

From Hype to Value: Mastering Gen AI Outcomes Through Evaluations



June 5th 2025

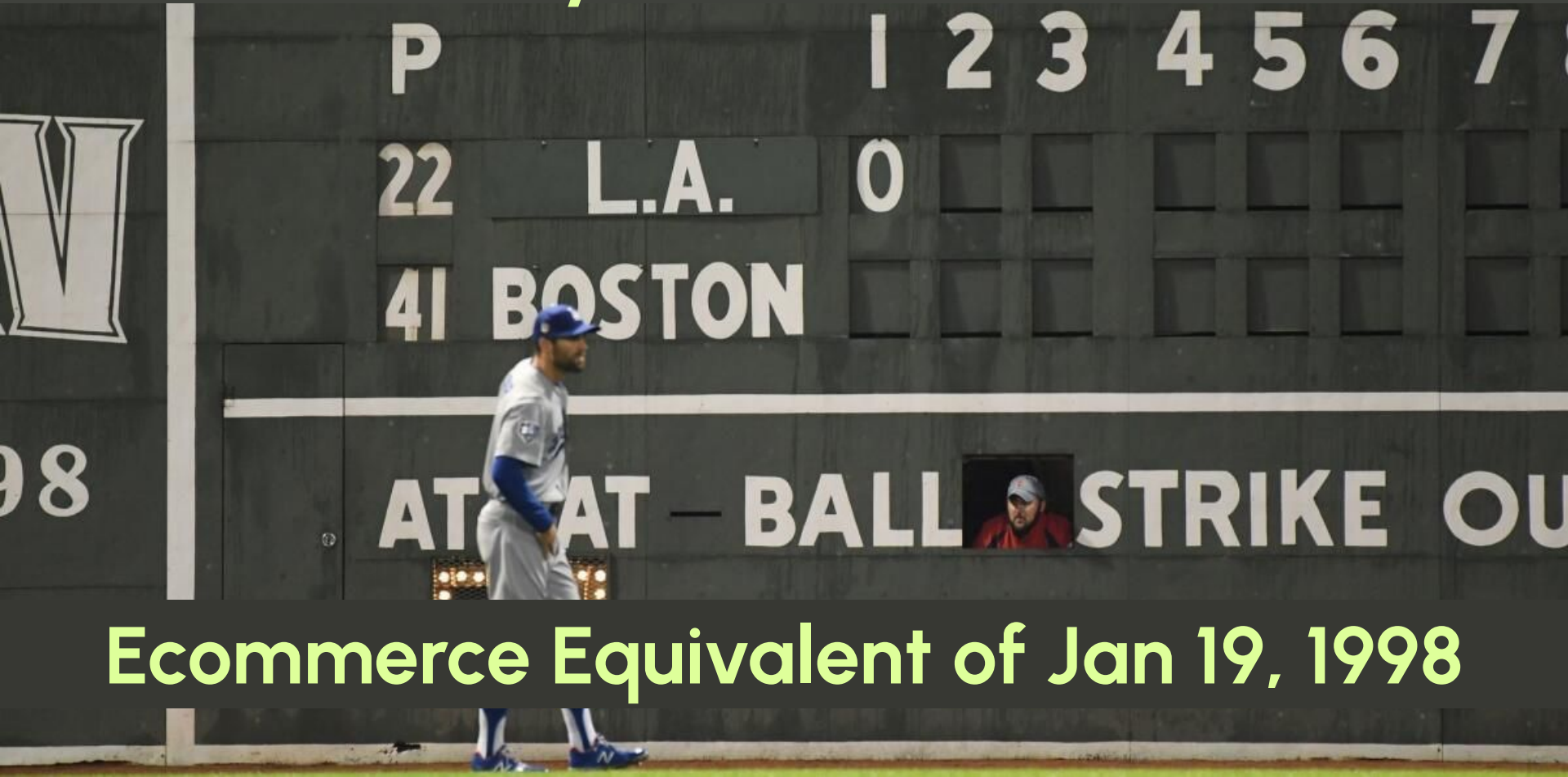
Presented by

Bill Gold
BillCGold@gmail.com



918 Days

918 Days, Since Chat GPT



Ecommerce Equivalent of Jan 19, 1998

Agenda

Gen AI Evaluations

1 Evaluations & Intuitions



2 Approaches to Consider



3 Better Practices



4 Q & A



Underwhelming Gen AI Results?

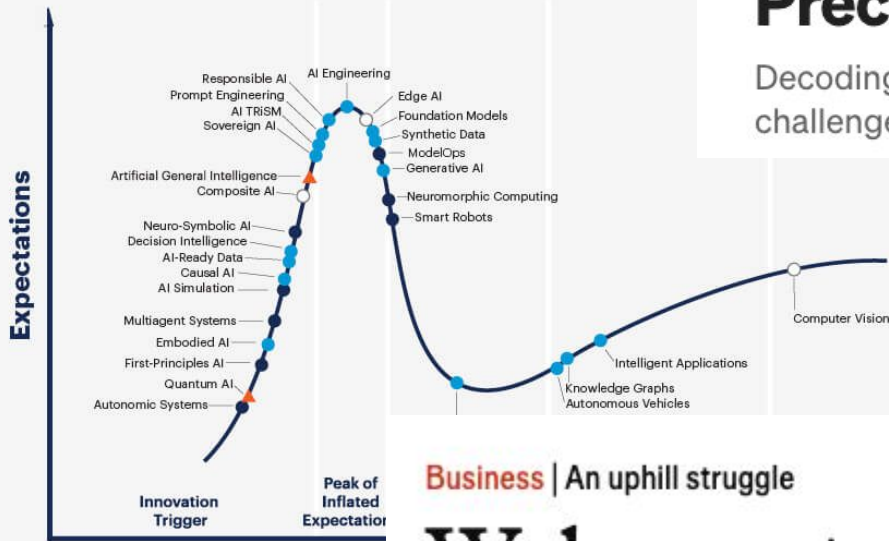


Google Gemini 2.5



Chat GPT 4.5

Hype Cycle for Artificial Intelligence



Source: Gartner
Commercial reuse requires approval from Gartner and must comply with the Gartner Content Compliance Policy on [gartner.com](https://www.gartner.com/content-compliance-policy).
© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. GTS_3282450

The AI Hype Cycle: Are We on the Precipice of Disillusionment?

Decoding generative AI's profitability dilemma and regulatory challenges

The Economist

Business | An uphill struggle

Welcome to the AI trough of disillusionment

Generative
of disillusionment

LLMs GET LOST IN MULTI-TURN CONVERSATION

1. Effective Evaluations, Intuitions

Exceeds Expectations ...

✦ GPT-4

BI

python nltk code to extract first name, last name and job title from free form text



The task you're trying to achieve is a common one in the field of natural language processing and is often called named entity recognition (NER). While NLTK does have some capabilities in this area, it doesn't provide the most accurate or nuanced results for NER tasks out of the box. There are other libraries, like Spacy, that often perform better on these tasks.

However, you specifically asked for a solution using NLTK, so I'll provide that first. Please note that NLTK's ne_chunk method is quite basic, and the following code will only work reliably for very standardized input where the name is directly followed by the job title.

```
python import nltk
```

```
import nltk
```


Underwhelming
too...

INNING BY-INNING												
INNING 1	RUR			PAM			RUNS			RLS		
1	1						1					
BOBUME	2	11	4	9	3	5	6	7	HITS	ERRORS		
BABEL	1	2	8	0	1	0	1	1		1		
N-YTYR	N	1	2	0	4	0	1	0	1			
CIAMESS												
INNINS GOS												
INNING 2	8			RUNS			HITS			HITS		
0												
INNING 3	1	5	7	5	4	8	18	13	HITS	ERRORS		
BOPALL	3	4	5	7	4	0	5	6	1	1		
LGAND	4	7	6	4	9	8	1					
FRATTER	6	6	7	9	0	3	2			0		
CRALS	4	5	6	7	0	8	5					

INNING BY-INNING												
INNING 1	RUR			PAM			RUNS			RLS		
	2			0			1			3		
INNMINC	2	11	4	9	3	5	6	7	HITS	ERRORS		
CHOME											4	
CANDES												
CIAMESS											0	
INNINS EOS												
INNING 2	8			RUNS			HITS			HITS		
							3			6		
INNING 3	1	5	7	5	4	8	18	13	HITS	ERRORS		
BOPALL										1	0	5
LGAND												
FRATTER										1	0	5
CRALS										6	1	5

B

produce an image of a baseball inning by inning score board,
focus on the early inning 1 thru 3




Quality Variances Exist

A Key Success Differentiator

Quality Drivers

- Lossy Models
- Scaled Wisdom: Reinforcement Learning from Human Feedback (RLHF), In Some Domains
- Data (Documents) Not Gen AI Friendly, Yet
- Immature Products
- Skill Gaps, e.g. Prompting
- Quality Metrics Gaps

LLMs are **Lossy** an Intuition

Input Size	Training	Output Size
10 TB 	6k GPU  12 Days \$2M	140 GB 
Text, Chunk of the Internet (Common Crawl)	Cost to Train Base Model	Llama 2, 70B Parameters

Source: Andrej Karpathy Intro to LLMs https://www.youtube.com/watch?v=zjkBMFhNj_g&t=909s

Bill Gold - June 5th 2025

What is Reinforcement Learning from Human Feedback (RLHF)?

One Prompt, Multiple Answers, Experts Rank | Rate ...
“Wisdom” is captured and fed back into model training.

I. Illustration, One Prompt: "What are the best ways to learn a new language?"

II. LLM Multiple Responses:

1. "Use apps."
2. "Immerse yourself by traveling, use language learning apps, practice with native speakers, and take classes."
3. "Watch movies in that language."
4. "Just read books."

III. Human Ranking: [Response 2, Response 3, Response 1, Response 4]

Domains Feedback Combinations Matters

713



SIC Codes (6 Digit)

105+



Georgia Tech Majors & Minors

100+



Agencies, Federal Government

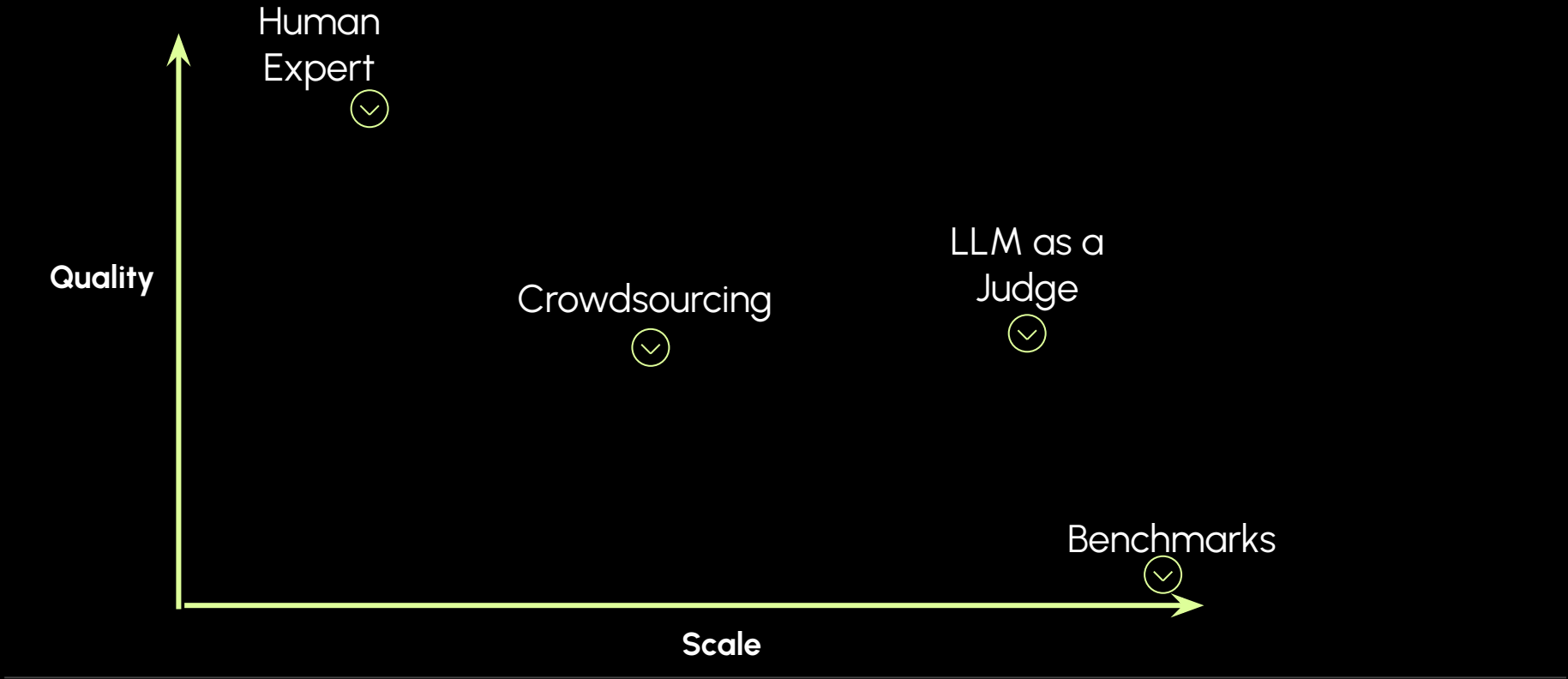
Which Domain(s) are Critical?
Which Domains are Being Invested In?
Do They Align?

Quality Drivers

- Lossy Models
 - Scaled Wisdom: Reinforcement Learning from Human Feedback (RLHF), In Some Domains
- Data (Documents) Not Gen AI Friendly, Yet
 - Immature Products
 - Prompting Skills
 - Existing quality metrics

2. Approaches to Consider

Evaluation Approaches



Benchmarks

Examples

- HumanEval
- StaticEval

Trade Offs

Plus

- Some Scale
- Economical
- Coding Oriented

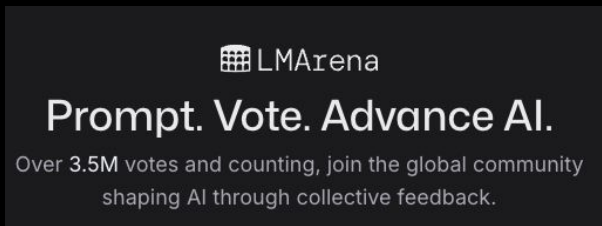
Minus

- Coding Oriented
- LLM Tuned to Benchmark

Crowdsourced

Examples

- LM Arena
- HALIE



Trade Offs

Plus

- Scale
- Easy 1st impression

Minus

- Do questions align with use case?
Do results align with use case?











Observations from HALIE:

A Closer Look at Human-LM Interactions in Information-Seeking Contexts

Crowdsourced (Cont'd)

LM Arena Categories e.g.

- Text
- WebDev
- Vision
- Search
- Co Pilot
- Text to Image

✍ Text View →			
Rank (UB) ↑	Model ↑	Score ↑	Votes ↑
1	 gemini-2.5-pro-preview-05-06	1446	9,503
1	 o3-2025-04-16	1442	13,133
3	 chatgpt-4o-latest-20250326	1429	17,656
3	 gpt-4.5-preview-2025-02-27	1424	15,271
3	 gemini-2.5-flash-preview-05-...	1418	8,669
4	 claude-opus-4-20250514	1414	7,729
7	 gemini-2.5-flash-preview-04-...	1400	12,720
7	 gpt-4.1-2025-04-14	1399	11,773
7	 grok-3-preview-02-24	1397	19,977
7	 claude-sonnet-4-20250514	1390	6,384

Crowdsourced (Cont'd)

Trade Offs

Plus

- Scale
- Easy first impression

Minus

- Do questions & use case align?

 **Conversational View**

1 / 200


Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

LLM As a Judge & EVALS

Examples

**Judging LLM-as-a-Judge
with MT-Bench and Chatbot Arena**




 [openai / evals](#) Public


Vertex AI: Gemini Evaluations Playbook

LLM-as-a-judge on Amazon Bedrock Model Evaluation




 [anthropics / evals](#) Public



 [langchain-ai / auto-evaluator](#) Public



 [SalesforceAIResearch / FaithEval](#) Public

Trade Offs

Plus

- Scale
- Vary Granularity
- Cross Model

Minus

- Cost
- Lack of Scale

Human Expert

Key Characteristics

- Gold Standard Answers
- Grading Scale

Trade Offs

Plus

- User Buy In
- High Quality
- Construct Core Team

Minus

- Time Consuming, Relevant SMEs
- Lack of Scale

3. Better Practices

Better Practices

- *Domains:* Select carefully, Align evaluation approach
- *Intuitions:* Grow LLM intuitions broadly across organizations
- *Early:* Evaluate early. Consider hands on vendor workshops
- *Alignment:* Align Evaluation Approaches and Use Cases
- *LLM as Judge:* Vary Granularities
- *Human Experts:* Gold Standards, Common Criteria, Key Metrics

Q&A

- *Domains:* Select carefully, Align evaluation approach
- *Intuitions:* Grow LLM intuitions broadly across organizations
- *Early:* Evaluate early. Consider hands on vendor workshops
- *Alignment:* Align Evaluation Approaches and Use Cases
- *LLM as Judge:* Vary Granularities
- *Human Experts:* Gold Standards, Common Criteria, Key Metrics