruções consistentes com o estilo do benchmark ("Escolhe entre A, B, C, D ou E"), e a resposta foi registada para comparação.
- **Comparação com respostas:** As respostas geradas foram comparadas com as respostas corretas presentes no JSON do dataset. A verificação de acerto foi binária (certo/errado).
- **Resultados Observados:** O GPT-4o apresentou desempenho superior aos modelos GPT-4 e LLAVA7B no total de tarefas resolvidas corretamente (46.99%), destacando-se nas categorias dedutiva (72.04%), mecânica (64.86%) e numérica (52.61%). Em comparação, o GPT-4 obteve 31.91% no total, e o LLAVA7B apenas 18.48%. O GPT-4o também demonstrou melhor capacidade nas tarefas com imagens de OCR (72.8%) e diagramas (60.06%), indicando maior robustez tanto no raciocínio como na interpretação multimodal.

```
Model: gpt-4o
Reasoning Skill Acc:
Total: 46.99331848552338% (211/449)
Inductive: 27.7777777777778% (30/108)
Deductive: 72.04301075268818% (67/93)
Numerical: 52.63157894736842% (50/95)
Spatial: 20.253164556962027% (16/79)
Mechanical: 64.86486486486487% (48/74)
Capability Acc:
diagram: 36.708860759493675% (87/237)
ocr: 72.8% (91/125)
```

```
Model: GPT4
Reasoning Skill Acc:
Total: 31.919642857142854% (143/448)
Inductive: 20.5607476635514% (22/107)
Deductive: 53.76344086021505% (50/93)
Numerical: 24.210526315789473% (23/95)
Spatial: 21.518987341772153% (17/79)
Mechanical: 41.891891891891895% (31/74)
Capability Acc:
diagram: 26.006191950464398% (84/323)
ocr: 37.735849056603776% (80/212)
```

```
Model: LLAVA7B
Reasoning Skill Acc:
Total: 18.485523385300667% (83/449)
Inductive: 29.629629629629626% (32/108)
Deductive: 18.27956989247312% (17/93)
Numerical: 7.368421052631578% (7/95)
Spatial: 17.72151898734177% (14/79)
Mechanical: 17.56756756756757% (13/74)
Capability Acc:
diagram: 22.362869198312236% (53/237)
ocr: 16.8% (21/125)
```

## 5 - Testes com Imagens Externas



Instituto Politécnico  
de Viana do Castelo

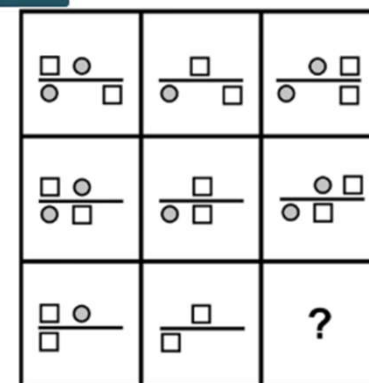
- **Procedimento:** Foi utilizado um exercício de raciocínio lógico visual retirado do Reddit, semelhante em estrutura aos exemplos do benchmark.
- **Execução:** A imagem foi submetida ao GPT-4o através de um script personalizado (run\_gpt4v\_on\_image.py). O modelo analisou o padrão visual e forneceu uma explicação textual do raciocínio, seguida da escolha da resposta correta.
- **Resultado Obtido:** O GPT-4o identificou corretamente a lógica de troca entre quadrados e círculos, selecionando a opção E, que corresponde à resposta correta. O modelo justificou a sua escolha com uma explicação coerente com o padrão visual.
- **Interpretação:** O desempenho positivo neste exemplo sugere que o GPT-4o consegue aplicar raciocínio lógico a imagens fora do conjunto original, mesmo sem ajustes específicos no prompt ou no formato dos dados.

```
PS C:\EI\IS\MLLM_Final_Project\LogicaVista\eval> python run_gpt4v_on_image.py
GPT-4 Vision Answer: The pattern involves arranging circles and squares above and below the line. In each row, the arrangement shifts consistently. The missing piece should follow this logic, ensuring the squares switch places while circles stay in their row's pattern.
```

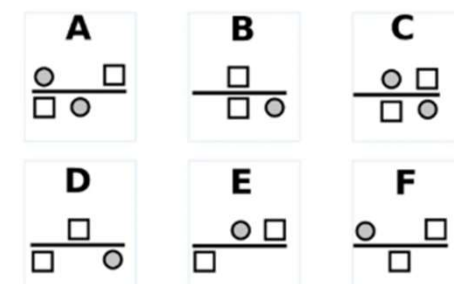
```
The answer is: **E**
```

19:44

Exercise 22



Select answer



## 5 - Testes com Imagens Externas



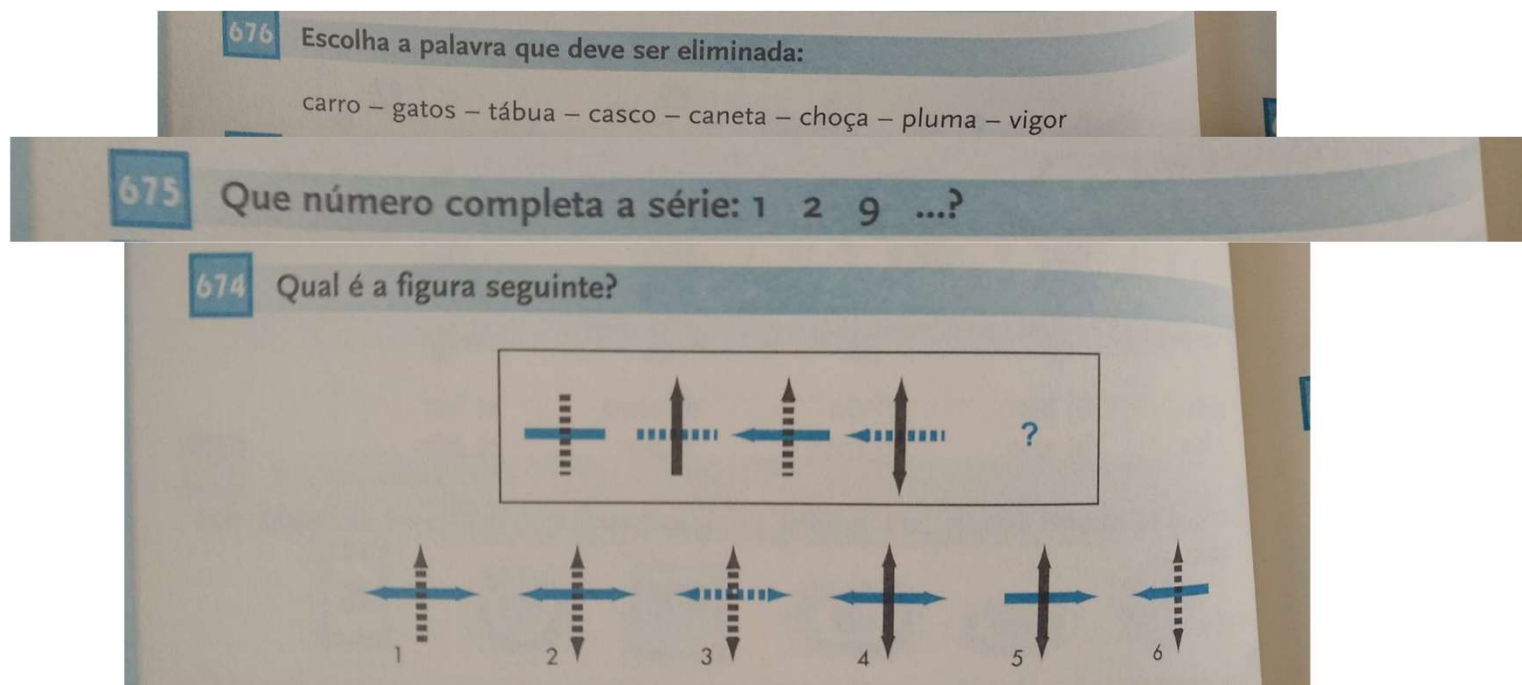
- **Procedimento:** Foi criado um novo conjunto de desafios visuais em formato JSON (new-challenges-dataset-(in)correct-base.json), contendo as imagens mostradas anteriormente e respetivas instruções de pergunta. Estas imagens foram guardadas na pasta data/new-challenges-images.
- **Execução:** O script `generate_gpt4o_new_challenges_answers.py` foi modificado para processar este novo dataset. A principal alteração consistiu na leitura do novo ficheiro `new-challenges-dataset-(in)correct-base.json` e na análise das imagens correspondentes utilizando o modelo GPT-4o via API da OpenAI. O script carrega cada imagem, codifica-a em base64, e envia um prompt combinado (imagem + questão) ao modelo. O GPT-4o responde com um raciocínio resumido seguido da escolha da opção correta.
- **Resultado Obtido:** As respostas foram automaticamente guardadas no ficheiro `gpt-4o-new-challenges-answers.json`, dentro da pasta `eval/answers`. Cada entrada inclui o identificador da imagem e a resposta textual gerada pelo modelo.
- **Interpretação:** A análise preliminar mostrou que o GPT-4o falhou nos cinco exemplos testados, apesar de apresentar justificações coerentes. Os erros envolveram interpretação visual incorreta, falhas na ordenação alfabética e categorização semântica, sugerindo limitações na generalização.



## 5 - Testes com Imagens Externas



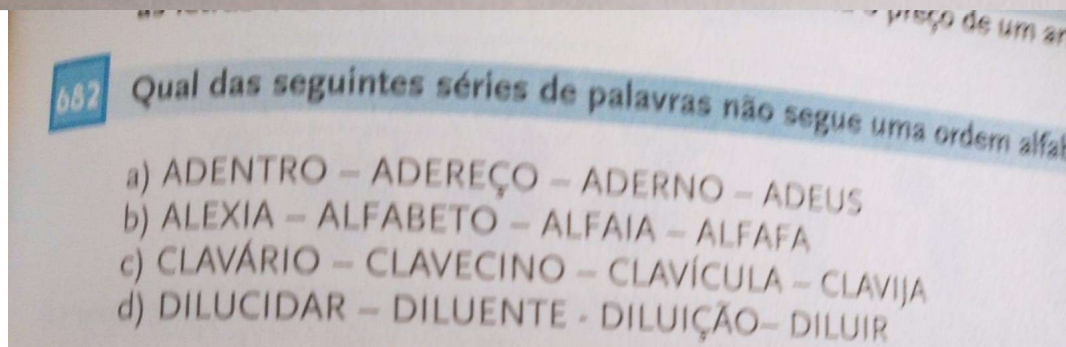
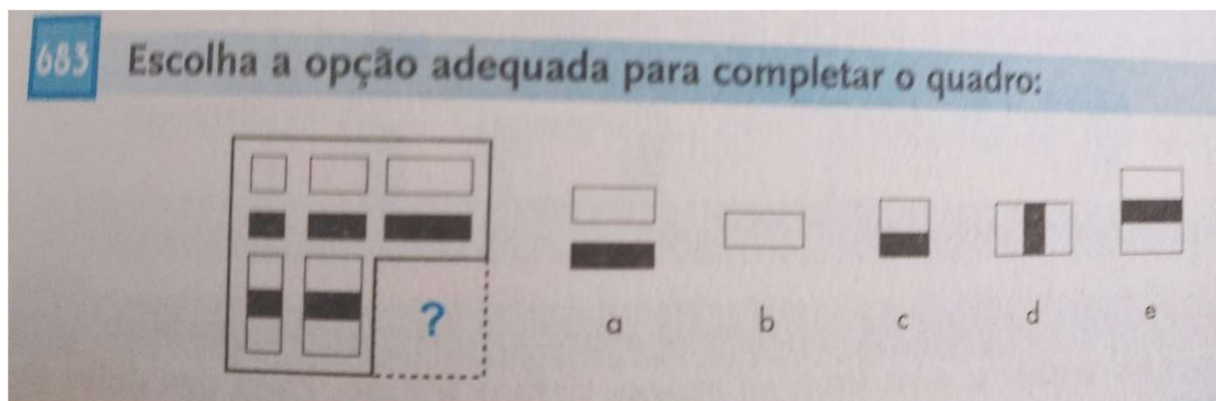
- Desafios:



## 5 - Testes com Imagens Externas



- Desafios:



## 5 - Testes com Imagens Externas



- **Respostas do modelo com respostas corretas no dataset:**

**v0\_676:** A palavra \"gatos\" deve ser eliminada, pois todas as outras palavras são objetos inanimados, enquanto \"gatos\" refere-se a seres vivos.\n\n Letra: gatos.

**v0\_675:** Para completar a série 1, 2, 9, a lógica parece ser multiplicar por números consecutivos: 1 (multiplicado por 1), 2 (multiplicado por 2), e depois 9 (multiplicado por 3), sugerindo que o próximo número é 36 (9 multiplicado por 4).\n\n D.

**v0\_674:** A sequência alterna entre setas vermelhas horizontais e pretas verticais, enquanto o padrão das barras internas também alterna. O próximo na sequência deve ser uma seta preta com linhas pontilhadas.\n Resposta: 4

**v0\_683:** Para completar o padrão, a coluna segue a lógica de alternar entre duas e três linhas preenchidas. A primeira coluna tem três partes brancas e duas pretas; a segunda tem três partes pretas; e a última deve seguir a sequência tendo três partes brancas. Portanto, a opção adequada é a mesma que a primeira coluna invertida horizontalmente.\n\n Resposta: \*\*b\*\*

**v0\_682:** A série de palavras na opção \*\*d\*\* não segue uma ordem alfabética, pois \"DILUIÇÃO\" deveria vir após \"DILUIR\". \n\n d.

## 5 - Testes com Imagens Externas

- **Procedimento:** As novas imagens e respetivos metadados foram adicionados ao dataset.json, juntando-as aos desafios originais. Esta combinação permitiu uma avaliação conjunta da performance do modelo GPT-4o.
- **Execução:** Após gerar as respostas através do script generate\_gpt4o\_answers.py, foi utilizado o extract\_accuracy.py para calcular a percentagem de acertos por tipo de raciocínio e capacidade. Em seguida, o script analyze\_accuracy.py foi executado para gerar os resultados agregados. A imagem 1 mostra os resultados com o conjunto expandido, enquanto a imagem 2 apresenta os resultados apenas com os dados originais.
- **Resultado Obtido:** A inclusão dos novos desafios causou uma ligeira descida global na precisão do modelo (de 46.99% para 46.69%), com impacto mais acentuado nas categorias “Inductive” e “Diagram”. O número total de exemplos subiu de 449 para 454, evidenciando que os novos exemplos foram corretamente incorporados.
- **Interpretação:** A redução na performance sugere que os novos exercícios introduzem uma complexidade adicional, à qual o GPT-4o ainda não responde de forma robusta. Em particular, os desafios visuais e de inferência indutiva continuam a representar um ponto fraco do modelo.

1

```
Model: gpt-4o
Reasoning Skill Acc:
Total: 46.69603524229075% (212/454)
Inductive: 28.18181818181818% (31/110)
Deductive: 71.27659574468085% (67/94)
Numerical: 52.63157894736842% (50/95)
Spatial: 20.253164556962027% (16/79)
Mechanical: 64.86486486486487% (48/74)
Capability Acc:
diagram: 36.82008368200837% (88/239)
ocr: 72.8% (91/125)
```

2

```
Model: gpt-4o
Reasoning Skill Acc:
Total: 46.99331848552338% (211/449)
Inductive: 27.77777777777778% (30/108)
Deductive: 72.04301075268818% (67/93)
Numerical: 52.63157894736842% (50/95)
Spatial: 20.253164556962027% (16/79)
Mechanical: 64.86486486486487% (48/74)
Capability Acc:
diagram: 36.708860759493675% (87/237)
ocr: 72.8% (91/125)
```

## 5 - Testes com Imagens Externas



- **Respostas do modelo com respostas incorretas no dataset:**

**v0\_676:** As palavras listadas são substantivos, exceto "vigor", que é um substantivo que implica uma condição ou estado. As demais são objetos ou animais concretos.  
Letra: vigor

**v0\_675:** A sequência é formada por números ao quadrado:  $1^2 = 1$ ,  $2^2 = 4$ ,  $3^2 = 9$ . O próximo termo seria  $4^2 = 16$ .  
D.

**v0\_674:** A sequência alterna entre setas verticais e horizontais, mantendo o padrão de cor e listras. A próxima figura segue contendo uma seta horizontal azul com listras pretas.  
5

**v0\_683:** A lógica do quadro mostra que cada linha repete a mesma distribuição de preenchimento nas colunas horizontais. Na terceira linha, o padrão de preenchimento vertical deve ser proporcional ao da segunda linha. Assim, a opção correta deve ter uma barra preta grossa no meio.  
A resposta correta é a letra **e**.

**v0\_682:** A série que não segue a ordem alfabética é a letra a) porque "ADENTRO" deveria vir depois de "ADEREÇO", seguindo a ordem alfabética correta.  
a)

## 5 - Testes com Imagens Externas

- **Procedimento:** Após integrar as novas questões no dataset principal, realizaram-se dois testes distintos com o modelo GPT-4o. Na imagem 1, o conjunto utilizou as respostas corretas para todos os exemplos. Na imagem 2, as respostas dessas novas questões foram intencionalmente adulteradas com os erros cometidos previamente pelo próprio modelo, de modo a verificar se este corrigia os inputs incorretos com base na análise da imagem.
- **Resultados Obtidos:** Com as respostas corretas, o GPT-4o obteve 212 acertos (46.69%), com 28.18% em raciocínio indutivo e 36.82% em diagramas. Com as respostas incorretas, alcançou 213 acertos (46.91%), com 29.09% em indutivo e 37.23% em diagramas.
- **Interpretação:** O aumento, ainda que pequeno, do número total de acertos após a introdução de respostas deliberadamente erradas no dataset sugere que o GPT-4o foi capaz de corrigir parcialmente os inputs manipulados, indicando alguma autonomia na interpretação visual. Esta resistência a ruído nos dados reforça a hipótese de que o modelo recorre ao conteúdo da imagem e não apenas ao prompt textual para formular as respostas. No entanto, os ganhos são marginais, o que indica que essa capacidade ainda é limitada e dependente da natureza do desafio.

1

```
Model: gpt-4o
Reasoning Skill Acc:
Total: 46.69603524229075% (212/454)
Inductive: 28.18181818181818% (31/110)
Deductive: 71.27659574468085% (67/94)
Numerical: 52.63157894736842% (50/95)
Spatial: 20.253164556962027% (16/79)
Mechanical: 64.86486486486487% (48/74)
Capability Acc:
diagram: 36.82008368200837% (88/239)
ocr: 72.8% (91/125)
```

2

```
Model: gpt-4o
Reasoning Skill Acc:
Total: 46.91629955947136% (213/454)
Inductive: 29.09090909090909% (32/110)
Deductive: 71.27659574468085% (67/94)
Numerical: 52.63157894736842% (50/95)
Spatial: 20.253164556962027% (16/79)
Mechanical: 64.86486486486487% (48/74)
Capability Acc:
diagram: 37.238493723849366% (89/239)
ocr: 72.8% (91/125)
```



## 6 – Interface inicial

- Nesta interface o utilizador deve **definir as suas chaves de API** da [OpenAI](#) e da [Google AI Studio](#) para poder fazer chamadas aos modelos atualmente implementados, **GPT-4o** e **Gemini-1.5-flash**, e carregar em **Save** para que as seguintes persistam.
- Pode também fazer um pedido a ambos modelos **individualmente** ou **em conjunto** de modo a **comparar os resultados** facilmente. Para isso basta selecionar o modelo com qual pretendemos comparar (é possível criar lógica parecida à dos modelos usados para poder comparar com outros modelos).

**Add Entry** **Evaluate**

Qual a palavra que deve ser eliminada?

676.png

Choose File

Model API Keys

GPT-4o  
key

Gemini  
key

Save Clear Submit

A palavra que deve ser eliminada é **\*\*gatos\*\***, pois todas as outras são substantivos do gênero masculino.

**Add Entry** **Evaluate**

Qual a palavra que deve ser eliminada?

GPT-4o

Compare with:  
Gemini

Model API Keys

GPT-4o  
key

Gemini  
key

Save Clear Submit

**gpt-4o**  
A palavra que deve ser eliminada é "vigor". As outras são substantivos concretos, enquanto "vigor" é um substantivo abstrato.

**gemini**  
A palavra que deve ser eliminada é **\*\*vigor\*\***. As outras palavras (carro, gatos, tábuas, cascos, canetas, choças, pluma) referem-se a objetos ou coisas que podem ser vistos ou tocados. Vigor é um conceito abstrato, relacionado à força e energia.


## 6 – Adicionar imagem no dataset para avaliação

- Nesta interface o utilizador pode **carregar uma imagem** no conjunto já disponibilizado pelo **LogicVista**, para ser usada na avaliação do modelo. Este passo cria uma entrada em **data/images/** e em **data/dataset.json** com as opções definidas pelo utilizador.

**Dataset Entry**

Fill in the information to add a new dataset entry.

**Image**



Choose File

674.png

**Id**

674

**Question**

Qual é a figura seguinte?

**Answer**

4

**Reasoning**

Há dois padrões alternados:  
Um com a seta preta na vertical e a azul na horizontal, e  
Outro com a seta azul na vertical e a preta na horizontal.  
Além disso, as setas mudam de direção a cada passo (ex.: horizontal azul: →, ←, →...).  
A figura final segue o padrão alternado e direcional. A resposta correta é a figura 4, pois mantém a seta azul na horizontal (→) e a seta preta na vertical (↑), invertendo ambas em relação à terceira figura.

**Sourcelink**

**Image Source**

manual

**Licensed**

false

**Skill**

inductive

**Broad Capability**

diagram

**Specific Capability**

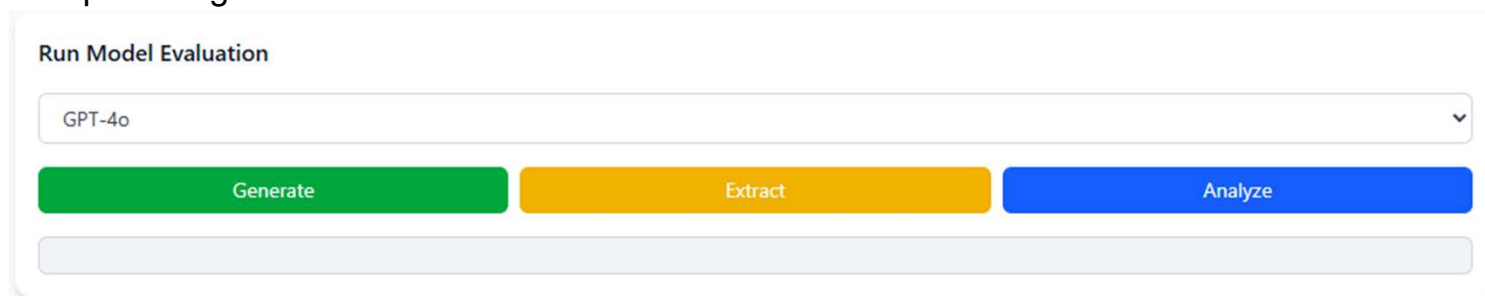
pattern

Cancel Submit



## 6 – Pagina de Avaliação

- Nesta interface o utilizador pode **gerar as respostas** do modelo seleccionado. Pode **extrair a precisão das respostas** para o ficheiro **csv** correspondente e **analisar e comparar** ambos os modelos.
- Para gerar as respostas, analisamos o dataset e comparamos para obter o numero de respostas que falta. De seguida as que são identificadas são respondidas pelos respetivos modelos e inseridas em **eval/answers/<model>-answers.json**
- Para extrair a sua precisão extraímos a última linha da resposta gerada e identificamos letras correspondentes às opções (A–E). De seguida comparamos com a resposta correta do dataset e os resultados são guardados em **eval/results/<modelo>-accuracy-results.csv**.
- Para analisar o modelo, calculamos a precisão total do modelo com base no ficheiro de resultados, agrupamos o desempenho por tipo de raciocínio e capacidade avaliada e a saída é um resumo em JSON com número de acertos por categoria.



Run Model Evaluation

GPT-4o

Generate Extract Analyze

## 6 – Pagina de Avaliação

- Continuamos o teste com a imagem inserida no slide anterior. Geramos as respostas de ambos os modelos (encontrou a nossa entrada – ‘Found 1 unanswered entries’):

Run Model Evaluation

GPT-4o

Generate

[GENERATE]:  
Using model: gpt-4o  
Found 1 unanswered entries.  
Done.

Run Model Evaluation

Gemini

Generate

[GENERATE]:  
Using model: gemini-1.5-flash  
Found 1 unanswered entries.  
Done.

- De seguida extraímos a precisão:

Run Model Evaluation

GPT-4o

Generate Extract

[GENERATE]:  
Using model: gpt-4o  
Found 1 unanswered entries.  
Done.

[EXTRACT]:  
Saved to C:\EI\IS\MLLM\_Final\_Project\LogicVista\eval\results\gpt-4o-accuracy-results.csv

Run Model Evaluation

Gemini

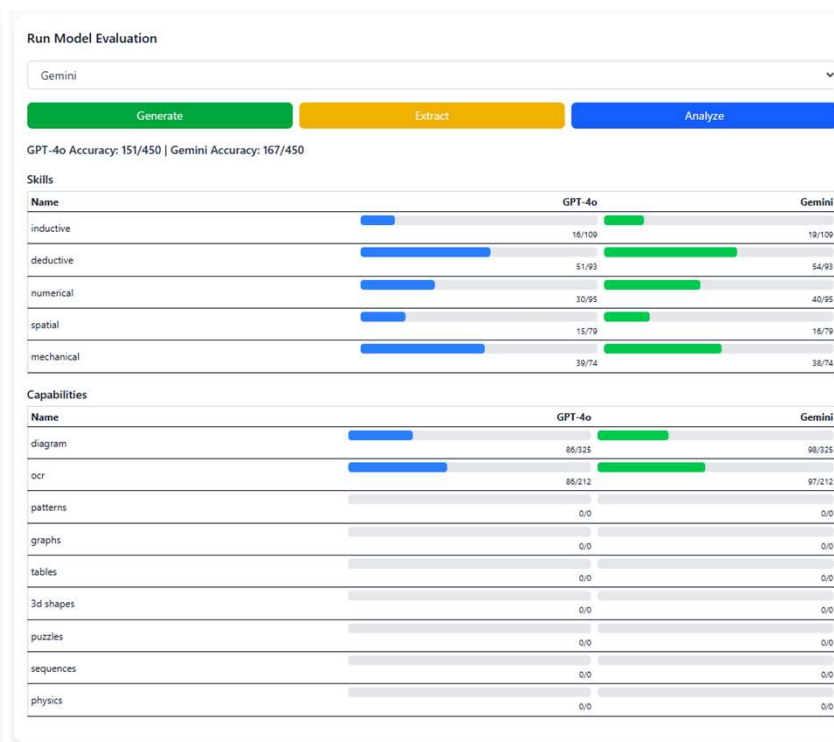
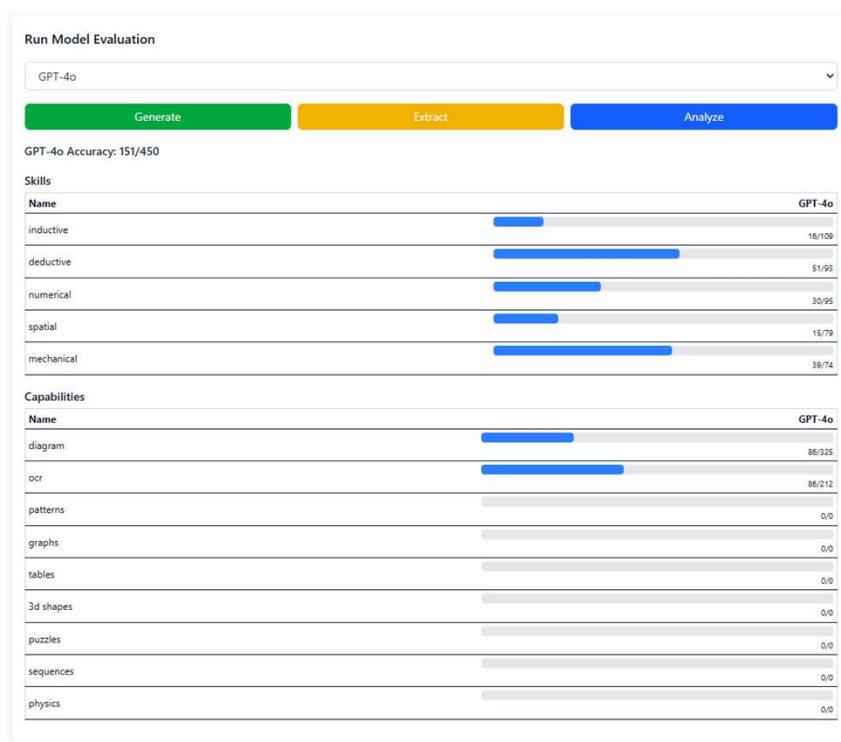
Generate Extract

[GENERATE]:  
Using model: gemini-1.5-flash  
Found 1 unanswered entries.  
Done.

[EXTRACT]:  
Saved to C:\EI\IS\MLLM\_Final\_Project\LogicVista\eval\results\gemini-accuracy-results.csv

## 6 – Pagina de Avaliação

- E por fim analisamos as capacidades do modelo: (para ter a vista de comparação, analisamos primeiro um modelo, depois trocamos para o outro e carregamos em **'Analyze'** de novo)



## 6 – Notas da Plataforma



- Esta interface permite **inserir novas imagens** manualmente para enriquecer o dataset e melhorar a avaliação dos modelos. As imagens devem corresponder a perguntas com **opções visuais claras**, como padrões, diagramas, sequências ou raciocínio espacial. Na secção Add Entry, pode preencher os campos de forma detalhada, selecionando a **Skill**, **Broad Capability** e **Specific Capability** mais adequadas ao tipo de desafio apresentado. Estas categorias são essenciais para análise posterior por tipo de raciocínio.
- O sistema foi desenvolvido com **Next.js** e integra **scripts Python diretamente via API** para: gerar respostas dos modelos (**generate**), extrair a resposta final (**extract**) e calcular a precisão total e por categoria (**analyze**).
- Os dados são guardados em ficheiros estruturados: **eval/answers/{modelo}-answers.json**, **eval/results/{modelo}-accuracy-results**.
- Para correr o sistema **localmente**, devemos aceder à pasta **aivista** e executar o comando: **npm run dev**.

## 7 - Conclusão



A experiência com o benchmark LogicVista demonstrou que o GPT-4o é capaz de interpretar padrões visuais e aplicar raciocínio lógico em contextos estruturados. O modelo respondeu corretamente a diversos desafios, incluindo imagens fora do dataset original, mostrando sinais de generalização. Em termos quantitativos, o GPT-4o obteve o melhor desempenho global entre os modelos avaliados, com destaque para as categorias dedutiva, mecânica e numérica. Além disso, mostrou maior robustez na interpretação de imagens com texto (OCR) e diagramas, o que evidencia a sua capacidade multimodal. Estes resultados indicam que o GPT-4o possui um potencial superior para tarefas que exigem raciocínio lógico visual, superando tanto modelos open-source como closed-source em vários domínios testados. No entanto, limitações como a ausência de integração direta no pipeline oficial do LogicVista e a falta de testes em larga escala sugerem que novas avaliações são necessárias para consolidar estas conclusões.

Num segundo momento, foi testada a robustez do GPT-4o através da manipulação intencional de respostas no dataset. O modelo demonstrou capacidade parcial de corrigir respostas erradas com base na interpretação da imagem, o que reforça a sua autonomia no raciocínio visual. A diferença nos resultados foi marginal, mas suficiente para indicar que o modelo não depende exclusivamente do texto fornecido, validando o seu potencial de generalização e resiliência a ruído nos dados.

Num terceiro momento, foi desenvolvida a interface AlVista, construída com Next.js e scripts Python integrados. Esta plataforma permite inserir novas imagens manualmente, gerar respostas, extrair precisão e comparar o desempenho dos modelos de forma interativa. A AlVista amplia significativamente a utilidade do benchmark, permitindo testes contínuos e colaborativos com maior flexibilidade e profundidade analítica.

o teu • de partida



Instituto Politécnico  
de Viana do Castelo

[www.ipvc.pt](http://www.ipvc.pt)