Team ID: 26

Quantifying Creativity in LLMs

Abeen Bhattacharya

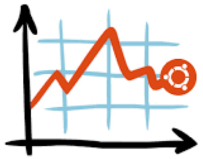
Motivation

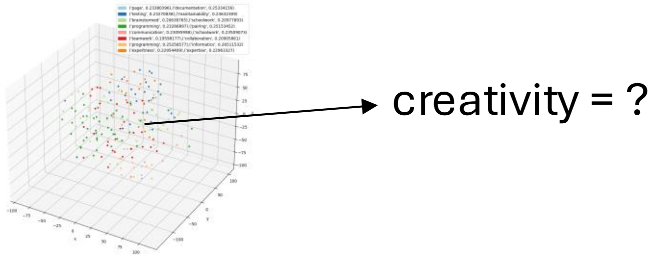
Team ID: 26

What is "Creativity"?
How/Can we quantify it?

Can we create an index using metrics to quantify "Creativity"?

What does "Creativity" mean to LLMs? Does it align with our metrics?





Quantifying Creativity

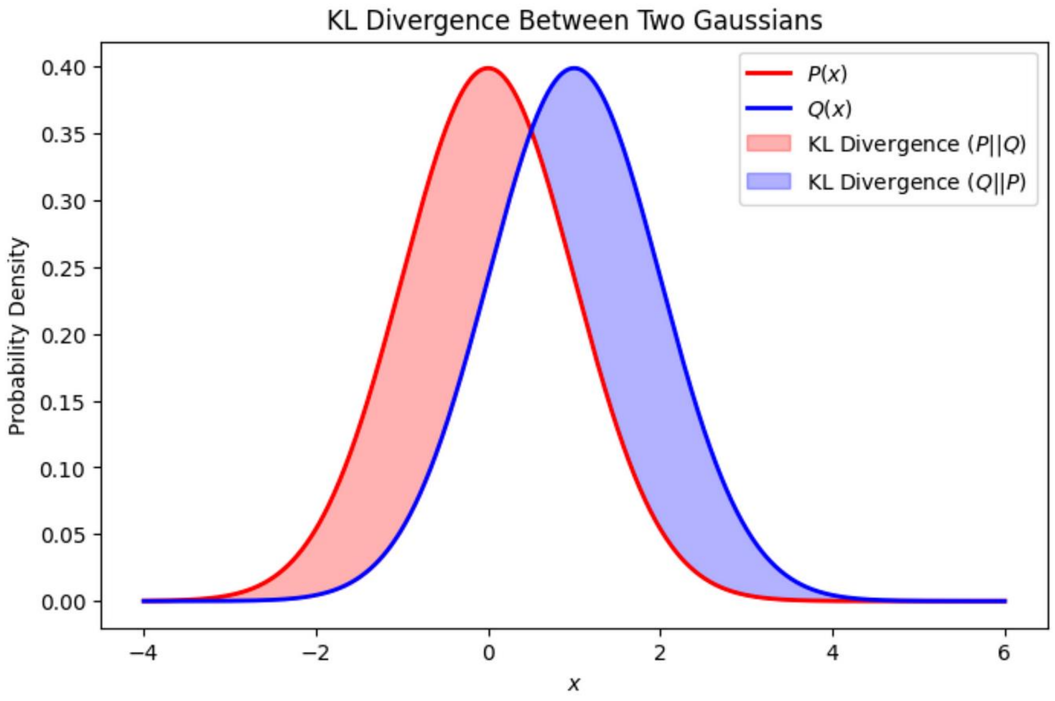
Team ID: 26

1. Novelty

Novelty is a cornerstone of creativity, reflecting the ability to produce original ideas or expressions that diverge from existing patterns. A creative text should introduce fresh perspectives or uncommon combinations of words, distinguishing it from routine or derivative outputs

Metric Used: Kullback-Leibler (KL) Divergence
Measures the difference between the token distribution of a text segment and a baseline corpus distribution

$$D_{\mathrm{KL}}\left(p(x)\,||\,q(x)\right) = \sum_{x \in X} p(x) ln \frac{p(x)}{q(x)}$$



Quantifying Creativity

Team ID: 26

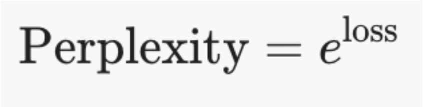
2. Coherence

Coherence ensures that a creative output is not just novel but also comprehensible and logically structured. Creativity without coherence can result in disjointed or nonsensical text, undermining its value. A creative text should maintain fluency and narrative flow, balancing originality with intelligibility, which is critical for effective communication of innovative ideas.

Metric Used: Perplexity (Inverted)

Quantifies how well a language model (GPT-2) predicts the text

Coherence is derived as 1 / (perplexity + 1e-6)



$$\text{Perplexity} = e^{\text{loss}}$$

where loss is the cross-entropy loss from GPT2LMHeadModel

Quantifying Creativity

Team ID: 26

3. Contextual Fit

Contextual fit assesses whether the creative output aligns with the given prompt or context, ensuring relevance. Creativity isn’t just about randomness; it requires purposeful deviation that still resonates with the intended theme or task. This metric ensures that novel ideas remain grounded in the creative intent

Metric Used: Cosine Similarity
Measures the semantic similarity between embeddings of the prompt and text segment, computed using Sentence Transformers

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||||\mathbf{B}||}$$

where A and B are the embedding vectors of the prompt & text respectively

Quantifying Creativity

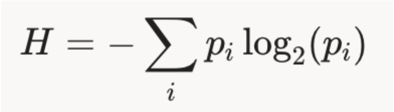
Team ID: 26

4. Syntactic Complexity

Syntactic complexity reflects the sophistication and variety of grammatical structures, which can indicate creative expression through intricate or unconventional phrasing. Creative writing often employs complex syntax to convey nuanced ideas or evoke specific effects, distinguishing it from simplistic or formulaic text.

Metric Used: Mean of Scaled Sub-Metrics

- **Average Sentence Length**: Words per sentence
- **Average Clause Count:** Clauses per sentence (approximated via dependency tags like advcl, ccomp)
- **Average Parse Tree Depth:** Maximum depth of dependency tree
- **POS Entropy:** Shannon entropy (H) of Parts of Speech tag distribution: Measures the level of uncertainty or randomness in the sequence of word types within a sentence



$$H = -\sum_{i} p_i \log_2(p_i)$$

Quantifying Creativity

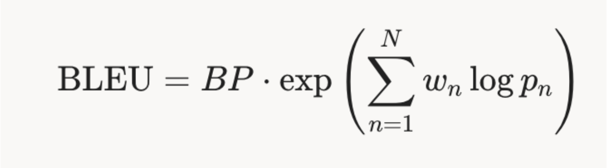
Team ID: 26

5. Lexical Diversity

Lexical diversity measures vocabulary richness and variety, a hallmark of creative language use. A creative text avoids repetition and employs a broad range of words, enhancing its expressiveness and originality. This metric captures the inventive use of language at the word level.

Metric Used: Scaled Weighted Mean of the following sub-metrics:

- **N-gram Diversity:** Ratio of unique n-grams to total n-grams (n=1, 2, 3)
- **Self-BLEU (Inverted):** Average BLEU score of each sentence against others, inverted as 10 * (1 - BLUE/ 100) .



$$\mathrm{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

(BP = brevity penalty, p_n = n-gram precision)

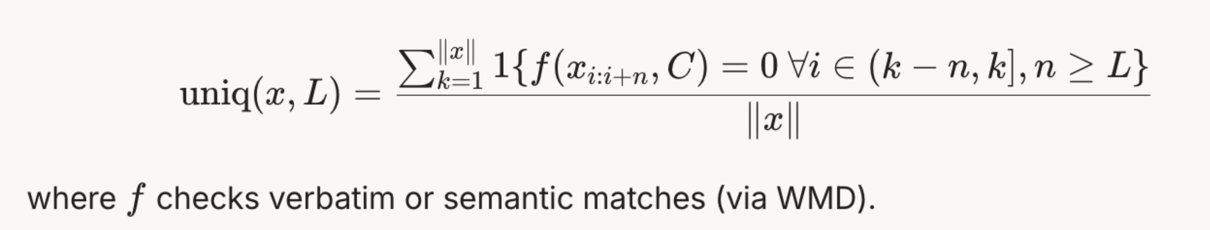
Quantifying Creativity

6. Lexical Novelty

Lexical novelty quantifies the uniqueness of phrasing by assessing how much of the text cannot be reconstructed from existing sources

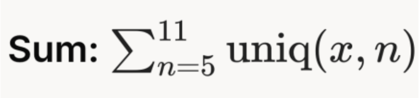
Metric Used: Sum of L-Uniqueness using **DJ Search**

- **L-Uniqueness**: Proportion of words not found in any n-gram (n ≥ L, L=5-11) in the baseline corpus, computed



$$\text{uniq}(x, L) = \frac{\sum_{k=1}^{\|x\|} 1\{f(x_{i:i+n}, C) = 0 \ \forall i \in (k-n, k], n \geq L\}}{\|x\|}$$

where f checks verbatim or semantic matches (via WMD).



Sum: $\sum_{n=5}^{11} \text{uniq}(x, n)$

Quantifying Creativity

Team ID: 26

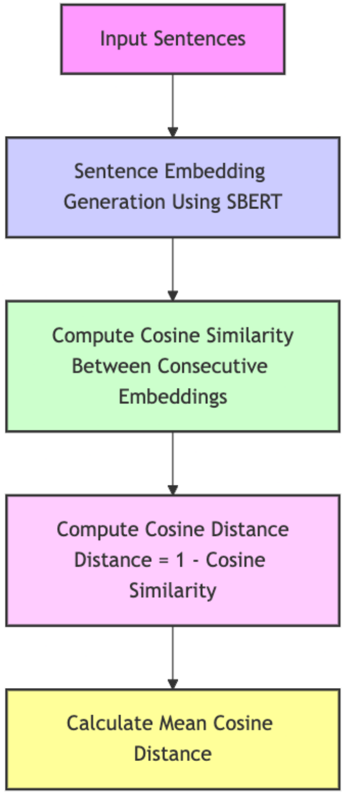
7. Surprise

Surprise measures unpredictability in text progression, a key creative trait that engages readers through unexpected twists or shifts. Creative works often defy conventional expectations.

Metric Used: Mean Cosine Distance between **consecutive sentence embeddings**

$$\frac{1}{n-1}\sum_{i=1}^{n-1}\text{Distance}(s_i, s_{i+1})$$

$$\text{Distance} = 1 - \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$



Input Sentences

Sentence Embedding Generation Using SBERT

Compute Cosine Similarity Between Consecutive Embeddings

Compute Cosine Distance Distance = 1 - Cosine Similarity

Calculate Mean Cosine Distance

Quantifying Creativity

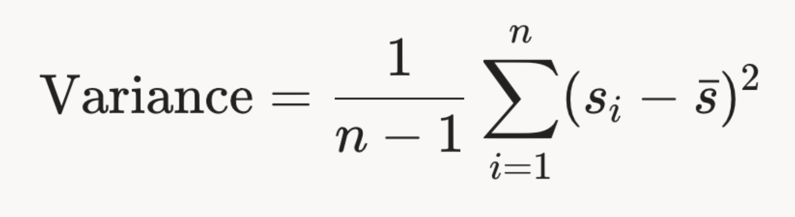
Team ID: 26

8. Emotional Expressiveness

Emotional expressiveness evaluates the intensity and variety of emotions conveyed, a critical aspect of creative writing that enhances impact and depth. Creative texts often evoke strong or diverse emotional responses, distinguishing them from flat or neutral outputs.

Metric Used: Variance of Sentiment Scores

VADER (Valence Aware Dictionary and sEntiment Reasoner) compound scores per sentence [−1,1]



$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(s_i - \bar{s})^2$$

where s_i is the sentiment score, $\bar{s}$ is the mean

Quantifying Creativity

Team ID: 26

9. Human-Likeness

Human-likeness ensures that creative output resembles natural human writing, balancing novelty with plausibility. Creativity should not devolve into incoherent randomness; this metric anchors the output in recognizable linguistic patterns, a practical constraint for usability.
Metric Used: Perplexity (Exponentiated Loss) / MAUVE

Why use both Coherence & Human-Likeness?:
Coherence ensures the creative output holds together as a meaningful whole, while **Human-Likeness** ensures it connects to human experience and expression. Both use GPT-2 perplexity, but **Coherence** inverts it to emphasize local sequence quality (sentence-level fluency), while **Human-Likeness** scales it exponentially to reflect global plausibility against a human-trained model
A text can fail one metric while passing the other, revealing distinct weaknesses:

- **High Coherence, Low Human-Likeness:** A logically structured but robotic text (e.g., "Step 1: Activate device. Step 2: Configure settings.") might score well on coherence but feel artificial, lacking creative flair.
- **High Human-Likeness, Low Coherence:** A conversational but disjointed rant (e.g., "Hey, cool stuff, oh wait, cats are great, um, what was I saying?") might sound human-like but fail to form a coherent narrative.

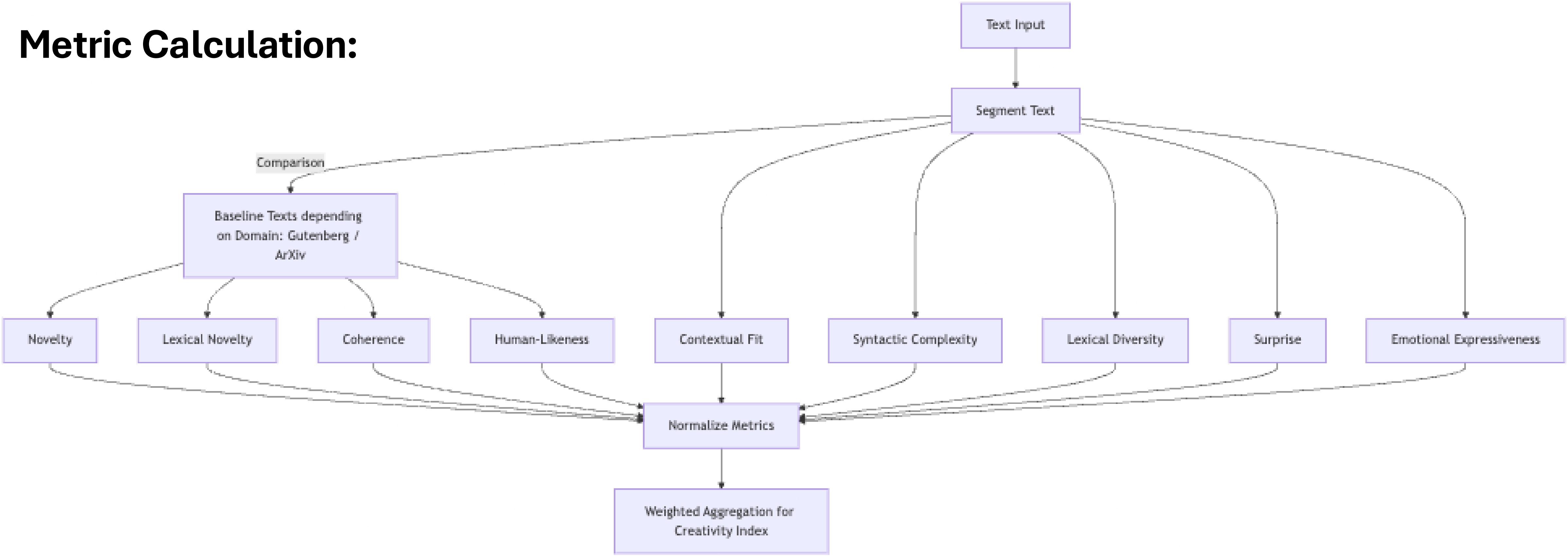
Defining the Baseline for Metric Calculation

Team ID: 26

Datasets for Baseline:

- Literary Domain: Gutenberg Corpus
- Technical Domain: arXiv Abstracts

Metric Calculation:



Defining the Creativity Index

Team ID: 26

Using all the metrics we defined, we can create a simple creativity index by:

$$\mathrm{CI} = \sum_{i=1}^{9} w_i \cdot \mathrm{norm}(m_i)$$

- m_i: The value of the i^{th} metric (e.g., Novelty, Coherence).
- norm(m_i): The normalized value of that metric.
- w_i: The weight for that metric
- The sum of weights (w_i) equals 1

The weights can be adjusted, depending on how important a particular metric is to you.

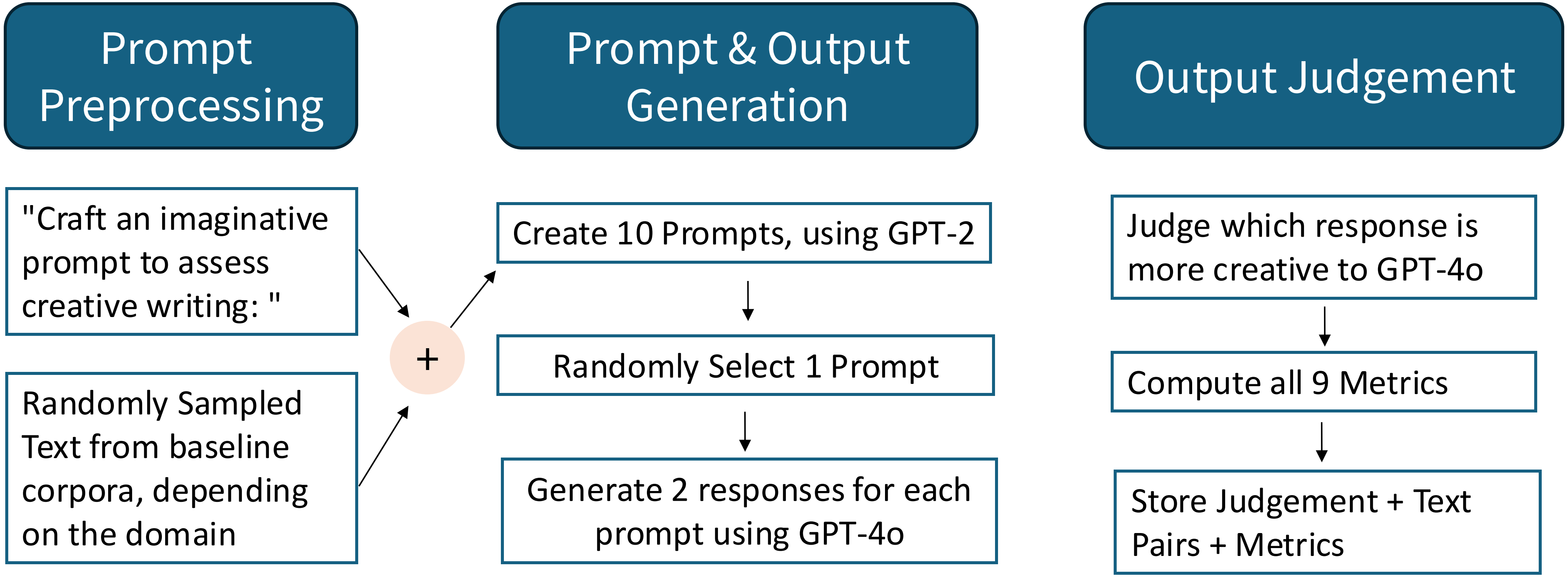
Now, that we have an index to measure creativity, we move on to:

- **Validate whether our index works?**
- **Understand what does "Creativity" mean to a Large Language Model?**
  - Create a dataset based on LLM's understanding of "Creativity"
  - Create ML models using the metrics as input to validate this "Creativity"

Understanding Creativity in LLMs

Team ID: 26

To validate our creativity index, with how an LLM views creativity
I created a **contrastive pair dataset, judged by GPT-4o** on which text of the pair is more "creative" according to it.



Prompt Preprocessing

"Craft an imaginative prompt to assess creative writing: "

+

Randomly Sampled Text from baseline corpora, depending on the domain

Prompt & Output Generation

Create 10 Prompts, using GPT-2

Randomly Select 1 Prompt

Generate 2 responses for each prompt using GPT-4o

Output Judgement

Judge which response is more creative to GPT-4o

Compute all 9 Metrics

Store Judgement + Text Pairs + Metrics

Different Models Used & Why

Team ID: 26

SimpleCreativityPredictor (Baseline Model)

- **Inputs**:
  - **Text Segments**: Processed into embeddings using **BERT**.
- **Processing**: Embeddings are processed by an **LSTM**.

Why?

- Tests whether **the text alone** (**without metrics**) can predict creativity, acting as a simpler baseline.
- Helps us understand if calculating metrics adds value

CompositeRegressor

- **Inputs**:
  - **Text Segments**: Processed into embeddings using **BERT**.
  - **Metric Features**: Creativity metrics fed into an **LSTM**.
- **Processing**: Embeddings and LSTM outputs are combined in a fusion layer.

Why?

- Blends raw text data with explicit metric values
- Starting point to see how well text and metrics together predict creativity.

TransformerCreativityAggregator

- **Inputs**:
  - **Text Segments**: Processed into embeddings using **BERT**.
  - **Metric Features**: Creativity metrics fed into a Transformer.
- **Processing**: Embeddings and Transformer outputs are combined in a fusion layer.

Why?

- Offers a more advanced alternative to the LSTM approach, potentially catching subtler patterns.
- Tests whether our metric performs better with a scale up in the architecture

TextBasedCreativityPredictor

- **Inputs**:
  - **Text Segments**: Processed into embeddings using **BERT**.
- **Processing**: Embeddings generate weights using LSTM, which are outputed and then **applied to CI Formula.**

Same as SimpleCreativityPredictor, but outputs index weights instead of direct score

Why?

- Adapts the importance of each metric based on the text's content
- Test whether **these metrics alone** can predict creativity
- Makes the evaluation more interpretable by showing which metrics matter most for a given text.

Different Models Used & Why

Team ID: 26

SimpleCreativityPredictor (Baseline Model)

- **Inputs**:
    - **Text Segments**: Processed into embeddings using **BERT**.
- **Processing**: Embeddings are processed by an **LSTM**.

Why?

- Tests whether **the text alone** (**without metrics**) can predict creativity, acting as a simpler baseline.
- Helps us understand if calculating metrics adds value

CompositeRegressor

- **Inputs**:
    - **Text Segments**: Processed into embeddings using **BERT**.
    - **Metric Features**: Creativity metrics fed into an **LSTM**.
- **Processing**: Embeddings and LSTM outputs are combined in a fusion layer.

Why?

- Blends raw text data with explicit metric values
- Starting point to see how well text and metrics together predict creativity.

TransformerCreativityAggregator

- **Inputs**:
    - **Text Segments**: Processed into embeddings using **BERT**.
    - **Metric Features**: Creativity metrics fed into a Transformer.
- **Processing**: Embeddings and Transformer outputs are combined in a fusion layer.

Why?

- Offers a more advanced alternative to the LSTM approach, potentially catching subtler patterns.
- Tests whether our metric performs better with a scale up in the architecture

TextBasedCreativityPredictor

- **Inputs**:
    - **Text Segments**: Processed into embeddings using **BERT**.
- **Processing**: Embeddings generate weights using LSTM, which are outputed and then **applied to CI Formula.**

Same as SimpleCreativityPredictor, but outputs index weights instead of direct score

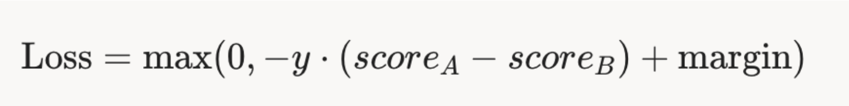
Why?

- Adapts the importance of each metric based on the text's content
- Test whether **these metrics alone** can predict creativity
- Makes the evaluation more interpretable by showing which metrics matter most for a given text.

Training

Team ID: 26

- **Contrastive Learning** approach to distinguish between more/less creative texts
- **Data**: **1000** pairs (500 literary, 500 technical), split 80% train, 20% validation, were prepared.
- Loss Function: **Margin Ranking Loss**



$$\text{Loss} = \max(0, -y \cdot (score_A - score_B) + \text{margin})$$

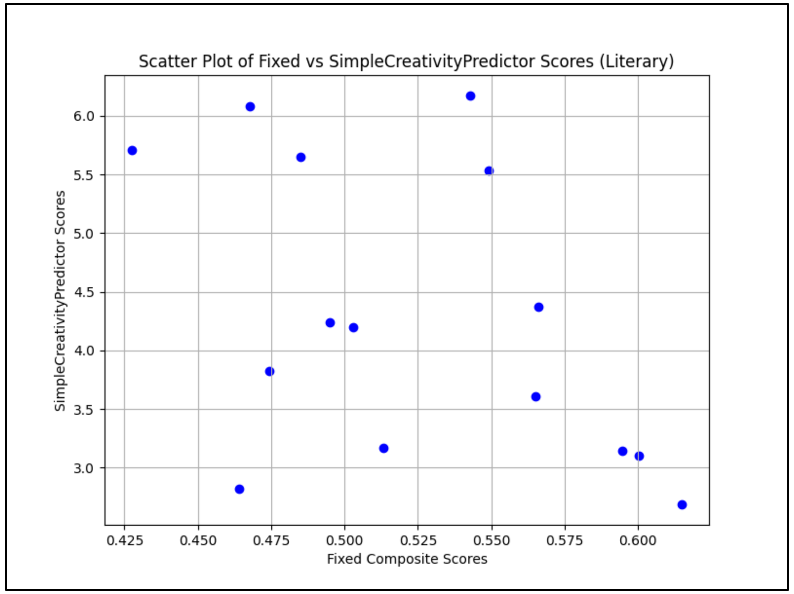
- ○ score_A : Creativity score for the preferred text.
- ○ score_B : Creativity score for the less preferred text.
- ○ y=1: Indicates $\text{score}_A > \text{score}_B$
- ○ margin=1.0: Ensures a minimum difference between scores.

Training Results

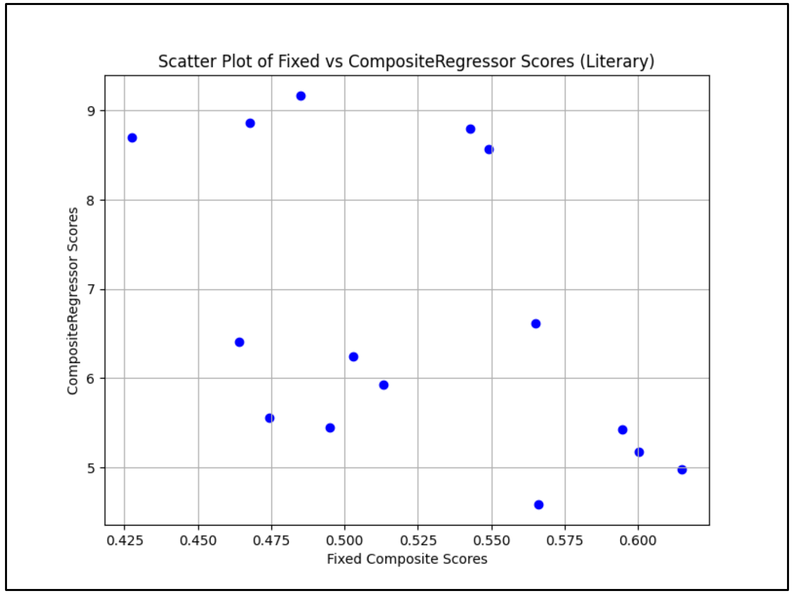
Model	Best Val Loss	Epoch	Train Loss
SimpleCreativityPredictor (Baseline)	0.7423	17	0.7222
CompositeRegressor	0.6259	14	0.6174
TransformerCreativityAggregator	0.5835	24	0.5956
TextBasedCreativityPredictor	0.6183	39	0.6139

Scatter Plots of Scores across Domains

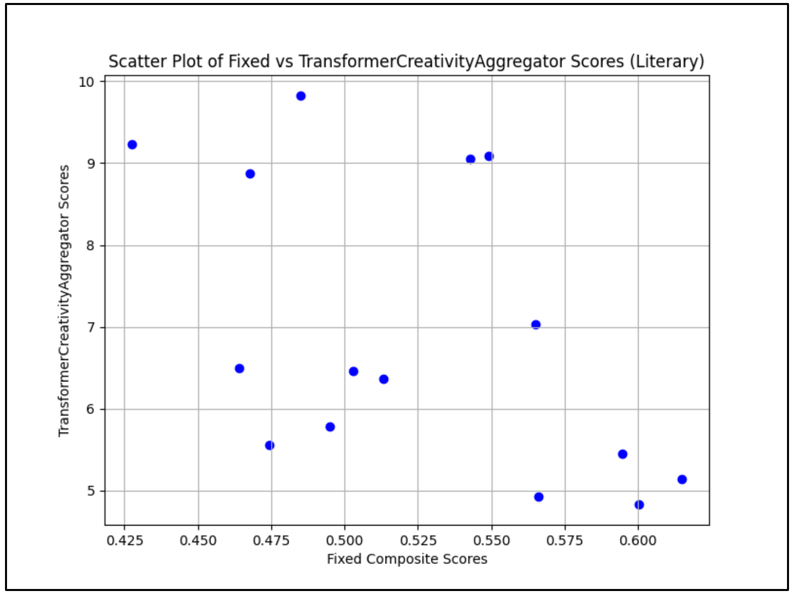
Team ID: 26



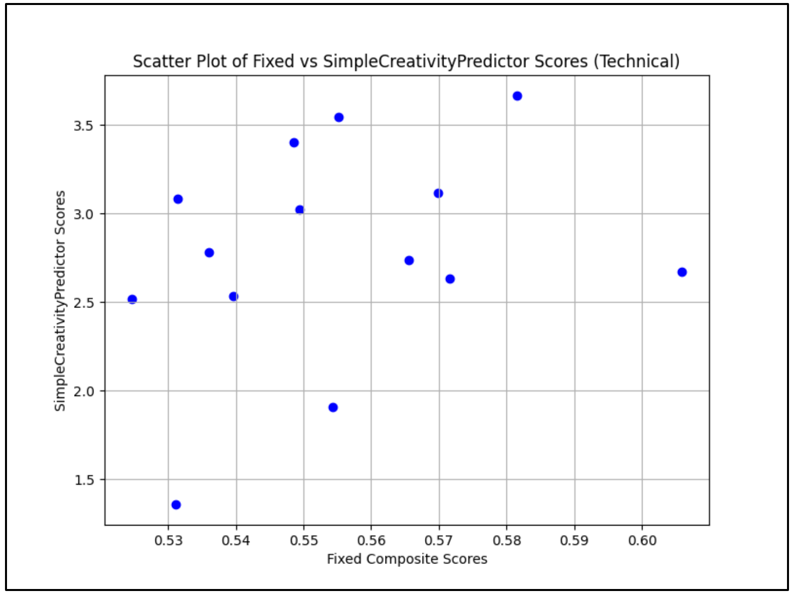
Scatter Plot of Fixed vs SimpleCreativityPredictor Scores (Literary)



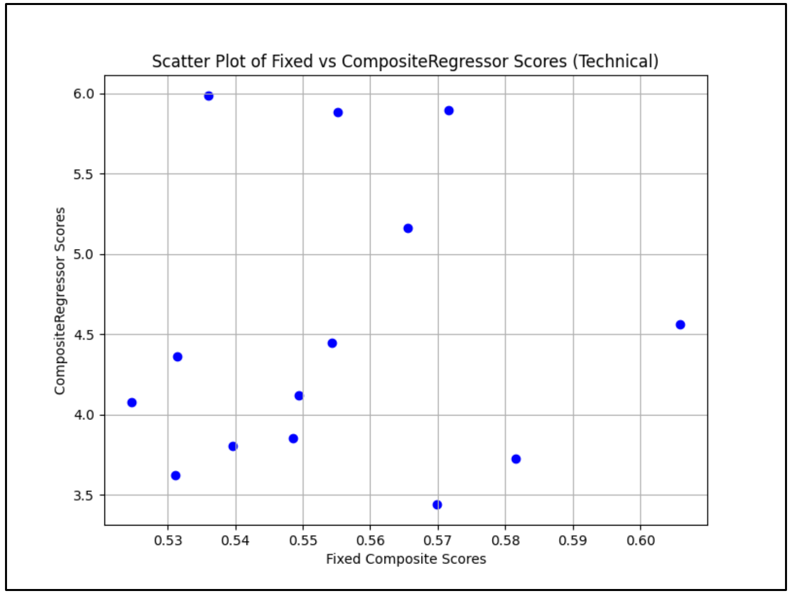
Scatter Plot of Fixed vs CompositeRegressor Scores (Literary)



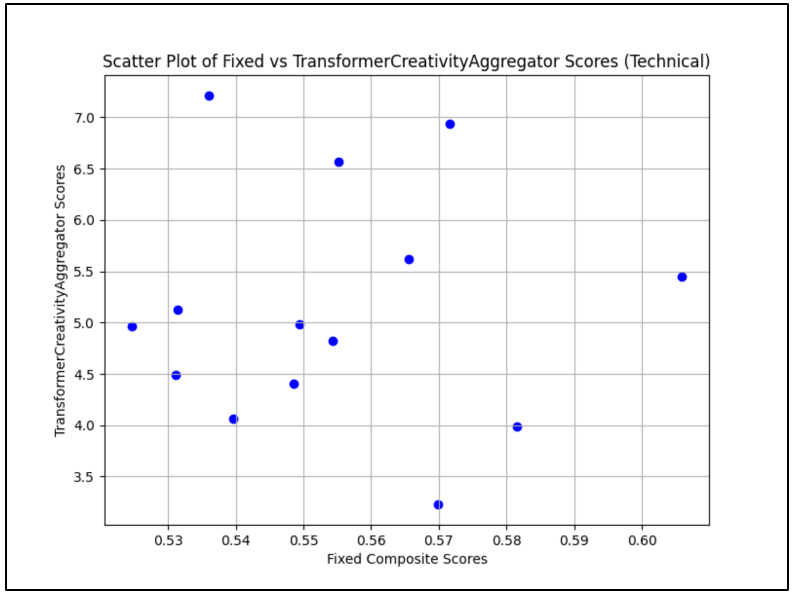
Scatter Plot of Fixed vs TransformerCreativityAggregator Scores (Literary)



Scatter Plot of Fixed vs SimpleCreativityPredictor Scores (Technical)



Scatter Plot of Fixed vs CompositeRegressor Scores (Technical)



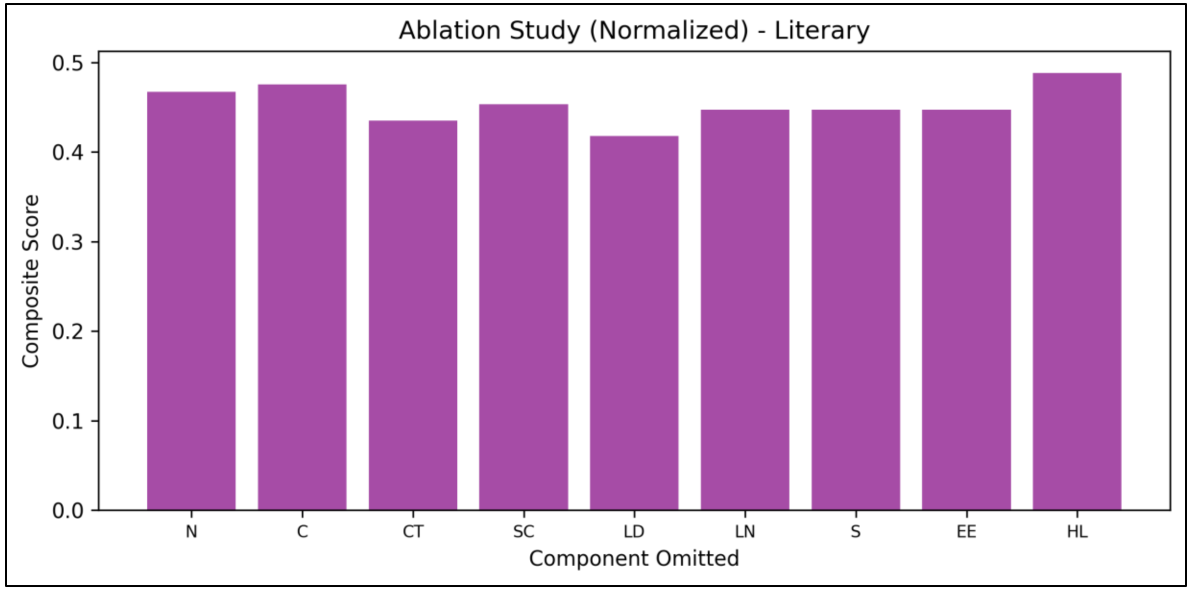
Scatter Plot of Fixed vs TransformerCreativityAggregator Scores (Technical)

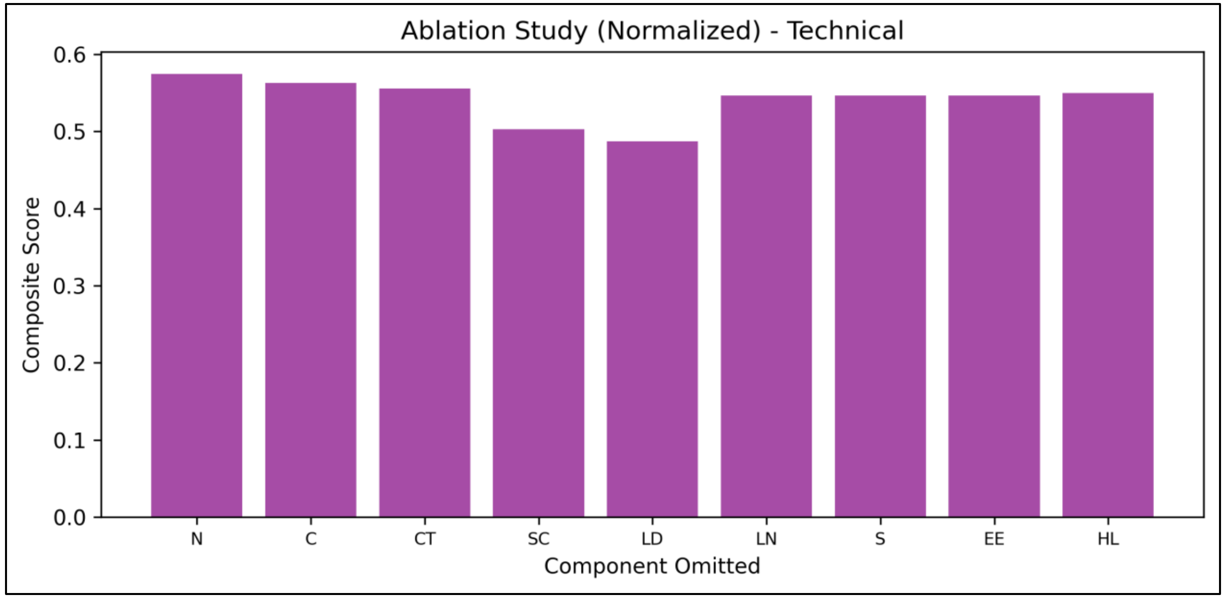
Ablation Studies

Team ID: 26

Done by systematically excluding one of the nine creativity metrics at a time, recomputing the composite score using the remaining normalized metrics with adjusted weights that sum to 1, and then comparing these scores to assess each metric's impact.

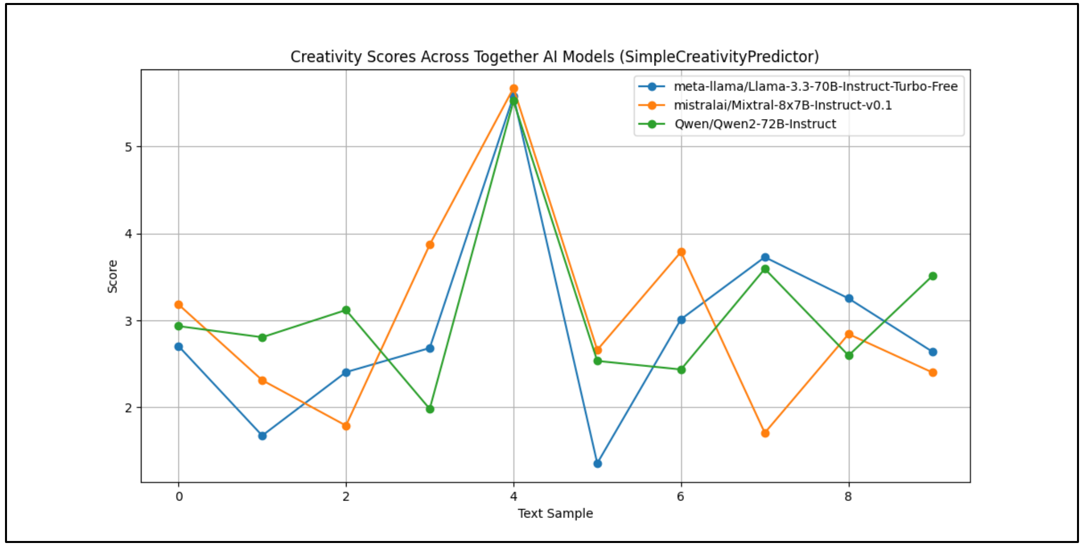
The metrics are averaged over the 4 model outputs and over 5 results for a prompt, and 3 prompts for each domain (15 results per model / domain or 60 results/domain)

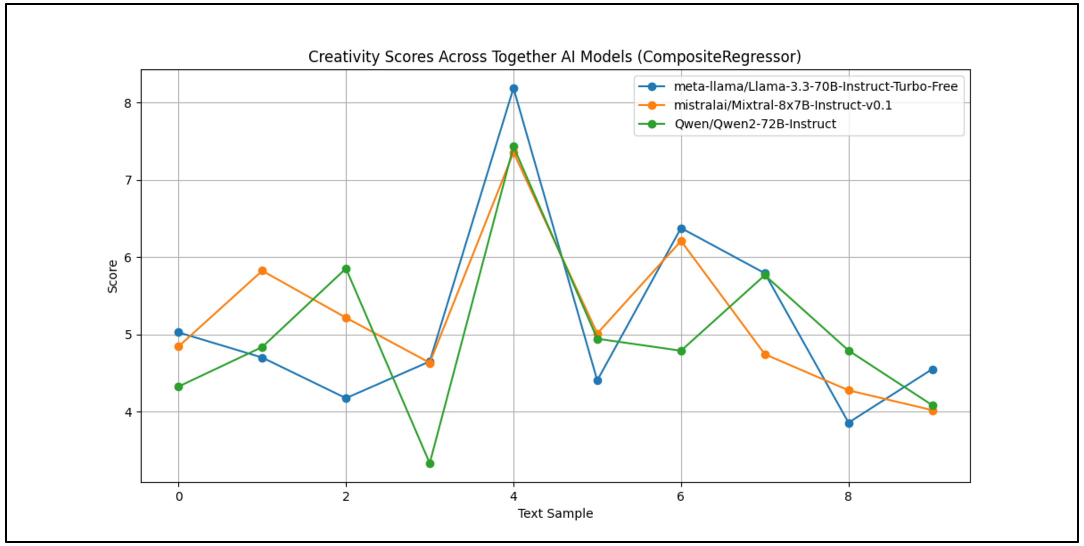


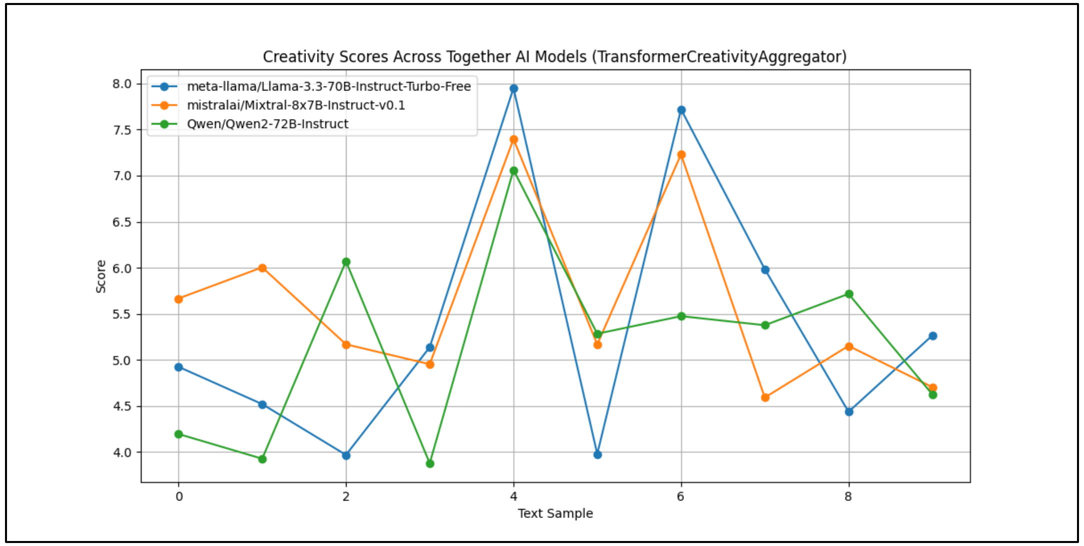


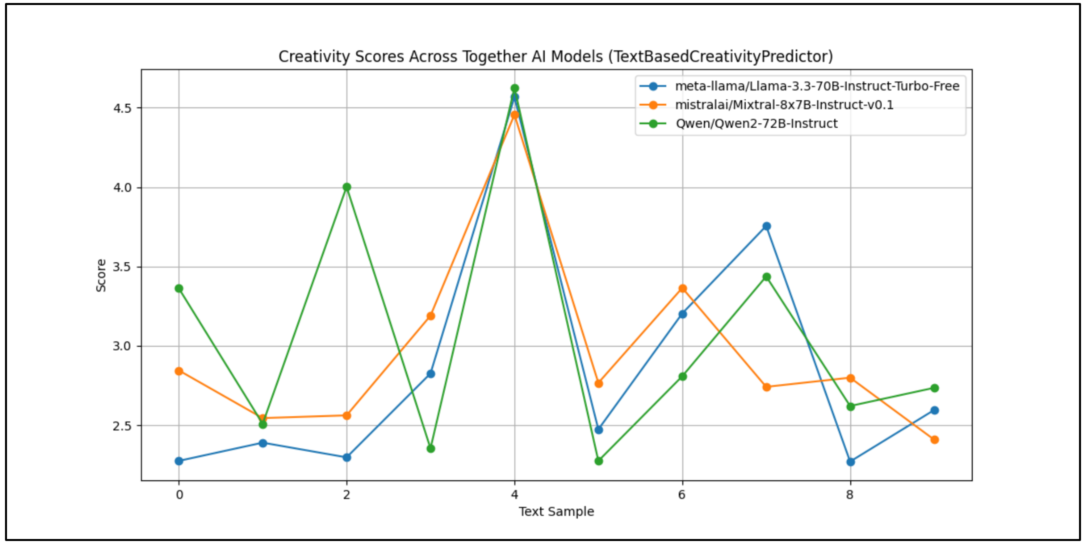
Comparison of Together AI Models across Metrics

Team ID: 26









Next Steps

Team ID: 26

1. Increase Training Dataset from 1000 pairs to 2000 pairs
2. Analyze Score Trends and Edge Cases
3. Evaluate Metrics against existing creativity indexes (Ex: from ˝AI as Humanity's Salieri")

Thank You

Have a Great Day!