AIware 2025

Human to Document, AI to Code: Comparing GenAI for Notebook Competitions

Tasha Settewong, Youmei Fan, Raula Gaikovina Kula, Kenichi Matsumoto



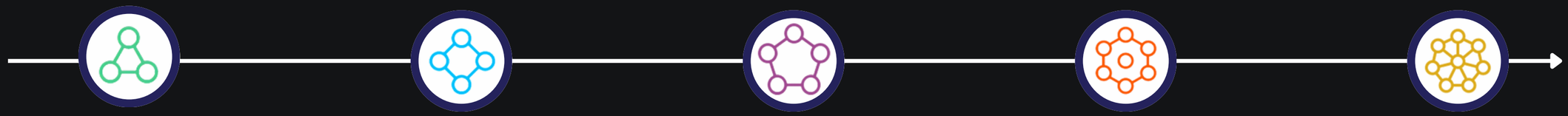

Pre-print

Computational notebook

> - Computational notebook is a well-known and well-adopted technology in tasks related to **data analysis**
> - A **Jupyter notebook** can be a central place for collaborative data analysis.

Kaggle

A cloud-based collaborative platform involving data analytics tasks using a computational notebook in practices



Novice

A new user who joins Kaggle

Contributor

A user who has completed a profile engaged with the community and fully explored the platform of Kaggle.

Expert

A user who receives 5 bronze medals

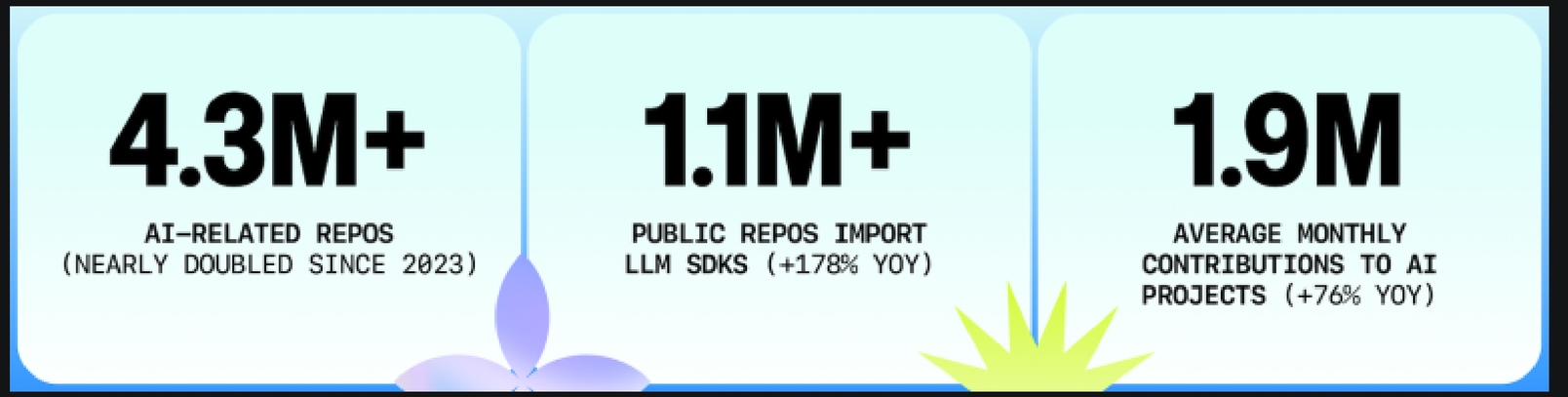
Master

A user who receives 10 silver medals

Grandmaster

A user who receives 15 gold medals

Problem Statements



1 To what extent can the notebook be considered as a **good quality notebook**

2 How can we **distinguish** between notebooks created by **humans** and those generated by **GenAI**?

GitHub. (n.d.). Generative AI and agentic workflows become ordinary engineering. https://github.blog/news-insights/octoverse/octoverse-a-new-developer-joins-github-every-second-as-ai-leads-typescript-to-1/

Research Questions

RQ1

What are **differences** between documentation and code for **human-written** notebook?

> What are the differences in code and documentation between **Gold-Medal** and **Non-Gold** human-written notebooks?

RQ2

What are the key **differences** between **GenAI** and **human-written** data science notebooks?

> How do **medal-winning** human notebooks compare to notebooks generated by **GenAI** models (GPT-4, Llama, Gemini)?

Methodology

Data Collection

Analyzed **465 medal-winning notebooks** from 3 most participant Kaggle competitions (KGTorrent dataset).

Generate GenAI Notebooks

Used 3 LLMs

- GPT-4.1
- Llama 4 Maverick
- Gemini 2.5 Pro Review

to generate **9 notebooks** for the same competitions using a single prompt.

Extract Features

Extracted **25 features** catagorized into

- Documentation-related attributes
- Code quality-related attributes

Analyze Results

Applied statistical tests to answer research questions

Results

RQ1: What are differences between documentation and code for human-written notebook?

Gold medalists are primarily distinguished by their **documentation**.
Compared to Non-Gold notebooks, Gold notebooks have significantly more:

- The number of markdown characters
- The number of markdown lines
- The number of sentences

Table1: Statistical Comparisons of Gold and Non-Gold Human-Written Notebooks

Feature	Home-Credit		Santander		IEEE-CIS	
	p-value	η^2	p-value	η^2	p-value	η^2
#Markdown Char	**0.0005*** **	0.085	**0.0004*** **	0.070	–	–
#Markdown Line	**0.001** **	0.074	**0.0004*** **	0.071	–	–
#Sentences	**0.001** **	0.073	**0.0004*** **	0.071	–	–
#Markdown Cell	**0.001** **	0.070	**0.001*** **	0.056	–	–
Avg Sentence	**0.013***	0.040	–	–	–	–
Duplicated Lines	**0.021***	0.033	**0.020***	0.026	**0.004** **	0.046
Duplicated Blocks	–	–	**0.009** **	0.035	**0.004** **	0.045
#Code Char	–	–	–	–	**0.011***	0.033
#Visual	–	–	–	–	**0.019***	0.027

* p-value < 0.05; ** p-value < 0.01; and *** p-value < 0.001.
The effect sizes; small, medium, large

Results

RQ1: What are differences between documentation and code for human-written notebook?

Gold medalists are primarily distinguished by their **documentation**.
Compared to Non-Gold notebooks, Gold notebooks have significantly more:

- The number of markdown characters
- The number of markdown lines
- The number of sentences

Table1: Statistical Comparisons of Gold and Non-Gold Human-Written Notebooks

Feature	Home-Credit		Santander		IEEE-CIS	
	p-value	η^2	p-value	η^2	p-value	η^2
#Markdown Char	**0.0005*** **	0.085	**0.0004*** **	0.070	–	–
#Markdown Line	**0.001** **	0.074	**0.0004*** **	0.071	–	–
#Sentences	**0.001** **	0.073	**0.0004*** **	0.071	–	–
#Markdown Cell	**0.001** **	0.070	**0.001*** **	0.056	–	–
Avg Sentence	**0.013***	0.040	–	–	–	–
Duplicated Lines	**0.021***	0.033	**0.020***	0.026	**0.004** **	0.046
Duplicated Blocks	–	–	**0.009** **	0.035	**0.004** **	0.045
#Code Char	–	–	–	–	**0.011***	0.033
#Visual	–	–	–	–	**0.019***	0.027

*** p-value < 0.05; ** p-value < 0.01; and *** p-value < 0.001.**
The effect sizes; small, medium, large

Table 2: Gold vs Non-Gold Feature Statistics Comparisons

Feature	Medal Type	Mean	Median	Std	Min	Max
#Markdown Char	Non-Gold	2688.10	934.0	6526.72	0	94377
	Gold	5369.35	2261.0	8071.63	0	45281
#Markdown Line	Non-Gold	35.02	16.0	67.15	0	748
	Gold	64.98	35.5	79.04	0	364
#Sentences	Non-Gold	27.39	12.0	44.14	0	308
	Gold	56.46	26.5	75.81	0	364
#Markdown Cell	Non-Gold	11.71	6.0	18.57	0	184
	Gold	21.12	12.0	27.64	0	160

Results

RQ2: What are the key differences between GenAI and human-written data science notebooks?

- **Human documentation** is **easier to read** (lower Gunning Fog score).
- **GenAI** notebooks produces **cleaner static code** with
  - fewer code smells, technical debt, and violations
- while human-written code has
  - higher function counts and statements.
- 3 out of 9 **GenAI** notebook **failed to run** due to logical or data-handling errors.

Table 3: Statistical Comparisons of GenAI and Each Medal Notebooks

Feature	Medal	Home-Credit		Santander		IEEE-CIS	
		p-value	η^2	p-value	η^2	p-value	η^2
Functions	Gold	**0.011***	0.197	–	–	**0.041***	0.078
	Silver	**0.009****	0.191	–	–	–	–
	Bronze	**0.023***	0.054	–	–	–	–
Statements	Gold	**0.025***	0.143	–	–	–	–
	Silver	**0.018***	0.148	–	–	–	–
	Bronze	–	–	–	–	–	–
Comment Lines	Gold	**0.035***	0.122	–	–	–	–
	Silver	–	–	–	–	–	–
	Bronze	–	–	–	–	–	–
#Code Cell	Gold	**0.049***	0.103	–	–	**0.025***	0.098
	Silver	–	–	–	–	–	–
	Bronze	–	–	–	–	–	–
Cyclomatic Complexity	Gold	**0.050***	0.102	–	–	–	–
	Silver	**0.028***	0.124	–	–	–	–
	Bronze	–	–	–	–	**0.043***	0.034
Gunning Fog	Gold	–	–	**0.017***	0.124	**0.032***	0.088
	Silver	–	–	–	–	**0.030***	0.107
	Bronze	–	–	**0.024***	0.046	**0.049***	0.031
Code Smells	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	**0.004****	0.209
	Bronze	–	–	–	–	–	–
Technical Debts	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	**0.004****	0.209
	Bronze	–	–	–	–	–	–
Violations	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	**0.006****	0.191
	Bronze	–	–	–	–	–	–

* p-value < 0.05; ** p-value < 0.01; and *** p-value < 0.001.
The effect sizes; small, medium, large

Results

RQ2: What are the key differences between GenAI and human-written data science notebooks?

- **Human documentation** is **easier to read** (lower Gunning Fog score).
- **GenAI** notebooks produces **cleaner static code** with
  - fewer code smells, technical debt, and violations
- while human-written code has
  - higher function counts and statements.
- 3 out of 9 **GenAI** notebook **failed to run** due to logical or data-handling errors.

Table 3: Statistical Comparisons of GenAI and Each Medal Notebooks

Feature	Medal	Home-Credit		Santander		IEEE-CIS	
		p-value	η^2	p-value	η^2	p-value	η^2
Functions	Gold	**0.011***	0.197	–	–	**0.041***	0.078
	Silver	**0.009****	0.191	–	–	–	–
	Bronze	**0.023***	0.054	–	–	–	–
Statements	Gold	**0.025***	0.143	–	–	–	–
	Silver	**0.018***	0.148	–	–	–	–
	Bronze	–	–	–	–	–	–
Comment Lines	Gold	**0.035***	0.122	–	–	–	–
	Silver	–	–	–	–	–	–
	Bronze	–	–	–	–	–	–
#Code Cell	Gold	**0.049***	0.103	–	–	**0.025***	0.098
	Silver	–	–	–	–	–	–
	Bronze	–	–	–	–	–	–
Cyclomatic Complexity	Gold	**0.050***	0.102	–	–	–	–
	Silver	**0.028***	0.124	–	–	–	–
	Bronze	–	–	–	–	**0.043***	0.034
Gunning Fog	Gold	–	–	**0.017***	0.124	**0.032***	0.088
	Silver	–	–	–	–	**0.030***	0.107
	Bronze	–	–	**0.024***	0.046	**0.049***	0.031
Code Smells	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	**0.004****	0.209
	Bronze	–	–	–	–	–	–
Technical Debts	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	**0.004****	0.209
	Bronze	–	–	–	–	–	–
Violations	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	**0.006****	0.191
	Bronze	–	–	–	–	–	–

* p-value < 0.05; ** p-value < 0.01; and *** p-value < 0.001.
The effect sizes; small, medium, large

Table 4: Human vs GenAI Feature Statistics Comparison in IEEE-CIS Competition

Feature	Source	Mean	Median	Min	Max
Functions	Human	3.16	2.0	0	46
	AI	0.67	0.0	0	2
Statements	Human	121.49	91.5	10	586
	AI	94.33	35.0	25	223
Comment Lines	Human	40.93	23.5	0	252
	AI	32.67	18.0	13	67
#Code Cell	Human	31.36	22.0	2	205
	AI	9.00	8.0	7	12
Cyclomatic Complexity	Human	14.30	9.5	0	126
	AI	14.67	0.0	0	44
Gunning Fog	Human	10.32	9.91	0	64.61
	AI	16.07	16.13	11.55	20.53
Code Smells	Human	29.25	0.0	0	3166
	AI	3.00	3.0	3	3
Technical Debts	Human	44.39	0.0	0	3228
	AI	15.00	15.0	15	15
Violations	Human	29.59	0.0	0	3167
	AI	3.00	3.0	3	3

Problem Statements

4.3M+
AI-RELATED REPOS (NEARLY DOUBLED SINCE 2023)

1.1M+
PUBLIC REPOS IMPORT LLM SDKS (+178% YOY)

1.9M
AVERAGE MONTHLY CONTRIBUTIONS TO AI PROJECTS (+76% YOY)

1 To what extent can the notebook be considered as a **good quality notebook**

2 How can we **distinguish** between notebooks created by **humans** and those generated by **GenAI**?

GitHub. (n.d.). Generative AI and agentic workflows become ordinary engineering. https://github.blog/news-insights/octoverse/octoverse-a-new-developer-joins-github-every-second-as-ai-leads-typescript-to-1/

Methodology

Data Collection
Analyzed **465 medal-winning notebooks** from 3 most participant Kaggle competitions (KGTorrent dataset).

PROMPT

Generate GenAI Notebooks
Used 3 LLMs
- GPT-4.1
- Llama 4 Maverick
- Gemini 2.5 Pro Review

to generate **9 notebooks** for the same competitions using a single prompt.

Extract Features
Extracted **25 features** catagorized into
- Documentation-related attributes
- Code quality-related attributes

Analyze Results
Applied statistical tests to answer research questions

Results

RQ1: What are differences between documentation and code for human-written notebook?

Gold medalists are primarily distinguished by their **documentation**.
Compared to Non-Gold notebooks, Gold notebooks have significantly more:
- The number of markdown characters
- The number of markdown lines
- The number of sentences

Statistical Comparisons of Gold and Non-Gold Human-Written Notebooks

Feature	Home-Credit		Santander		IEEE-CIS	
	p-value	η^2	p-value	η^2	p-value	η^2
#Markdown Char	0.0005***	0.085	0.0004***	0.070	–	–
#Markdown Line	0.001**	0.074	0.0004***	0.071	–	–
#Sentences	0.001**	0.073	0.0004***	0.071	–	–
#Markdown Cell	0.001**	0.070	0.001***	0.056	–	–
Avg Sentence	0.013*	0.040	–	–	–	–
Duplicated Lines	0.021*	0.033	0.020*	0.026	0.004**	0.046
Duplicated Blocks	–	–	0.009**	0.035	0.004**	0.045
#Code Char	–	–	–	–	0.011*	0.033
#Visual	–	–	–	–	0.019*	0.027

* p-value < 0.05; ** p-value < 0.01; and *** p-value < 0.001.
The effect sizes; small, medium, large

Gold vs Non-Gold Feature Statistics Comparisons

Feature	Medal Type	Mean	Median	Std	Min	Max
#Markdown Char	Non-Gold	2688.10	934.0	6526.72	0	94377
	Gold	5369.35	2261.0	8071.63	0	45281
#Markdown Line	Non-Gold	35.02	16.0	67.15	0	748
	Gold	64.98	35.5	79.04	0	364
#Sentences	Non-Gold	27.39	12.0	44.14	0	308
	Gold	56.46	26.5	75.81	0	364
#Markdown Cell	Non-Gold	11.71	6.0	18.57	0	184
	Gold	21.12	12.0	27.64	0	160

Results

RQ2: What are the key differences between GenAI and human-written data science notebooks?

- **Human documentation** is **easier to read** (lower Gunning Fog score).
- **GenAI** notebooks produces **cleaner static code** with
  - fewer code smells, technical debt, and violations
- while human-written code has
  - higher function counts and statements.
- 3 out of 9 **GenAI** notebook **failed to run** due to logical or data-handling errors.

Table 3: Statistical Comparisons of GenAI and Each Medal Notebooks

Feature	Medal	Home-Credit p-value	Home-Credit η^2	Santander p-value	Santander η^2	IEEE-CIS p-value	IEEE-CIS η^2
Functions	Gold	0.011*	0.107	–	–	0.041*	0.078
	Silver	0.009**	0.191	–	–	–	–
	Bronze	0.023*	0.054	–	–	–	–
Statements	Gold	0.025*	0.143	–	–	–	–
	Silver	0.018*	0.148	–	–	–	–
	Bronze	–	–	–	–	–	–
Comment Lines	Gold	0.035*	0.122	–	–	–	–
	Silver	–	–	–	–	–	–
	Bronze	–	–	–	–	–	–
#Code Cell	Gold	0.049*	0.103	–	–	0.025*	0.098
	Silver	–	–	–	–	–	–
	Bronze	–	–	–	–	–	–
Cyclomatic Complexity	Gold	0.050*	0.102	–	–	–	–
	Silver	0.028*	0.124	–	–	–	–
	Bronze	–	–	–	–	0.043*	0.034
Gunning Fog	Gold	–	–	0.017*	0.124	0.032*	0.088
	Silver	–	–	–	–	0.030*	0.107
	Bronze	–	–	0.024*	0.046	0.049*	0.031
Code Smells	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	0.004**	0.209
	Bronze	–	–	–	–	–	–
Technical Debts	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	0.004**	0.209
	Bronze	–	–	–	–	–	–
Violations	Gold	–	–	–	–	–	–
	Silver	–	–	–	–	0.006**	0.191
	Bronze	–	–	–	–	–	–

* p-value < 0.05; ** p-value < 0.01; and *** p-value < 0.001.
The effect sizes; small, medium, large

Table 4: Human vs GenAI Feature Statistics Comparison in IEEE-CIS Competition

Feature	Source	Mean	Median	Min	Max
Functions	Human	3.16	2.0	0	46
	AI	0.67	0.0	0	2
Statements	Human	121.49	91.5	10	586
	AI	94.33	35.0	25	223
Comment Lines	Human	40.93	23.5	0	252
	AI	32.67	18.0	13	67
#Code Cell	Human	31.36	22.0	2	205
	AI	9.00	8.0	7	12
Cyclomatic Complexity	Human	14.30	9.5	0	126
	AI	14.67	0.0	0	44
Gunning Fog	Human	10.32	9.91	0	64.61
	AI	16.07	16.13	11.55	20.53
Code Smells	Human	29.25	0.0	0	3166
	AI	3.00	3.0	3	3
Technical Debts	Human	44.39	0.0	0	3228
	AI	15.00	15.0	15	15
Violations	Human	29.59	0.0	0	3167
	AI	3.00	3.0	3	3