

# AVALIAÇÃO DE LLM GENERALISTA VS. ESPECIALIZADA NA SÍNTESE DE ESPECIFICAÇÕES FORMAIS

Felipe Tabosa

Filipe Eduardo

José Izaias

Karen Samara

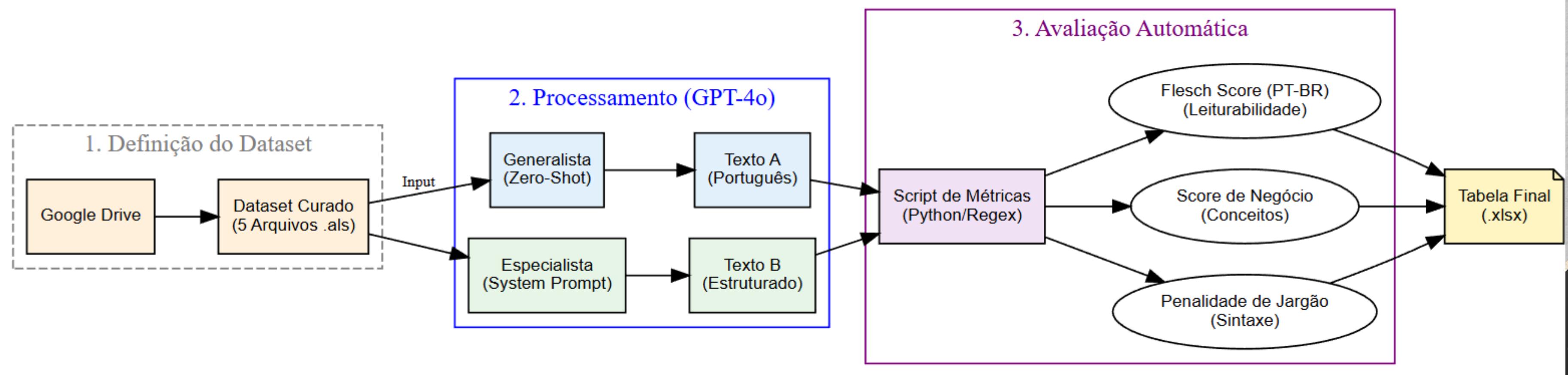


# Índice

- Contextualização
- Metodologia
- Dataset
- Código
- Explicação
- Resultados finais
- Conclusão



# Fluxograma



# Contextualização

Validação de modelos

GPT 4o

GPT 5

GPT 5.1

Gemini Flash

Gemini Pro

Claude Code



# Contextualização

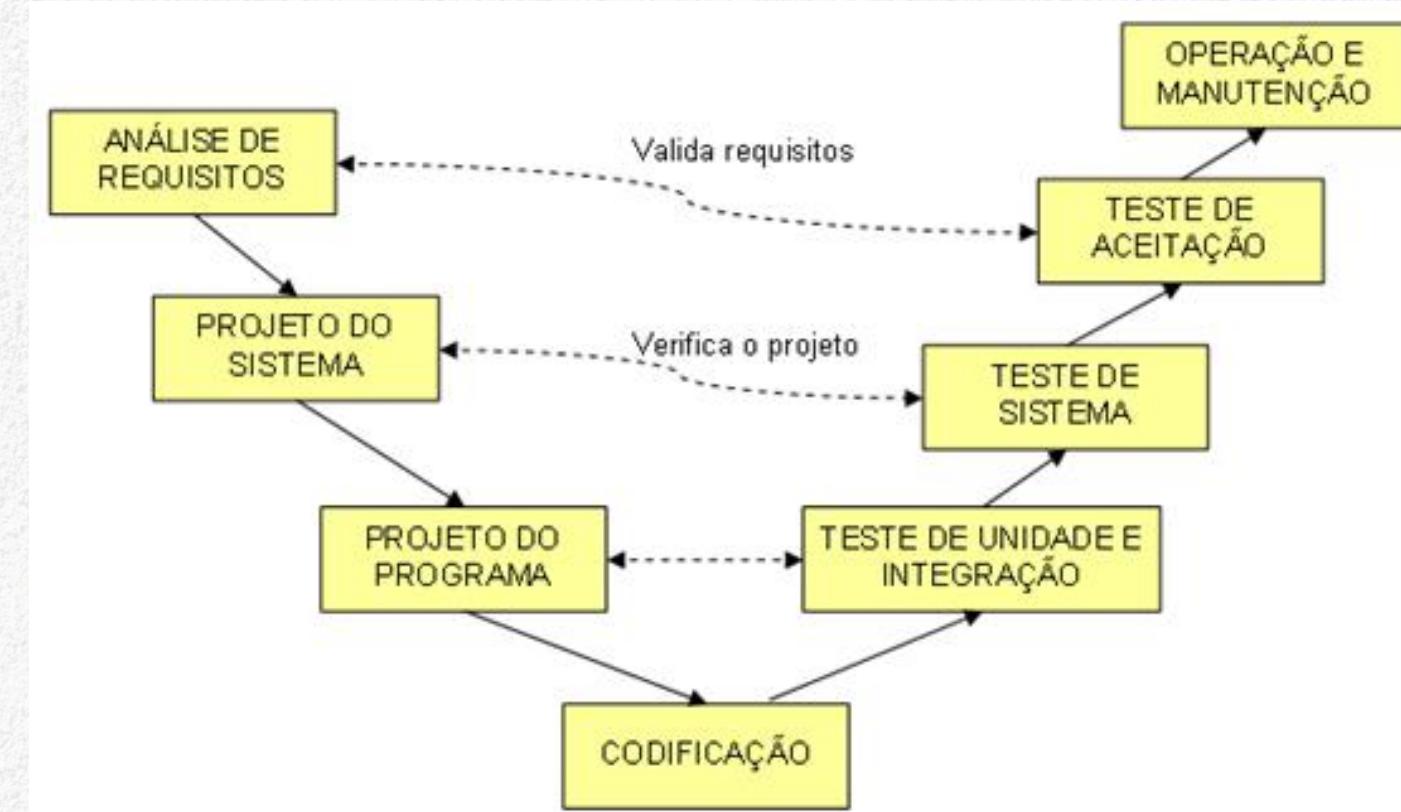
## Validação de modelos

GPT 4o foi o modelo que apresentou resultados mais coerentes entre as avaliações

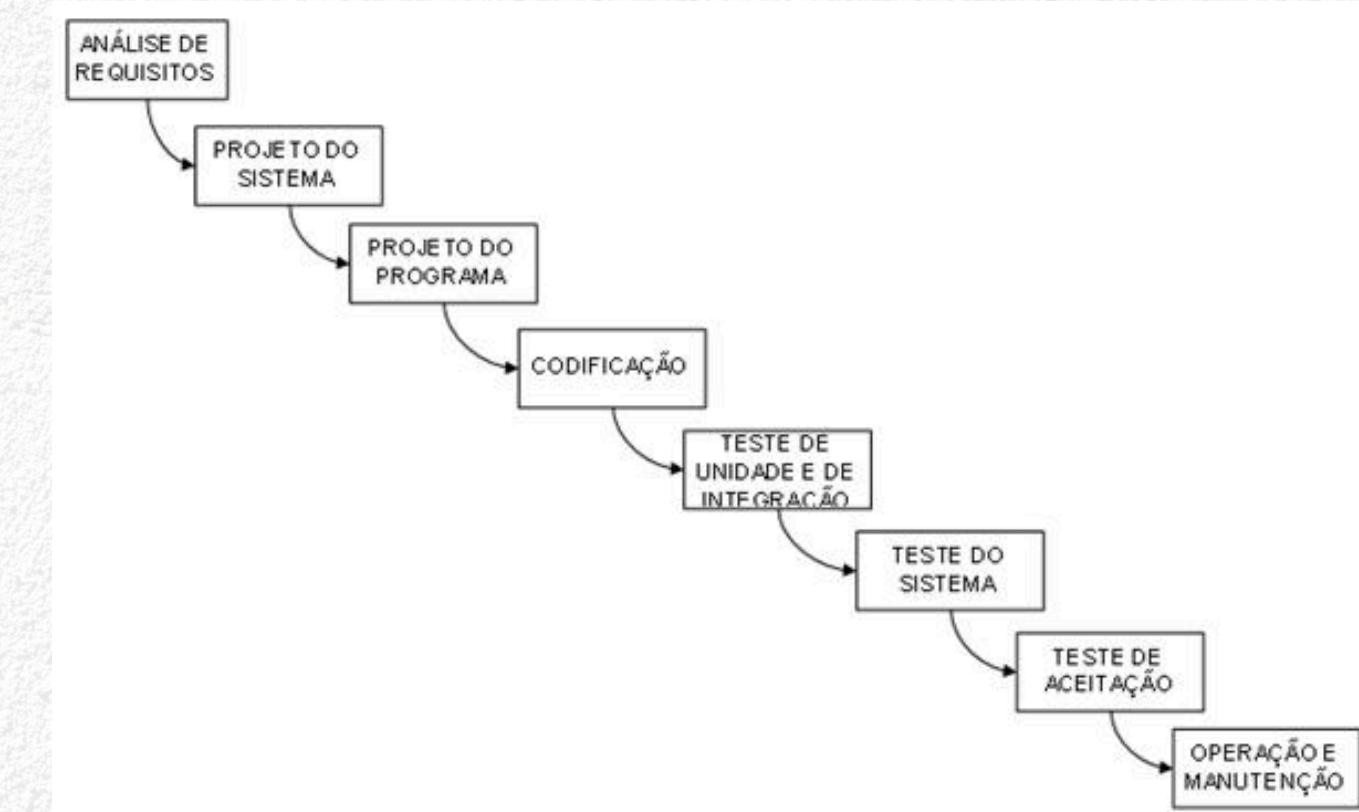


# Contextualização

## Ciclo de Software



Modelo em V



Modelo em Cascata



# **Contextualização**

## **Ciclo de Software**

**Criação de requisitos formais para garantir a funcionalidade.**

**Explicação dos conceitos estabelecidos para o stakeholder**

**Como facilitar essa etapa de explicação e  
possíveis erros de interpretação?**



# Metodologia

## Generalista

- Instruções escritas em formato informal
- Sem delimitações na saída
- Sem restrições de linguagem

## Modelo - GPT 4o

**"Explique em Português do Brasil o que esse código faz, lembre que não tenho conhecimento em programação e lógica"**



# Metodologia

## Especialista

- Explicação de como será a entrada
- Definição do limite da linguagem de saída, como por exemplo uma linguagem leve, sem jargões técnicos, interpretação e não apenas leitura
- Definição do formato de saída

## Modelo - GPT 4o

O formato final da resposta deve seguir exatamente:

1. Resumo Executivo
2. Explicação Detalhada da Especificação Formal
3. Pontos Críticos, Restrições e Alertas



# Metodologia

## Avaliação dos Resultados

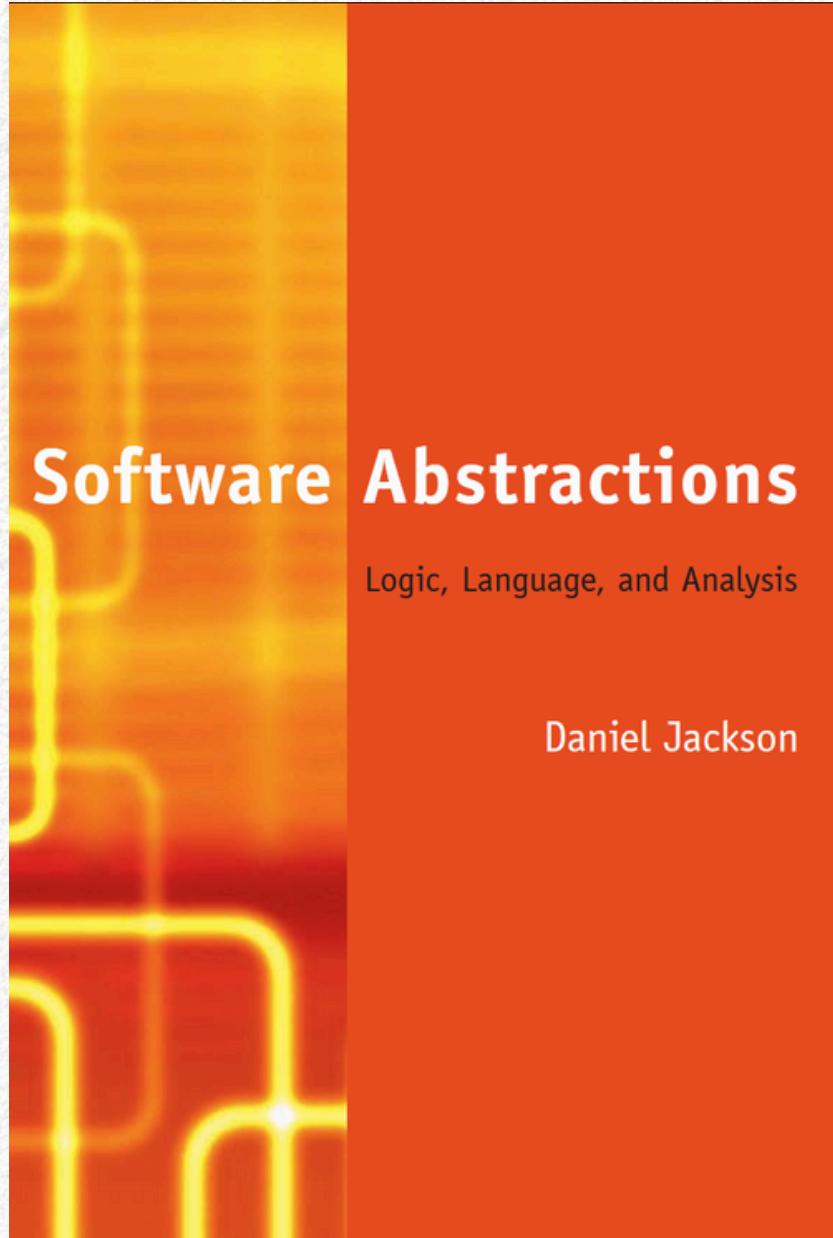
Modelo - GPT 4o

- Sistema de pontuação pela utilização de termos informais e jargões técnicos
- Utilização da métrica Flesch - Índice de Legibilidade Flesch, adaptada para o Português

Índice Flesch =  $227 - (1,04 \times \text{média de palavras por sentença}) - (72 \times \text{média de sílabas por palavra})$



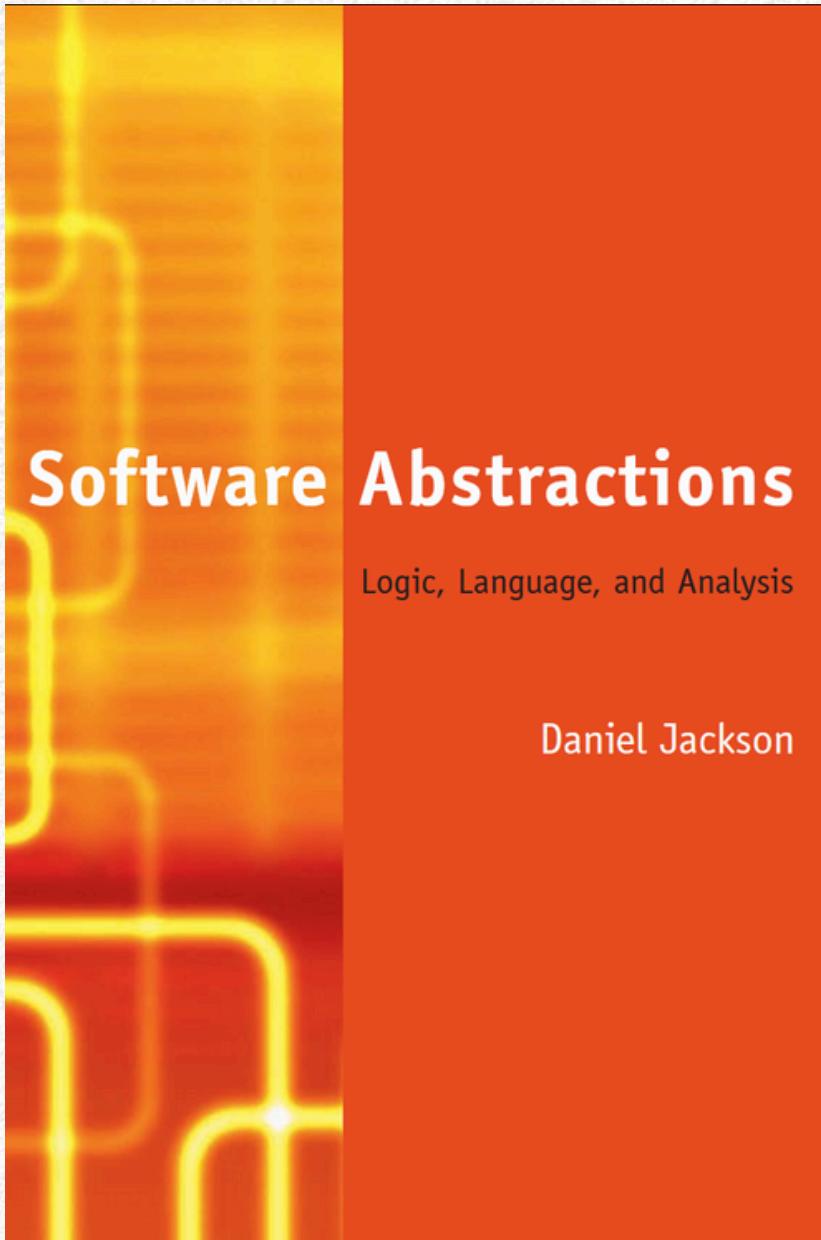
# Dataset



## Software Abstractions - MIT

Subconjunto dos modelos formais clássicos do  
livro de Daniel Jackson

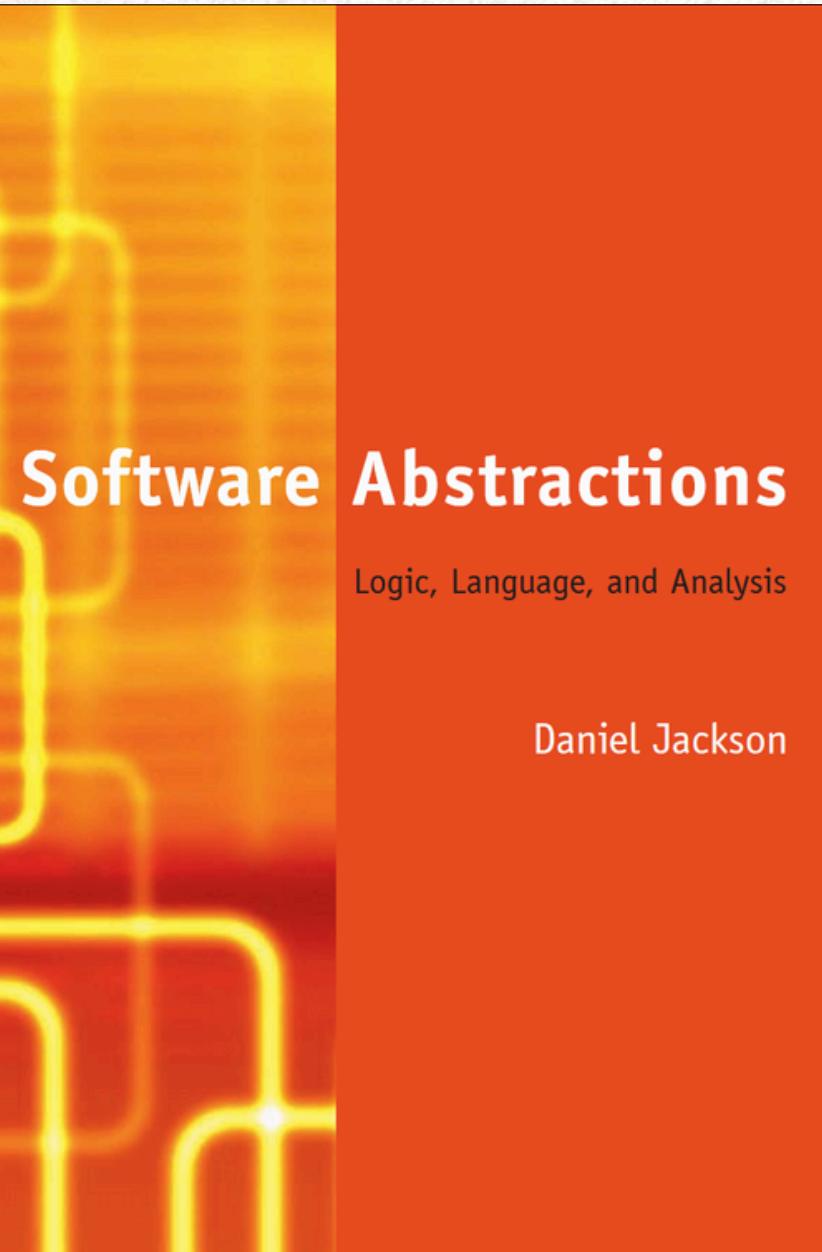
# Dataset



## Motivação

- Modelos amplamente usados em ensino/pesquisa
- Documentação confiável
- Diversidade semântica e estrutural
- Permite comparar LLMs em tarefas realmente formais
- Evita enviesamento por modelos inventados especificamente para o experimento

# Dataset



## Arquivos Selecionados

Arquivo	Dificuldade	Conceito Testado
<b>addressBook.als</b>	● Fácil	Mapeamento simples, conjuntos, funções parciais
<b>properties.als</b>	● Médio	Propriedades matemáticas: transitividade, injeção, axiomas
<b>filesystem.als</b>	● Médio	Hierarquia, recursão, fechamento transitivo (^)
<b>grandpa2.als</b>	● Difícil	Paradoxo genealógico, detecção de inconsistências semânticas
<b>ringElection2.als</b>	● Difícil	Algoritmos distribuídos e temporalidade

# Dataset

## grandpa2.als

```
abstract sig Person {
    father: lone Man,
    mother: lone Woman
}

sig Man extends Person {
    wife: lone Woman
}

sig Woman extends Person {
    husband: lone Man
}

fact {
    no p: Person | p in p.^{mother+father}
    wife = ~husband
}

assert NoSelfFather {
    no m: Man | m = m.father
}

fun grandpas [p: Person] : set Person {
    let parent = mother + father + father.wife + mother.husband
    |
    p.parent.parent & Man
}

pred ownGrandpa [p: Person] {
    p in p.grandpas
}
```



# O código - arquitetura da pipeline de avaliação

- Stack: Python 3.10 • Google Colab • OpenAI API • Pandas
- Automação e Ingestão:
  - Conexão direta com Google Drive via script.
  - Carregamento dinâmico de arquivos .als com filtro de prioridade.
- Controle de Variáveis (Agentes):
  - Funções modulares: Zero-Shot (Generalista) vs. System Prompting (Especialista).
- Motor de Métricas Customizado:
  - Implementação manual do algoritmo Flesch (PT-BR) para avaliar leitabilidade.
  - Análise léxica via Regex para pontuar "Foco no Negócio" vs. "Jargão Técnico".
- Processamento em Lote:
  - Execução sequencial automatizada com exportação de dados para Excel/CSV.



# Métricas



	Arquivo	Gen_Score_Negocio	Gen_Jargao	Gen_Leiturabilidade_PT	Esp_Score_Negocio	Esp_Jargao	Esp_Leiturabilidade_PT	Gen_Texto	Esp_Texto
0	grandpa2.als	10	14	72.985285	14	6	45.143530	Este código está escrito em Alloy, uma linguag...	1. Resumo Executivo\n\nEste modelo descreve ...
1	addressBook.als	3	10	81.000998	10	7	46.021160	Este código é escrito em Alloy, uma linguagem ...	1. Resumo Executivo\n\nEste modelo representa ...
2	filesystem.als	7	1	46.692143	8	3	26.827836	Vou explicar o que esse código faz de uma mane...	1. Resumo Executivo\n\nEste modelo representa ...
3	ringElection2.als	14	9	44.453243	18	4	42.681282	Este código é escrito em Alloy, uma linguagem ...	1. Resumo Executivo\n\nEste modelo representa ...
4	properties.als	4	11	46.824481	6	10	33.066855	Este código está escrito em uma linguagem de m...	1. Resumo Executivo\n\nEste modelo formal em A...

# Conclusão

Historicamente os Métodos Formais eram restritos a uma elite acadêmica ou industrial devido à curva de aprendizado. Utilizamos a IA como Ponte Semântica convertendo lógica matemática em requisitos de negócio legíveis.

Os resultados obtidos possuem diferenças notáveis.

- Generalista se utiliza de termos técnicos mas traz uma abordagem mais fluida para o leitor
- Especialista destrincha o conteúdo para uma linguagem e formato mais acessível, mas acaba deixando uma leitura mais densa pela quantidade de conteúdo.



# OBRIGADO!

---

