september 7, 2023

HITL Is All You Need

An In-Depth View of the Significant Role of Embedded Human Experts in the LLM Application Development Life Cycle.

prepared for

Voice & AI 2023

presented by

Sam Shamsan, Head of Data Science
Michel Lopez, CEO

about e2f

e2f helps people and machines communicate naturally regardless of language, content, or culture.

With expertise in data science – and deep roots providing agile translation in 200+ languages and dialects – e2f uniquely provides high-quality linguistic datasets of multilingual speech, text, annotation, and quality data required to help machines understand people.

Lately, e2f has sharpened Generative AI expertise through LLM projects for some of the largest and most innovative companies in the world.

fast facts

Established in 2004 in San Jose

Privately held

Fully remote, follow-the-sun global operations

Code of Ethics committed to the wellbeing of our human resources,

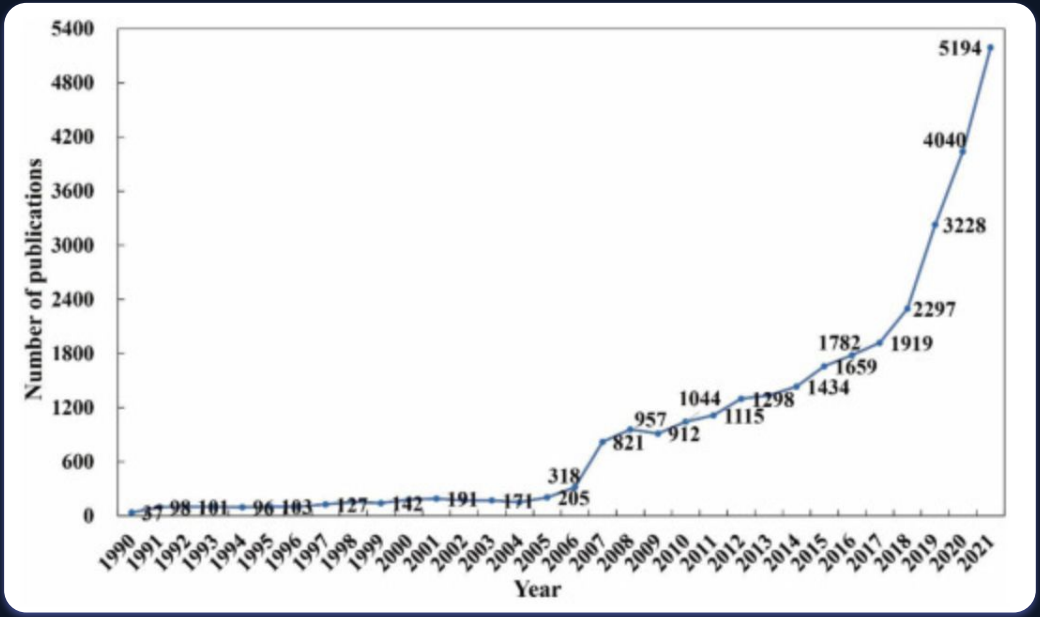
and to ensuring we deliver balanced, unbiased data

presentation beginning

HITL Is All You Need

An In-Depth View of the Significant Role of Embedded Human Experts in the LLM Application Development Life Cycle.

Publication trends of NLP

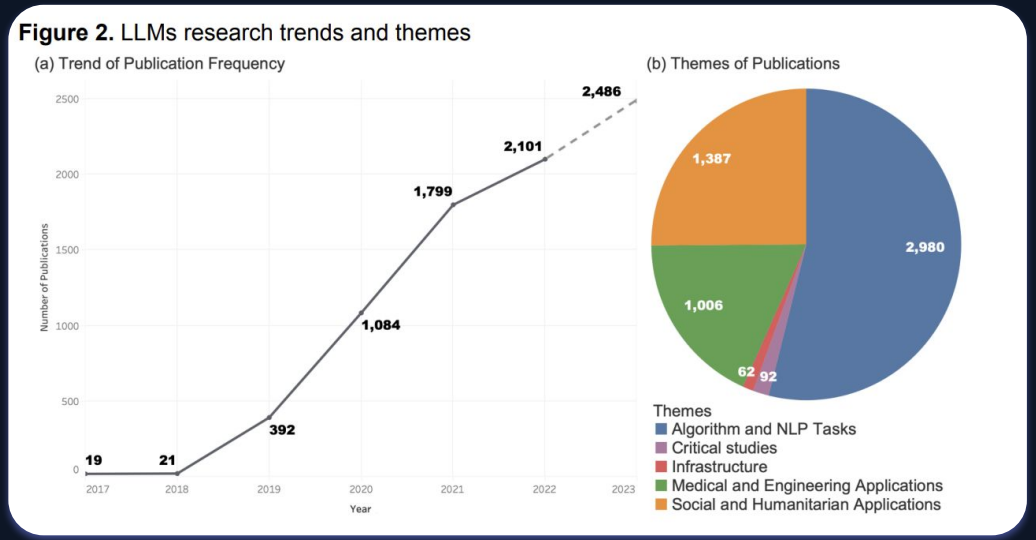


- A total of 31,485 NLP papers were analyzed, revealing a growing trend in NLP research from 1999 to 2021, with three distinct stages of development:
  - slow growth (1999–2005)
  - steady growth (2006–2016)
  - fast growth (2017–2021).

About 53% of the literature was published between 2017 and 2021, indicating that NLP is an increasingly active field of research.

Source: Vision, status, and research topics of Natural Language Processing

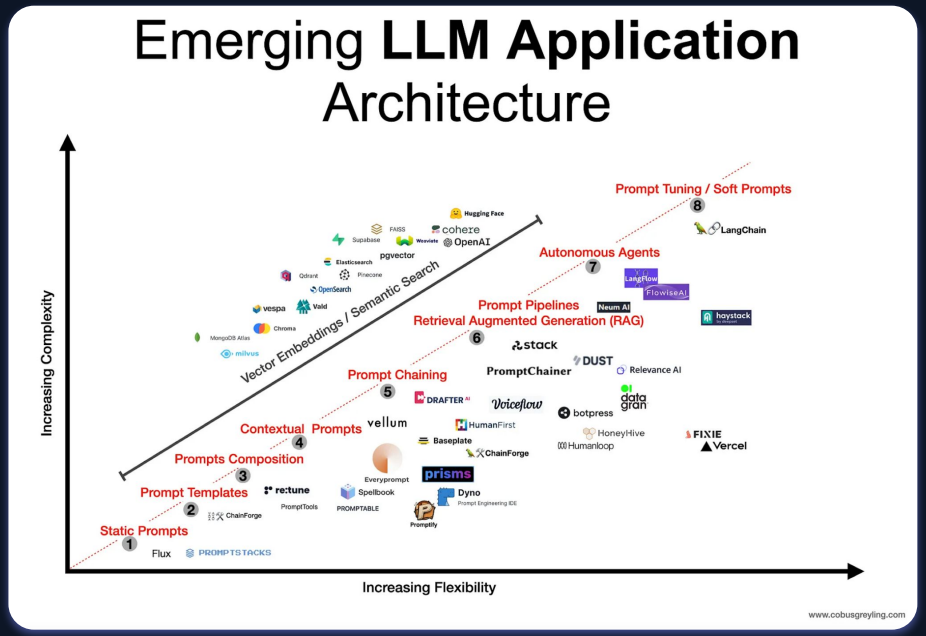
LLMs research trends and themes



- Advanced Models
- Diverse Research Themes: Research in LLMs spans across diverse themes including:
  - algorithms and NLP tasks (54%)
  - social and humanitarian applications (25%)
  - medical and engineering applications (18%),
  - critical studies, and infrastructures (each <2%).

Source: A Bibliometric Review of Large Language Models Research from 2017 to 2023 Lizhou Fan 1*, Lingyao Li 1, Zihui Ma 2, Sanggyu Lee 2, Huizi Yu 1, Libby Hemphill 1

The Rapid Growth Of LLMs



- Applications are increasing
- Last month, White house announced a responsible AI with big giant tech like: Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI.

Limitation of LLMs

	Reasoning	Knowledge	Conversation	Creativity	Personality	Storytelling	Empathy
LaMDA	0.84	0.69	1.0	0.53	0.85	0.58	0.94
ChatGPT	0.74	0.82	0.92	0.77	0.72	0.74	0.7
GPT-3	0.87	0.86	0.72	0.75	0.66	0.72	0.49
T5	0.7	0.6	0.19	0.51	0.1	0.36	0.04
PaLM	0.76	0.56	0.21	0.24	0.21	0.18	0.17
BLOOM	0.48	0.35	0.29	0.36	0.15	0.18	0.24
Turing-NLG	0.56	0.42	0.29	0.07	0.16	0.07	0.0

How popular LLMs score along human cognitive skills (source: semantic embedding analysis of ca. 400k AI-related online texts since 2021)

- **Outdated Responses:** LLMs are limited by the data they were trained on and may produce outdated responses if not frequently updated and retrained.
- **Lack of Domain-Specific Knowledge:** Generic LLMs lack the domain-specific knowledge required to provide contextually specific responses.
- **High Training Costs:** The large-scale nature of LLMs results in costly and resource-intensive training requirements for frequent knowledge updates.
- **Hallucinations:** Even when fine-tuned, LLMs can generate factually incorrect responses not aligned with the provided data.

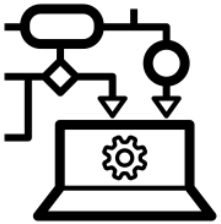
Human-Centric Approach

LLM Research Directions

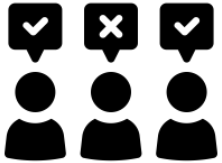
- ➔ Reducing Hallucination
- ➔ In context Learning
- ➔ LLMs for non english languages
- ➔ Multi-Modality
- ➔ New architecture
- ➔ Better hardware
- ➔ Learning personal preferences

- Can’t be solved with only technical solution.
- More investment in non-technical and human domain expert is needed.

Closing Gap



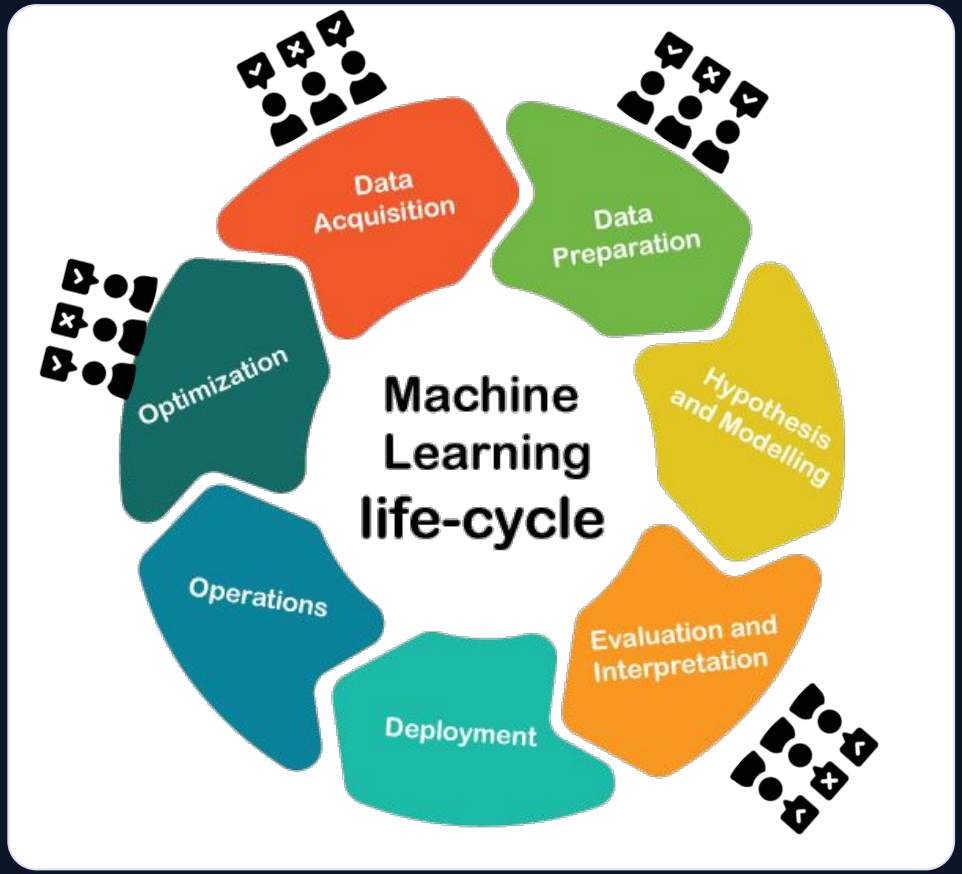
LLM Model



HITL

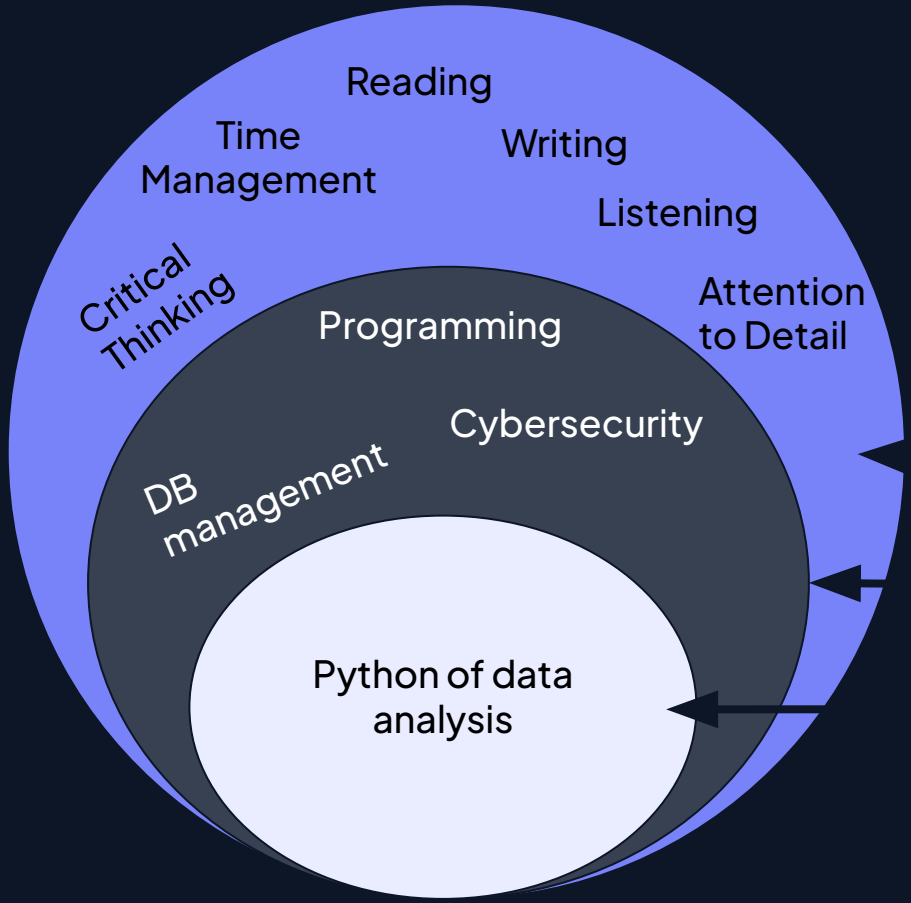
HITL: The overlooked factor

- Outdated Responses:
- Lack of Domain-Specific Knowledge
- High Training Costs:
- Hallucinations:



Engagement of Experts in ML Lifecycle

- Human engagement is crucial in two different stages of the ML lifecycle.
- **Pretraining:** The first stage is data preparation, which includes wrangling, accumulating, classifying, cleaning, and validating data before training, fine-tuning, or customizing the model.
- **Post training:** pre-deploying the model:
  - black boxes with a minimal degree of explainability.
  - Human experts vet the model output and evaluate.



Vetted Experts: Why they matter?

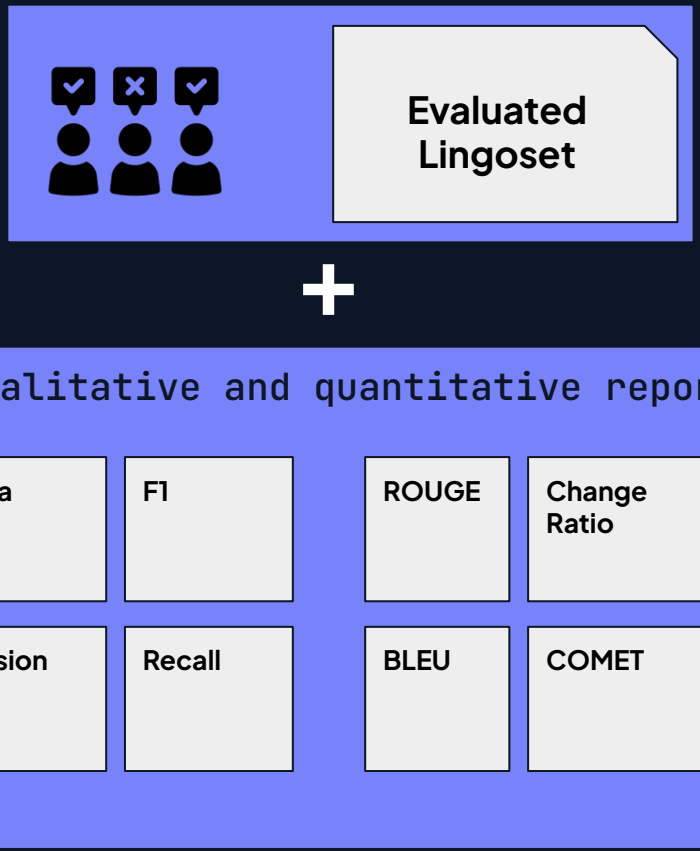
- Agile access to a pool of trained experts

General Test

Domain specific Test

Task specific Test

- Experts tested in generic tasks like reading, writing, and more domain-specific tasks like math, coding, insurance, medical, etc.



What should you expect?

- Evaluated Lingosets with all data point we evaluated per each category. You can have a third eye on it.
- Full quantitative report of the evaluation process:
  - How many data points?
  - What categories?
  - Crucial issues.
  - Data profiling.
- Certification that indicate the type of test and whether you passed or failed.

Standardize these tests.

Why You MUST evaluate LLMs App?

Relevance and Accuracy:

- ➔ Factuality
- ➔ Completeness
- ➔ Relevance
- ➔ Coherence
- ➔ Hallucination
- ➔ Helpfulness
- ➔ Domain-Specific Accuracy
- ➔ Formating
- ➔ Informativeness

Safety and Ethical Considerations:

- ➔ Harmfulness
- ➔ Kid's Safe
- ➔ Offensiveness
- ➔ Fairness and Bias Detection
- ➔ Ethical Considerations
- ➔ Resilience to Adversarial Inputs

A Reality Check on the Knowledge of LLMs

Understanding LLMs Limitations:

➔ Are LLMs as informed as we assume?

➔ Do they generate non-factual information?

- A Head-to-Toe benchmarking system with 18,000 question-answer pairs was used to evaluate 14 publicly available LLMs.
- Even the most advanced LLMs struggle with representing factual knowledge.

Model	All		Open		Movie		Book		Academics	
	A_{LM}	H_{LM}	A_{LM}	H_{LM}	A_{LM}	H_{LM}	A_{LM}	H_{LM}	A_{LM}	H_{LM}
ChatGPT	20.3	14.1	22.1	14.8	34.7	13.3	16.9	24.9	3.0	1.9
LLaMA (33B)	18.2	80.0	19.0	79.1	28.7	70.1	15.8	82.9	7.1	90.3

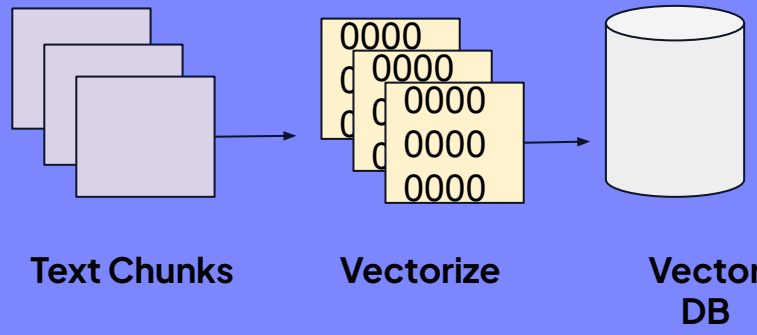
Table 3: Overall accuracy of the best two LLMs is only ~20% on Head-to-Tail. All numbers are in percentage (%).

[How Knowledgeable are Large Language Models (LLM)?]

in context learning

Factuality, Hallucination and Retrieval-Augmented Generation (RAG)

- The most common issue is the hallucination of non-factual answers.
- One of the most effective ways to reduce hallucinations is by retrieving useful, factual information and feed them to the prompt as context Using Standard vector search method.



Vector Search

Add to prompt as context

Why MUST you evaluate LLM Apps?

Model Functionality and Robustness:

- ➔ Adaptability
- ➔ Sensitivity to Feedback
- ➔ **Multi-Turn Conversation Capability**
- ➔ Response Latency (Speed)
- ➔ Model Robustness
- ➔ **Contextual Understanding**
- ➔ Argumentation Skills
- ➔ Emotional Intelligence
- ➔ **Cross-Lingual Understanding**

User Experience:

- ➔ **Naturalness**
- ➔ **Engagement**
- ➔ Personalization
- ➔ User Satisfaction
- ➔ **Consistency**
- ➔ Topic Transition
- ➔ Creativity

Why MUST you evaluate LLM Apps?

Data Privacy and Security:

- ➔ Data Protection
- ➔ Data Encryption
- ➔ Data Anonymization

Interoperability and Integration:

- ➔ API Design and Usability
- ➔ Compatibility
- ➔ Scalability

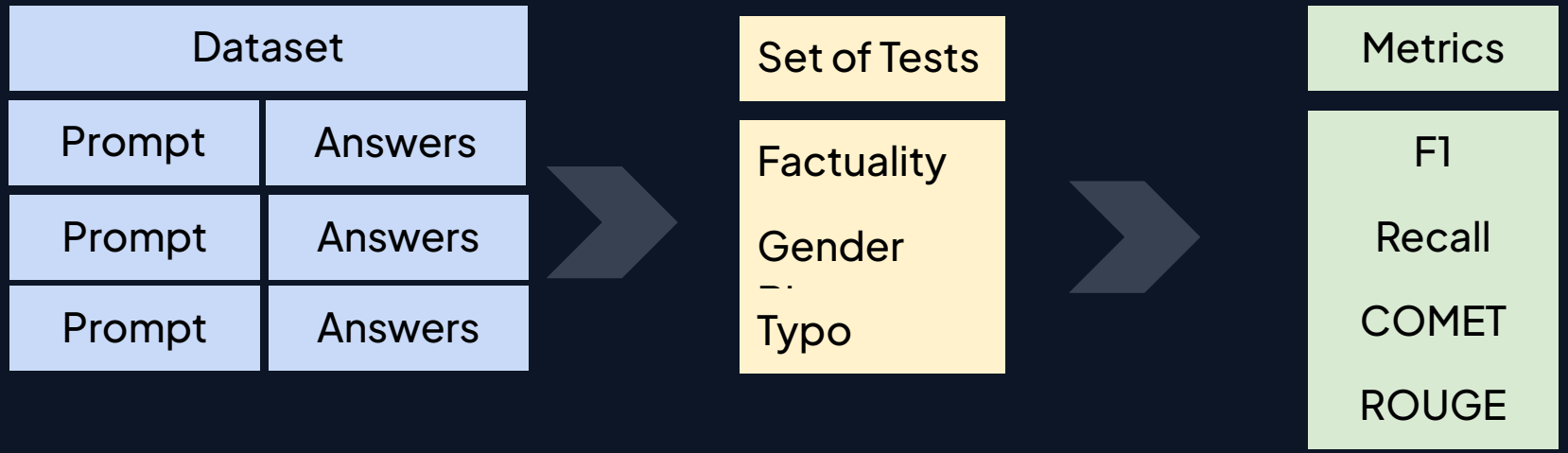
Policy Compliance and Legal Considerations:

- ➔ Complying with AI Policies
- ➔ Copyright
- ➔ Jailbreaking
- ➔ Legal and Regulatory Compliance

Performance Optimization:

- ➔ Resource Efficiency
- ➔ Cost-Effectiveness

LLMs Holistic Evaluation



- ➔ Question-answering based knowledge testing.
- ➔ One-turn style vs multi-turn dialogue.
- ➔ Well-structured benchmark allows an objective assessment.
- ➔ Remove human bias.

Issues with the current standards

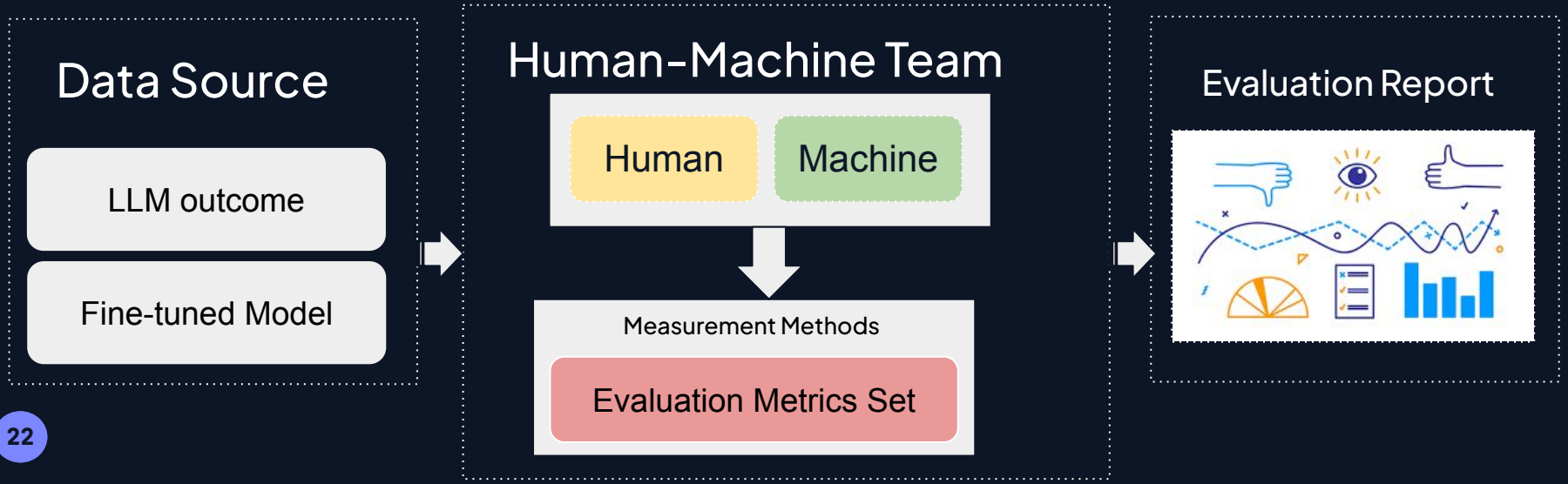
- ➔ Most of the tests follow one-turn style.
- ➔ Simple to manipulate benchmarks.
- ➔ Test dataset are compromised. Test cases are mixed in the training dataset.
- ➔ Open ended question metrics are not relevant such as subject and objective human grading.

Researcher are currently on aligning the LLM with High human evaluation.

Evaluating The HITL evaluation

Human-in-the-Loop Evaluation:

➔ Human drive and analysis the outcome of the automated metrics.In certain tasks, humans exclusively handle the entire evaluation process.



Job Type Based Metrics

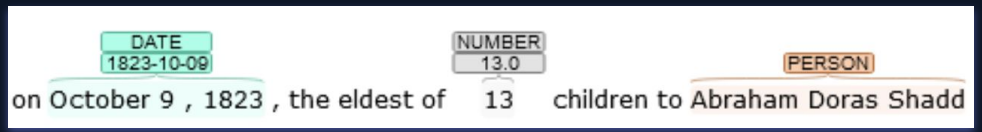
annotation

Binary

Ranking

Scale

Tagging



content editing

Content Enhancement

Content Correction

Content Translation

Content Creation

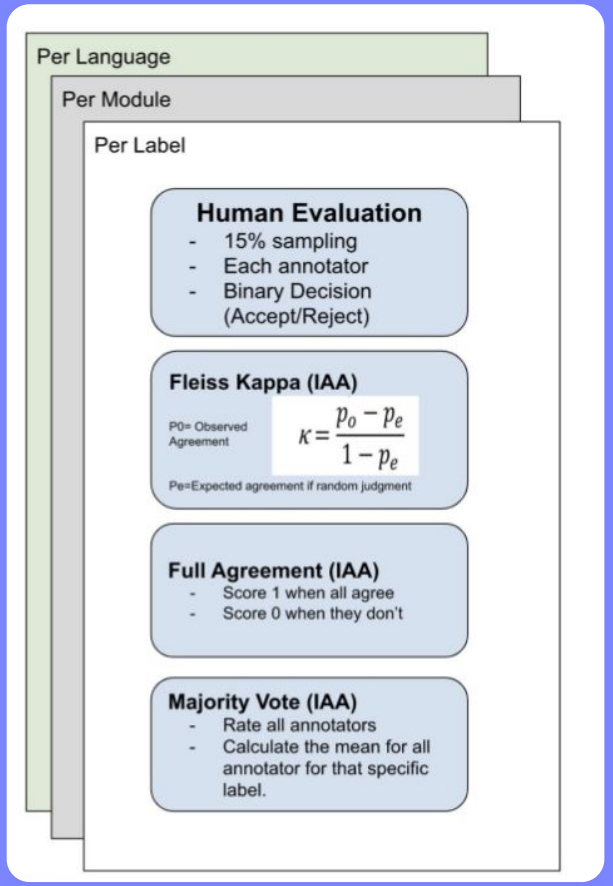
Inter Annotator Agreement

Definition: A statistical measure used to determine consistency or agreement between two or more annotators who label a set of items.
Importance:

- **Consistency: Ensures that different annotators interpret the annotation guidelines in the same way.**
- **Quality Control: Highlights potential ambiguities or issues in the guidelines.**
- **Reliability: Confirms that the annotations are reproducible and not subject to individual biases.**

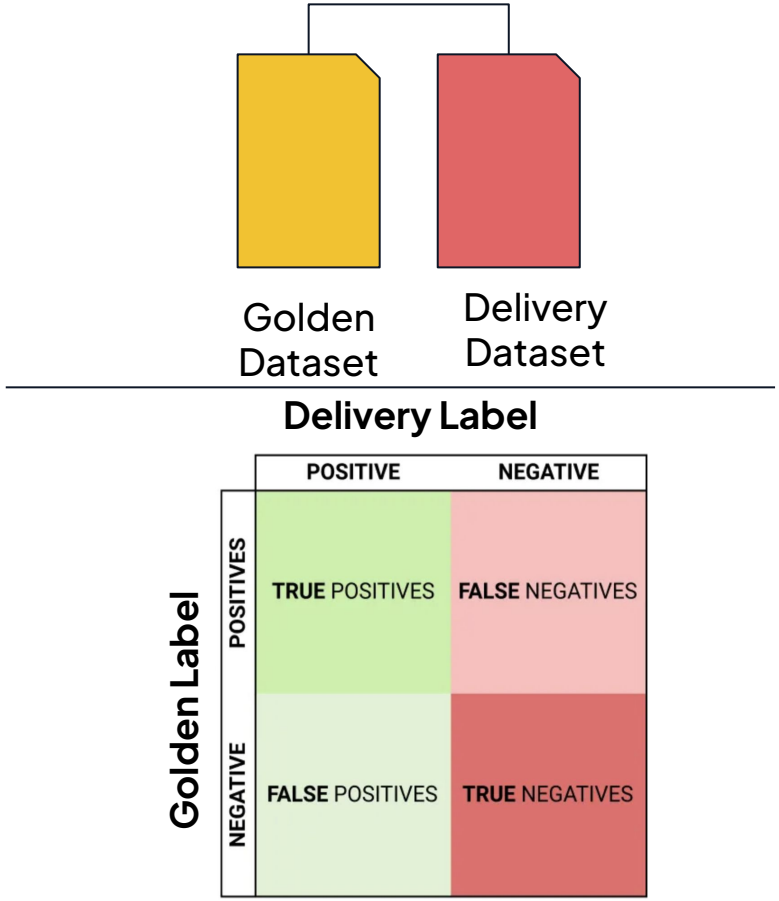
Steps to Improve IAA:

I. **Clear Guidelines**: Ensure annotation guidelines are comprehensive and unambiguous.
II. **Training Sessions**: Conduct training sessions and workshops for annotators.
III. **Regular Feedback**: Provide feedback and address any queries or concerns annotators might have.
IV. **Iterative Process**: Continuously refine the guidelines based on the feedback and challenges faced.



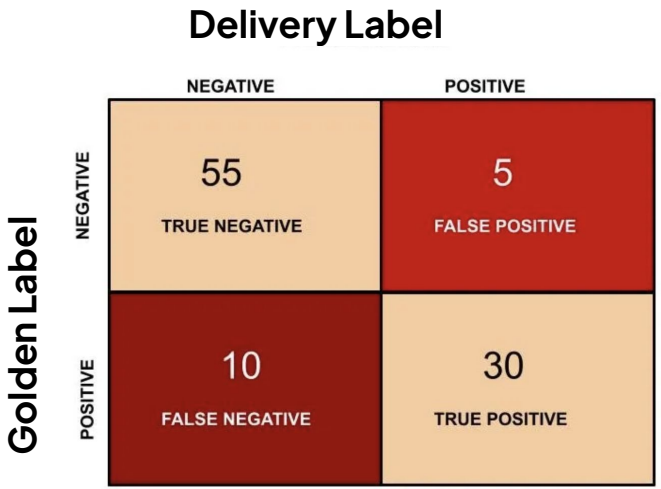
Classification Metrics

- These Metrics require golden dataset.
- Importance:
  - Serves as a "source of truth" for model evaluation.
  - Helps in calculating the classification metrics to estimate the model shortcomings.
  - Essential for benchmarking and comparison across different models.
- Challenge: It's not consistently accessible and requires time to develop for every delivery.
- Metrics:
  - Accuracy
  - Precision
  - Recall
  - F1 Score



Accuracy

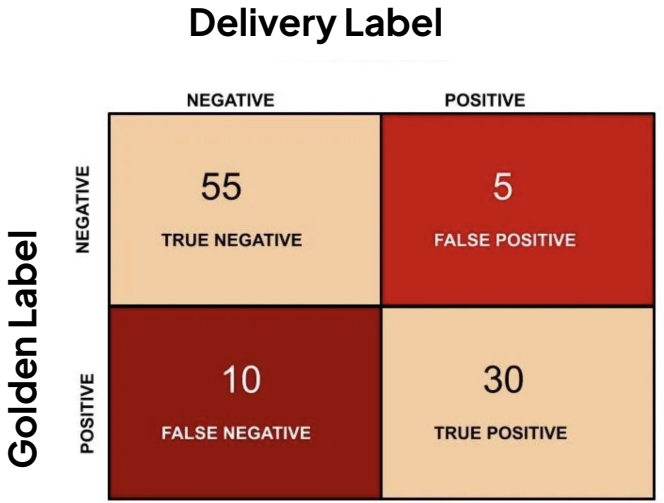
- Accuracy represents the number of correctly annotated data instances over the total number of data instances.
- Accuracy = (55 + 30)/(55 + 5 + 30 + 10) = 0.85
- Not a good measure if the dataset is not balanced



$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Precision

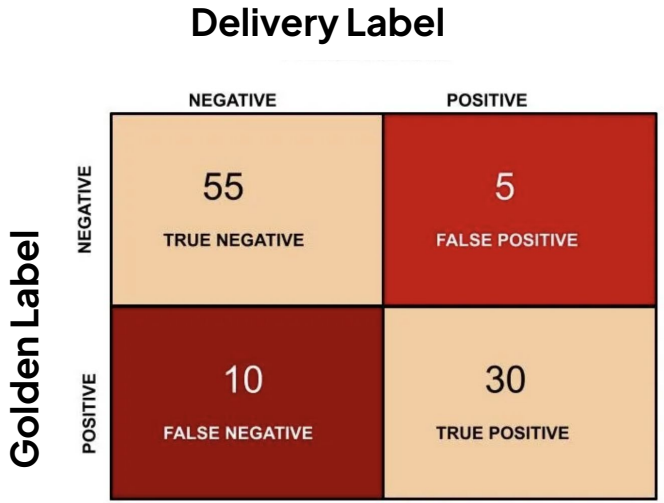
- ➔ Definition: Proportion of true positive annotated among all positive annotated.
- ➔ Importance: Indicates how many of the annotated positives are actually positive.
- ➔ precision = 30/(30+ 5) = 0.857



$$Precision = \frac{TP}{TP + FP}$$

Recall

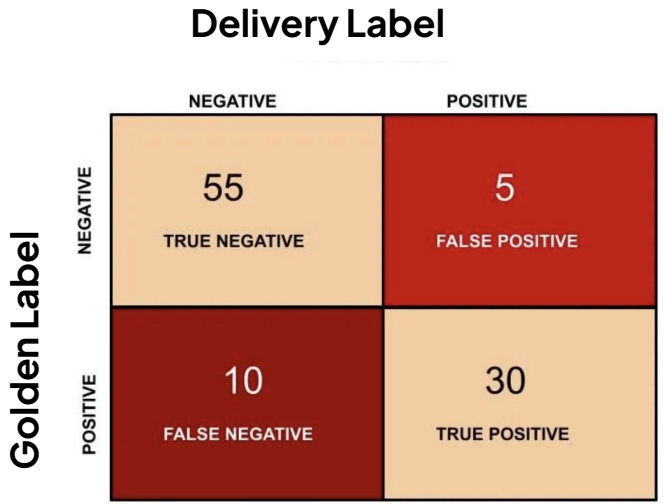
- ➔ Definition: Proportion of true positive annotations among all actual positives.
- ➔ Importance: Indicates how many of the golden positives were captured by the annotation.
- ➔ Recall = 30/(30+ 10) = 0.75



$$Recall = \frac{TP}{TP + FN}$$

F1 Score

- ➔ Definition: The harmonic mean of Precision and Recall, providing a balance between them.
- ➔ Importance: Helps evaluate annotation when class distributions are uneven.
- ➔ F1 Score = 2* (0.857 * 0.75)/(0.857 + 0.75) = 0.799

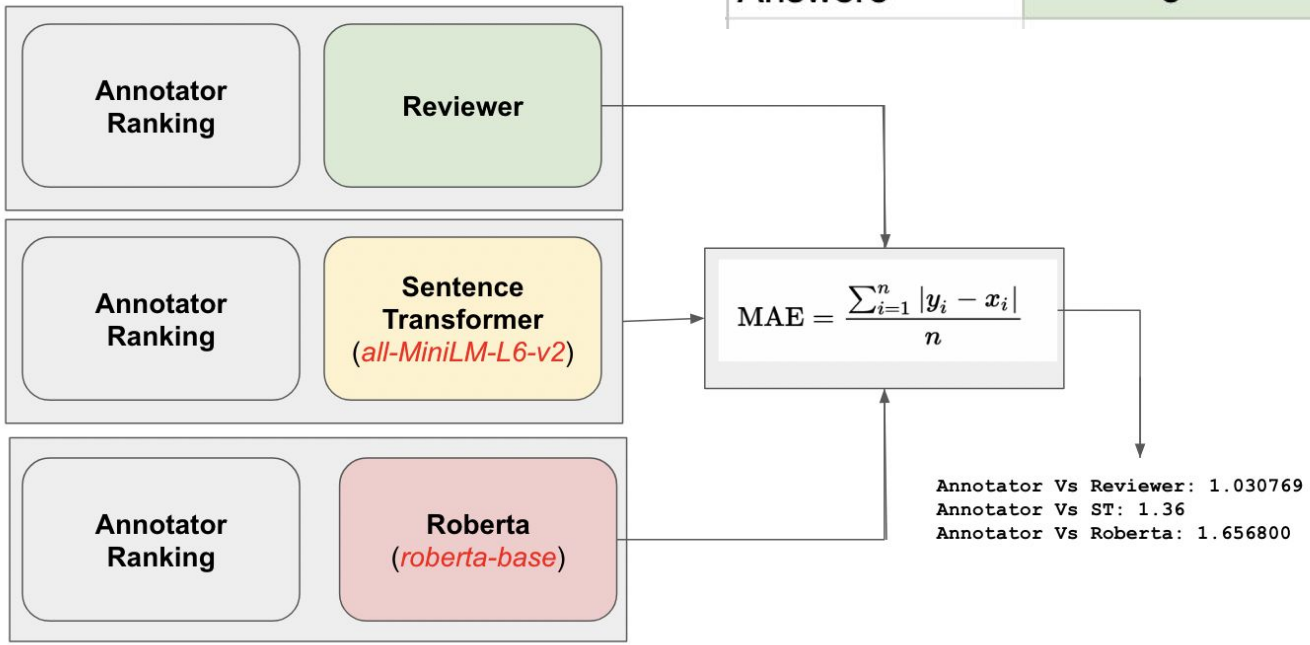


$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Ranking Metrics

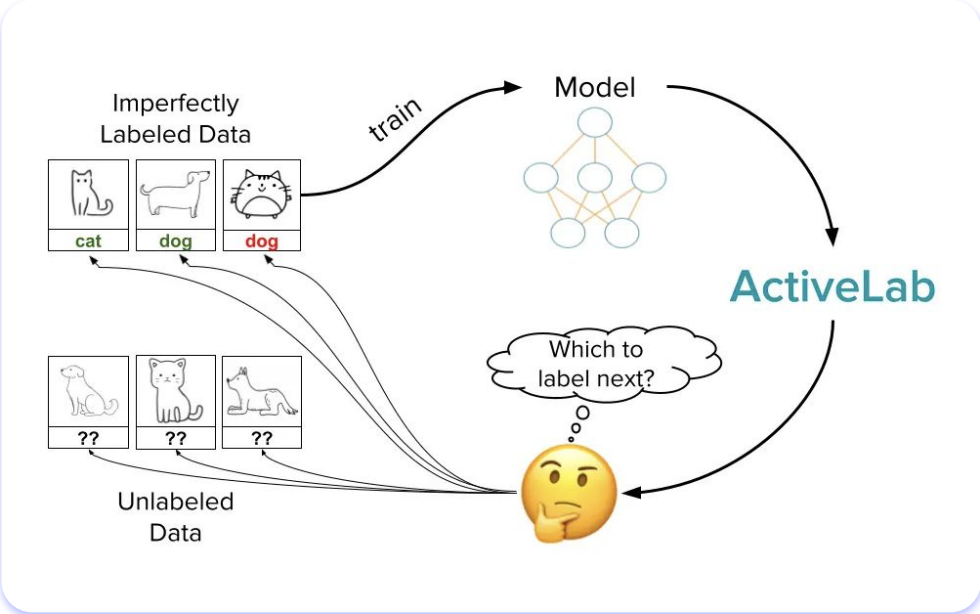
Answers	Ranking 1	Ranking 2
Answer1	3	4
Answer2	2	3
Answer3	4	1
Answer4	1	2
Answer5	5	1

Ranking Evaluation - Calculate MAE



Confident & Active Learning

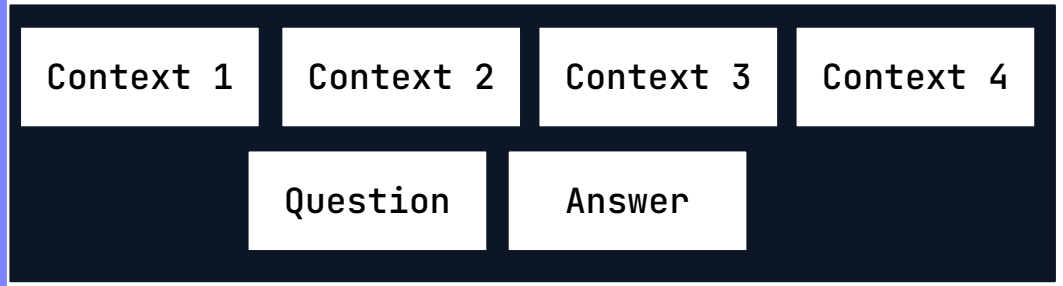
- ➔ Machine learning approach to detect potential noise.
- ➔ We can use it to measure the accuracy of each annotators.
- ➔ Another use case for it is to build machine learning based annotation.

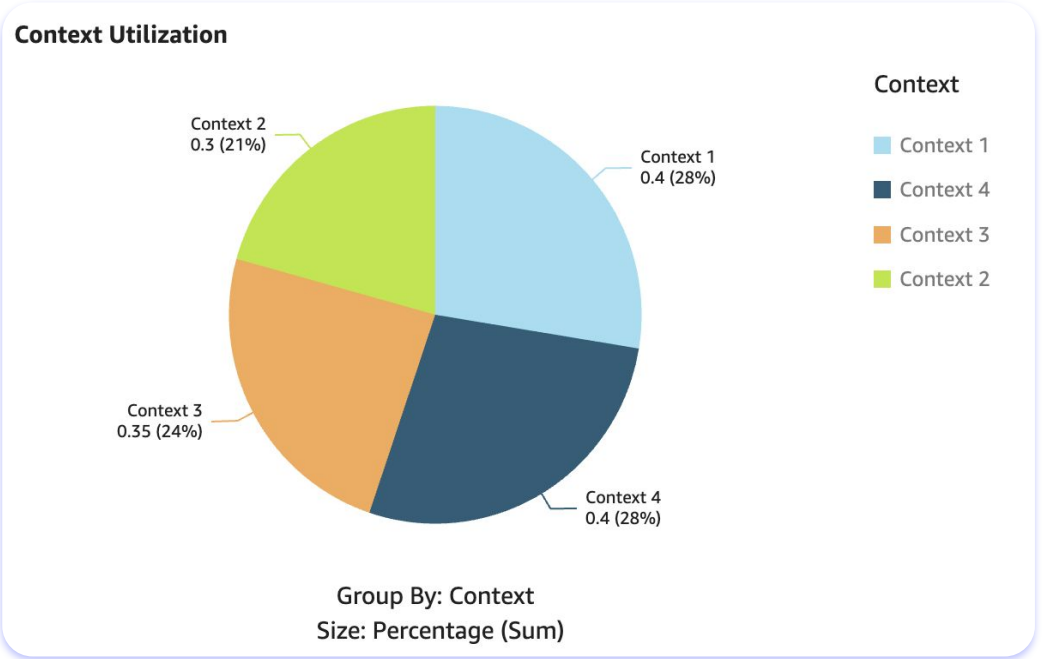


content editing

Context utilization & relevance

- ➔ How often context is used.
- ➔ Use of context vs use of external resources
- ➔ How often context is not usable at all.
- ➔ Help the client refine their context data generation.

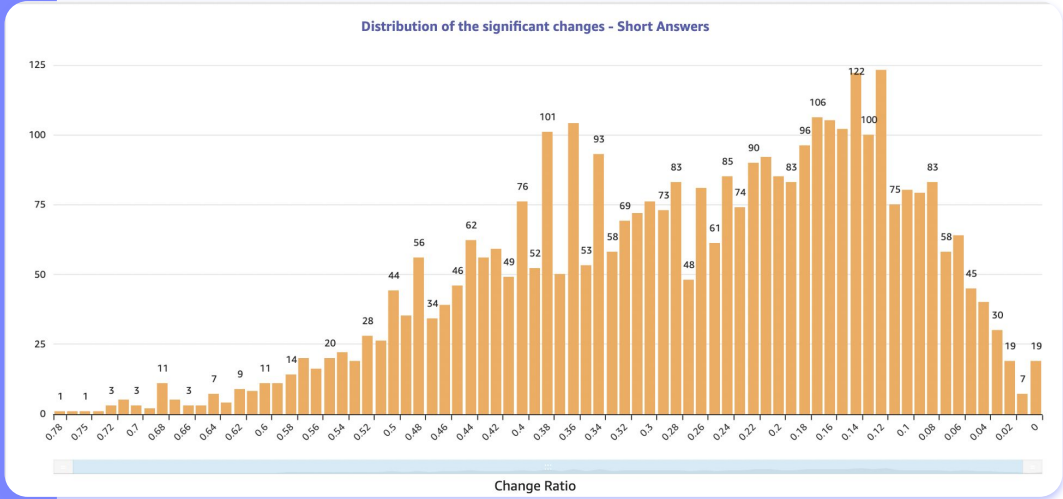




content editing

Significant changes

- ➔ Between Annotator and Reviewer (benefit of 2nd pass).
- ➔ Between client provided source and final result we deliver (our contribution)
- ➔ Use a combination of changes measurement such as sentiments, sentence structure, readability and edit distances.



Editing

Formatting: paragraphs, bullet points, introduction, conclusion

Improved long answer	Improved long answer (formatted)
Electric fireplaces offer homeowners the opportunity to enjoy the ambiance of a traditional fireplace without the hassle of installing a chimney. Although they don't generate significant heat, they utilize lights and mirrors to create the illusion of flickering flames, providing the visual warmth of a fireplace. With various design options available, electric fireplaces can be seamlessly integrated into any room's decor. They eliminate the need for fuel like wood or coal, reducing maintenance and cleaning requirements. Ultimately, electric fireplaces serve as an attractive centerpiece and offer a convenient and mess-free alternative for those seeking the cozy atmosphere of a fireplace.	Electric fireplaces offer homeowners the opportunity to enjoy the ambiance of a traditional fireplace without the hassle of installing a chimney. Although they don't generate significant heat, they utilize lights and mirrors to create the illusion of flickering flames, providing the visual warmth of a fireplace. Benefits of Electric Fireplaces: - Various design options available, electric fireplaces can be seamlessly integrated into any room's decor. - Eliminate the need for fuel like wood or coal, reducing maintenance and cleaning requirements.

chatgpt detector stat

Editing –ChatGPT Content Detection

ChatGPT Detector Stat

Label	Count
ChatGPT	191
Human	1,914

edited_long_answer	ChatGPT/Human
There are several ingredients that can be used as substitutes for beer in cooking or baking. For a light beer^ options include chicken broth^ ginger ale^ white grape juice^ or white wine^ while dark beer can be replaced with beef broth^ mushroom stock^ apple juice^ apple cider^ root beer^ or coke. It is important to replace the amount of beer called for in the recipe with an equal amount of the substitute [[1](https://www.thespruceeats.com/beer-substitute-1388879)]. Another possible substitute for beer is coffee grounds^ with 2 tablespoons per 1 cup (0.24 l) of brewed coffee used for every 6 ounces (0.23 kg) of beer [[2](https://eatdelights.com/cooking-with-beer-substitutes/)]. For those looking for a non-alcoholic replacement^ kombucha can be used in place of beer^ with lighte...	ChatGPT
Pregnancy typically lasts about 40 weeks^ from the first day of the last normal period to the time of delivery.[[1](https://www.womenshealth.gov/pregnancy/youre-pregnant-now-what/stages-pregnancy)] This is divided into three trimesters: the first trimester (weeks 1-12)^ the second trimester (week 13-28)^ and the third trimester (week 29-40). After 41 weeks^ a pregnancy is considered late-term^ and once it reaches 42 weeks or beyond^ it is called post-term. During late-term and post-term pregnancies^ the risk of certain health problems^ such as larger-than-average birth size^ increases^ which may necessitate induction of labor. [[2](https://www.mayoclinic.org/healthy-lifestyle/pregnancy-week-by-week/in-depth/overdue-...	Human

User with chatbot content

Label	user	Count
⊟ ChatGPT	DSAGU	2
	DSEUO	33
	DSGAA	28
	DSMBI	55
	DSNIB	57
	DSSEN	12
	DSSHC	4

Task specific metrics

word error rate WER

Calculate WER from

- Reference: golden set
- Hypothesis: original text to be evaluated

Threshold: <5%

```
WORD RECOGNITION PERFORMANCE

Percent Total Error       =    2.5%   ( 107)

Percent Correct           =   97.6%   (4161)

Percent Substitution      =    1.4%   (  60)
Percent Deletions         =    1.0%   (  42)
Percent Insertions        =    0.1%   (   5)
Percent Word Accuracy     =   97.5%
```

BLEU

Usage: Evaluates machine-generated text quality, commonly in translation and generation tasks.

Interpretation: Higher score indicates better similarity with reference texts (0 to 1 range). BLEU scores above 0.4–0.5 can indicate acceptable performance.

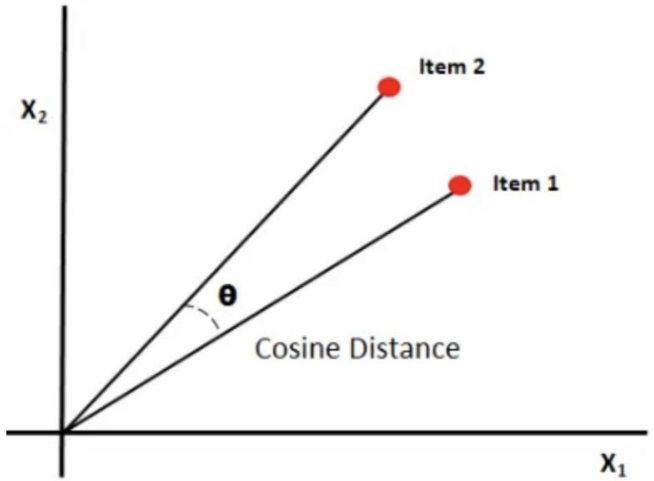
Limitations: Focuses on n-gram matches, lacks semantic context, insensitive to synonyms.

Shortcomings: Low scores don't always imply poor quality.

Complementary: Combine with other metrics for comprehensive assessment.

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

cosine similarity



Usage: Measures similarity between vectors in high-dimensional space, useful for text and content comparisons.

Interpretation: Higher score means greater similarity (range -1 to 1).

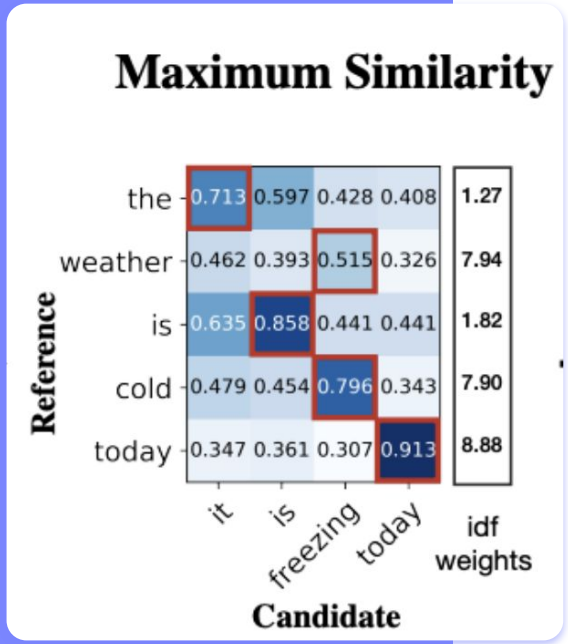
Limitations: Doesn't capture semantic context nor the negative data.

Normalization: Often used to standardize vectors.

Formula:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

cosine similarity



Usage: Metric for evaluating generated text quality using contextual embeddings from models like BERT.

Strength: Focuses on contextual understanding, sentence-level comparison, and reference embeddings.

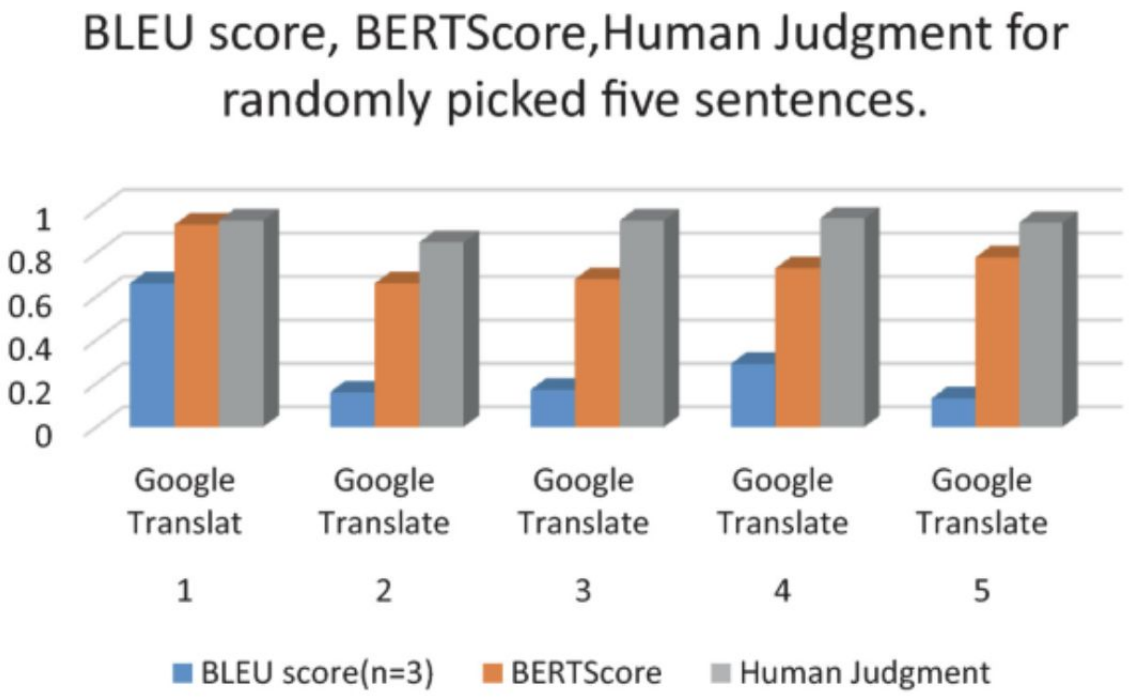
Interpretation: Higher score means greater similarity (range -1 to 1). Ideal for text generation, summarization, machine translation, etc.

Calculation:

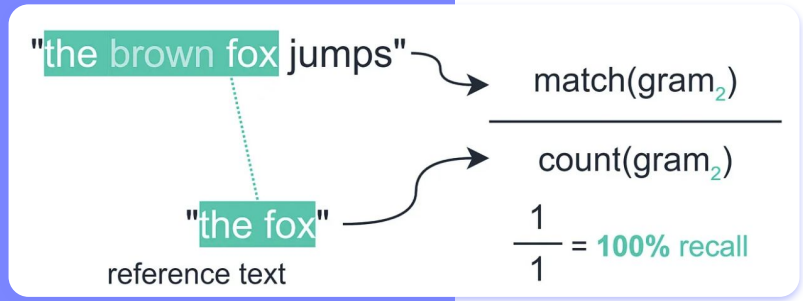
- Measure cosine similarity between reference and candidate embeddings.
- Aggregate sentence similarities for overall BERTScore.

Formula:

BLEU, BERTScore, and human judgement



rouge



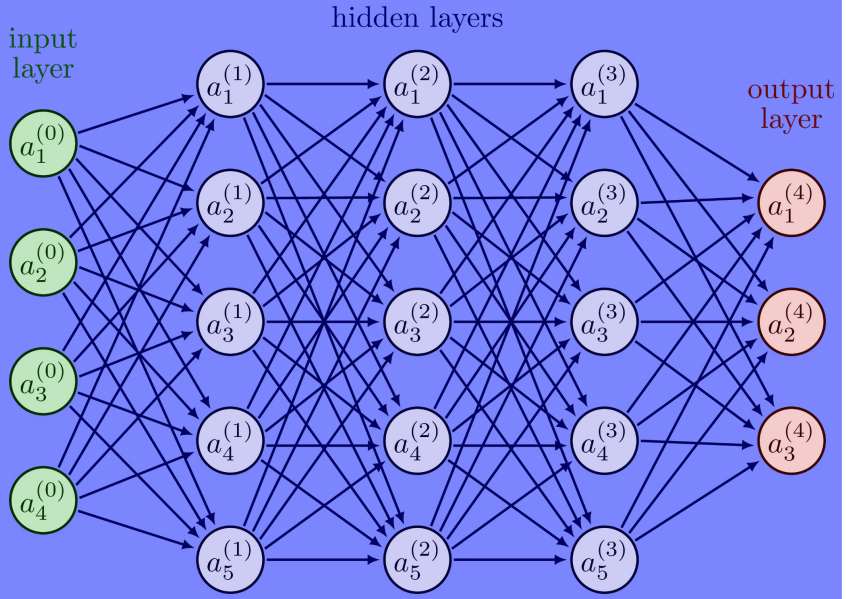
Usage: Metric used to assess machine-generated text quality, especially for summaries and translations.

Available in various versions: ROUGE-N, ROUGE-W, ROUGE-L, etc.

Calculation: Measures n-gram overlap between generated and reference texts.

Strength: ROUGE offers valuable insights into text quality based on n-gram overlap. Unlike BLEU that only match the n-gram. Works at the sentence level, suitable for diverse text comparisons. While BLEU is design for a chunk of text. Consider ROUGE for tasks where ***semantic context and sentence-level*** evaluation are crucial.

comet



Advantages over Traditional Metrics:

Semantic Similarity: Unlike earlier metrics (e.g., BLEU, chrF, METEOR), COMET goes beyond lexical-level features and captures semantic similarity between translations and human references.

Precision: COMET's incorporation of neural networks and human judgments enhances its precision in distinguishing between higher-performing and lower-performing translation systems.

Automated Evaluation: COMET's neural model automates the process of evaluating translation quality, reducing the need for manual annotation by human experts.

perplexity

- **Definition**: Perplexity calculates a language model's ability to predict a sample of text. Lower perplexity values indicate higher model performance.
- **Usage**: It's commonly used to assess the quality of language models in tasks like text generation, machine translation, and speech recognition.
- **Limitations**: Perplexity doesn't account for real-world meaning or context comprehension. A model with low perplexity might still produce text that lacks coherence and meaning.
- **Calculation**: Perplexity is calculated using the formula: Perplexity = 2^(Entropy), where Entropy measures the average uncertainty in predicting the next word based on the model's distribution.

HITL Is All You Need

Thank you

Stop by Booth 722 to learn more about our newest offering!