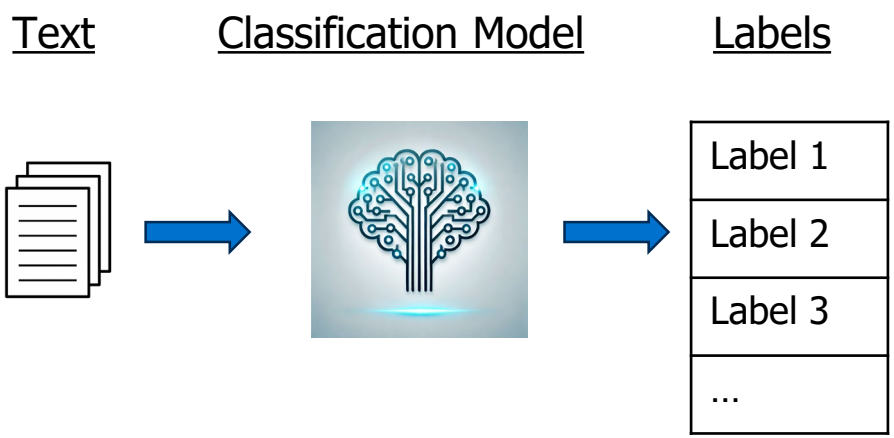
Applied Generative AI

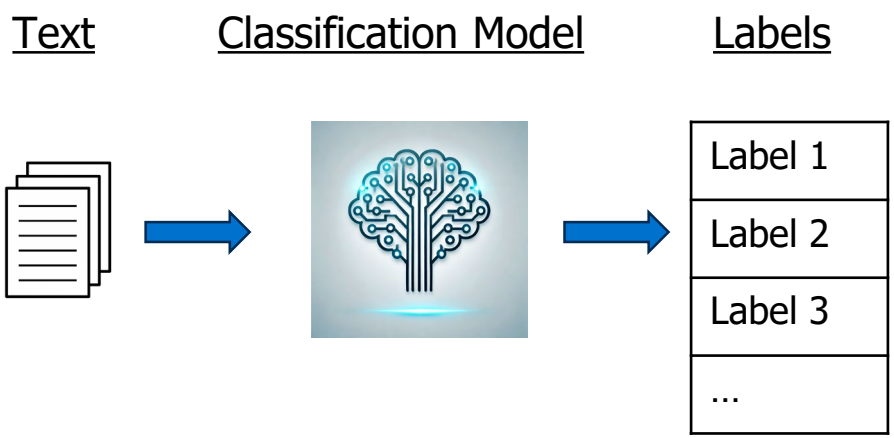
Text-to-Label Tasks: Flow and Applications

What are Text-to-Label Tasks?



- Tasks involving classification or label assignment to text inputs.
- Examples:
  - Sentiment analysis
  - Topic classification
  - Spam detection
  - Code bugs

What are Text-to-Label Tasks?



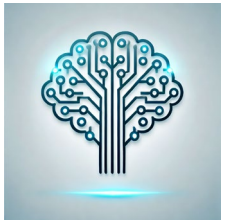
- Importance across several industries:
  - Customer service
  - E-commerce
  - Healthcare

What are Text-to-Label Tasks?

Text	Featurization	Classification Model	Labels (Sentiment)

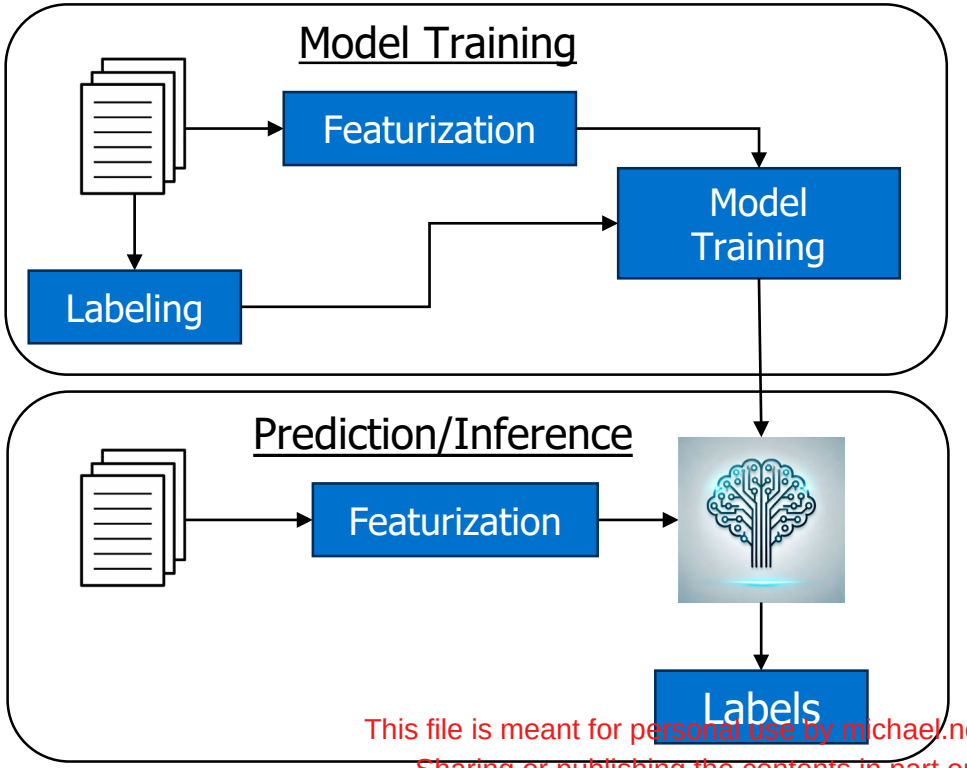
"The product arrived two days late, but it works perfectly."

product	1
package	0
arrive	1
perfect	1
...	



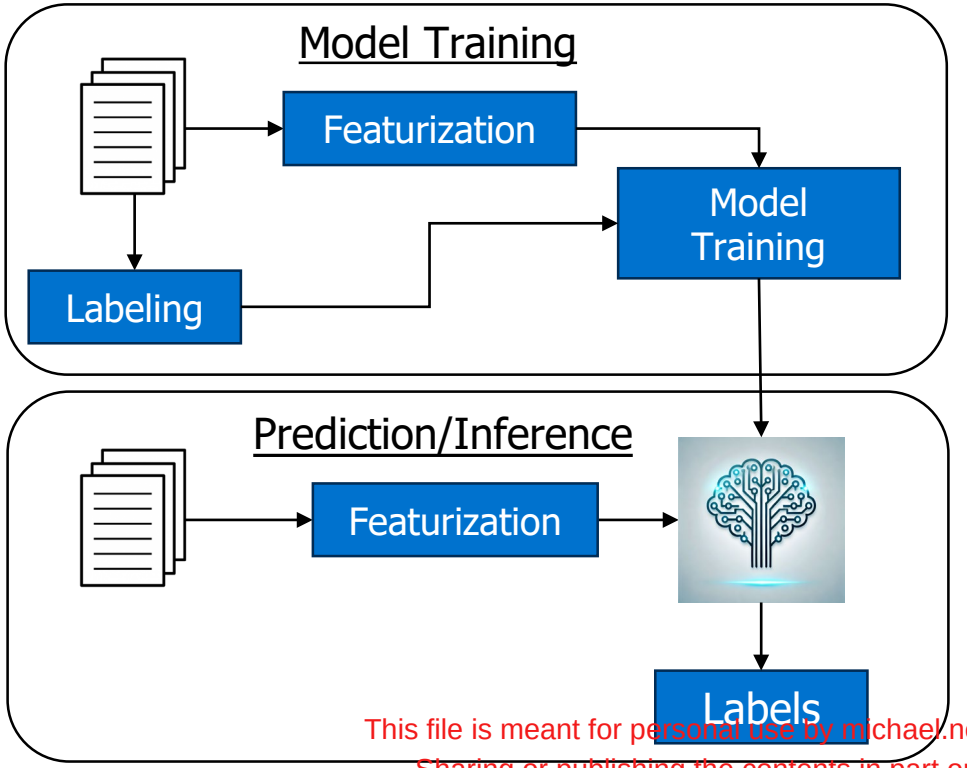
Mixed	88%
Negative	31%
Positive	18%
Neutral	4%

Traditional Approaches



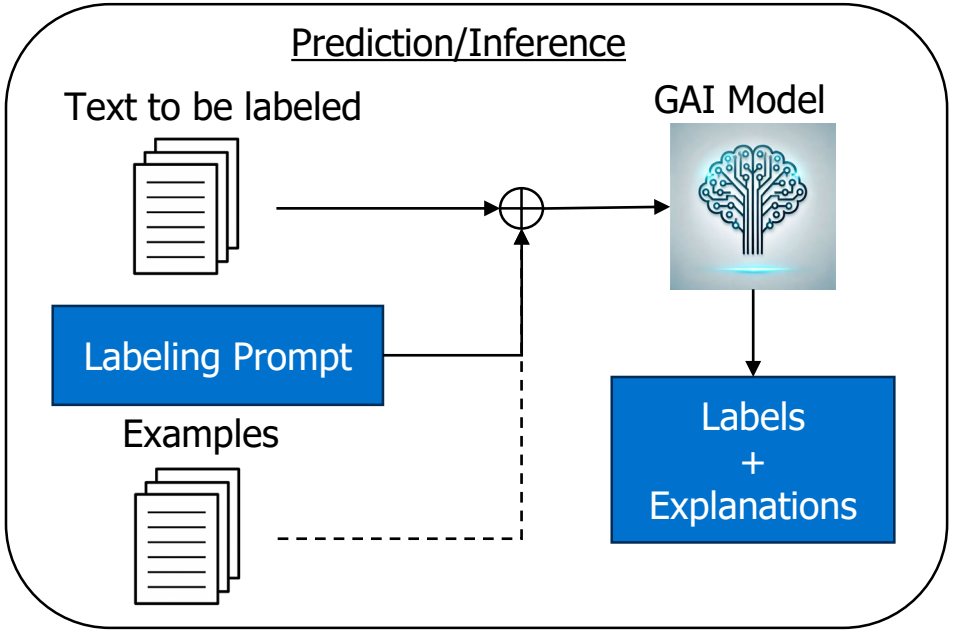
- Traditionally done by supervised machine learning methods:
  - Classical models like Naive Bayes, SVMs
  - Deep learning models
- Featurization is critical.

Traditional Approaches



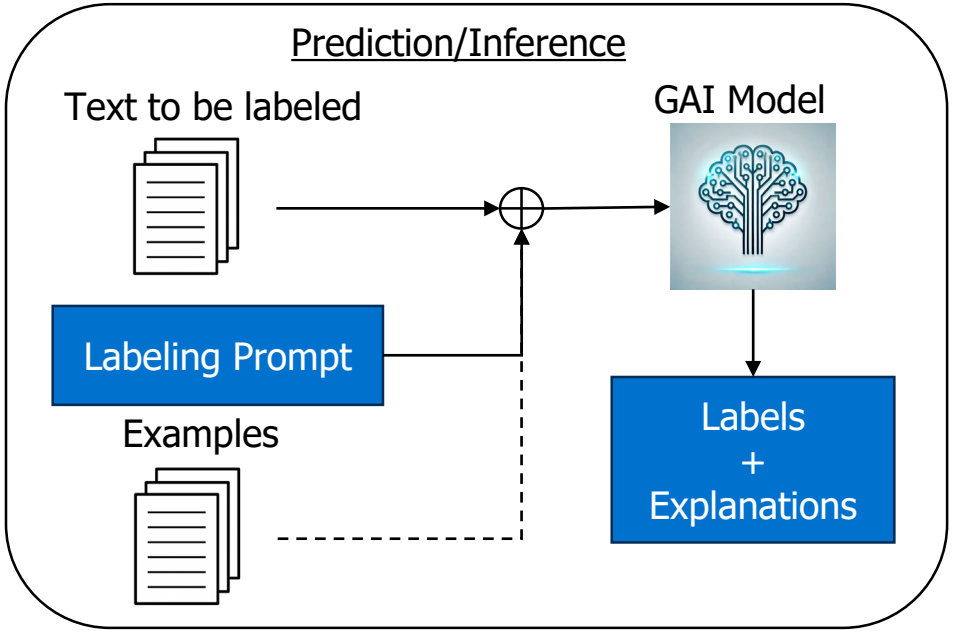
- Challenges with traditional methods:
  - Poor performance in low-resource scenarios.
  - Intensive model training and feature engineering.
  - Difficulty handling nuances like sarcasm or idioms.

Generative AI for Classification



- LLMs can frequently classify text without any training (zero-shot).
  - Can be augmented with few demonstrations (few-shot).
- Exploit open-endedness property of LLMs.

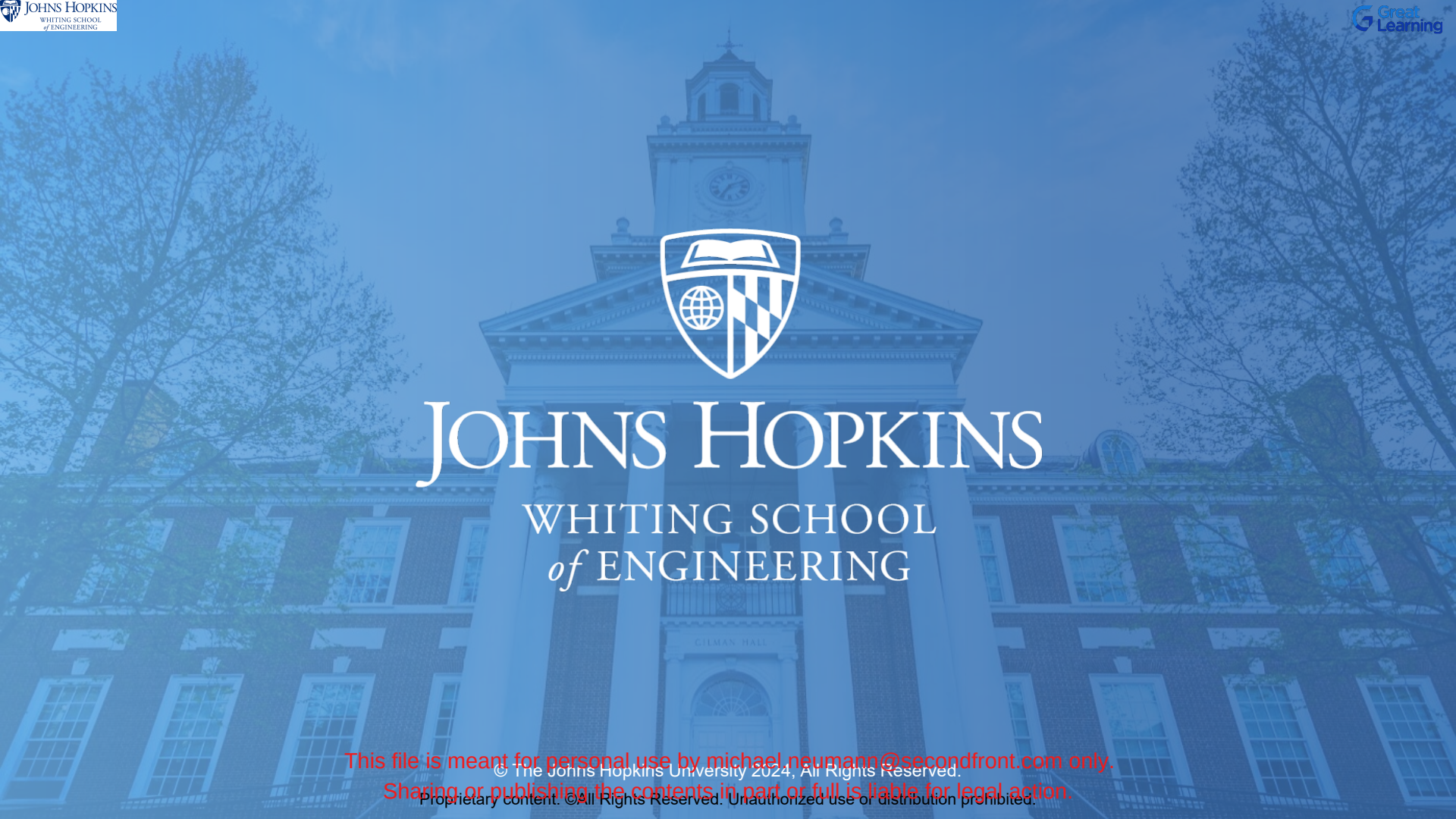
Generative AI for Classification



- Key Benefits:
  - Better handling of ambiguity and context.
  - Lower dependency on labeled datasets.
  - Provides explanations along with predictions.

Key Takeaways

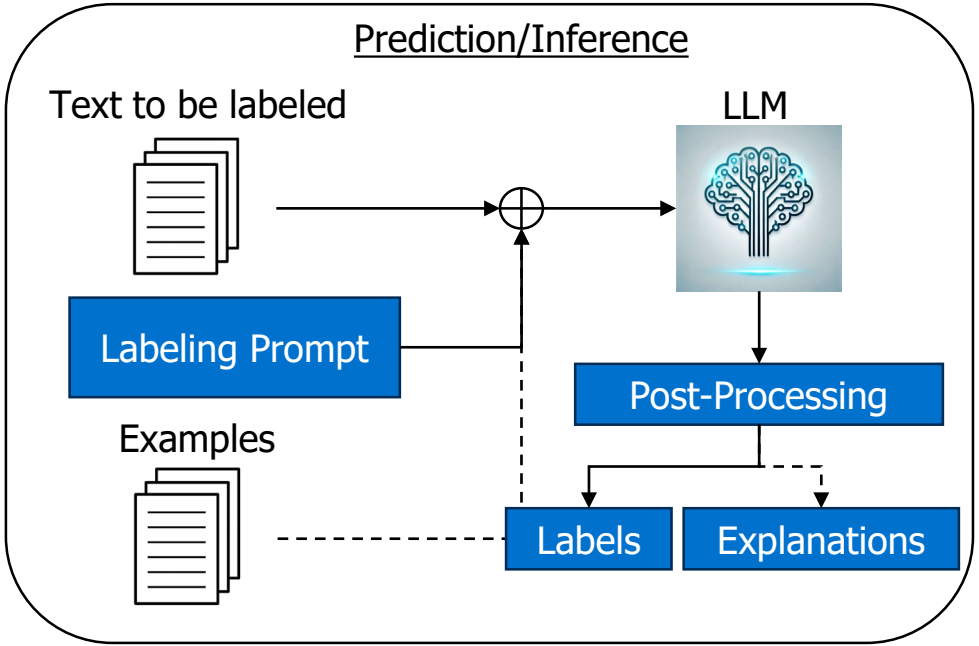
- Text-to-label tasks are crucial for many AI applications.
  - Ex: Sentiment analysis for product reviews.
- Traditional methods frequently require a supervised machine learning approach which involves labeling lots of data and training a model.
- Traditional methods can struggle with complex language (e.g., sarcasm) and out-of-domain or out-of-distribution data.
- Generative AI simplifies workflows, improves performance, and reduces labeling costs.



Applied Generative AI

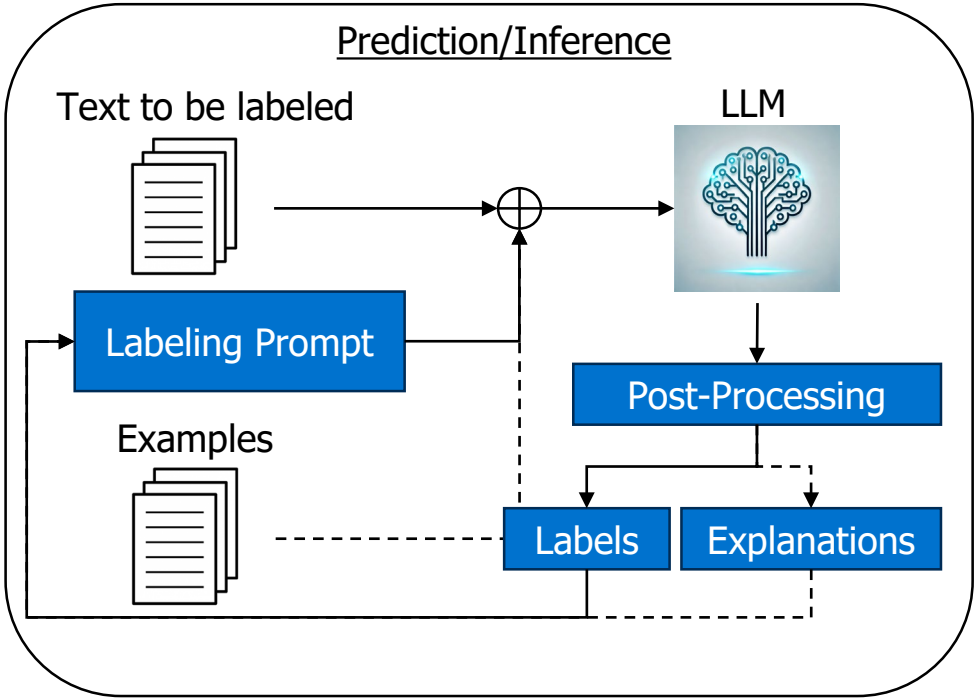
Using Generative AI for Classification: Benefits and Approaches

Generative AI for Classification



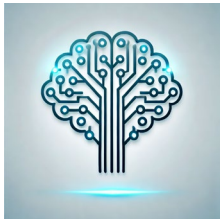
- LLMs can frequently classify text without any training (zero-shot).
  - Can be augmented with few demonstrations (few-shot).
- Key Components:
  - Labeling Prompt
  - LLM
  - Examples
  - Post-Processing

How does it work?

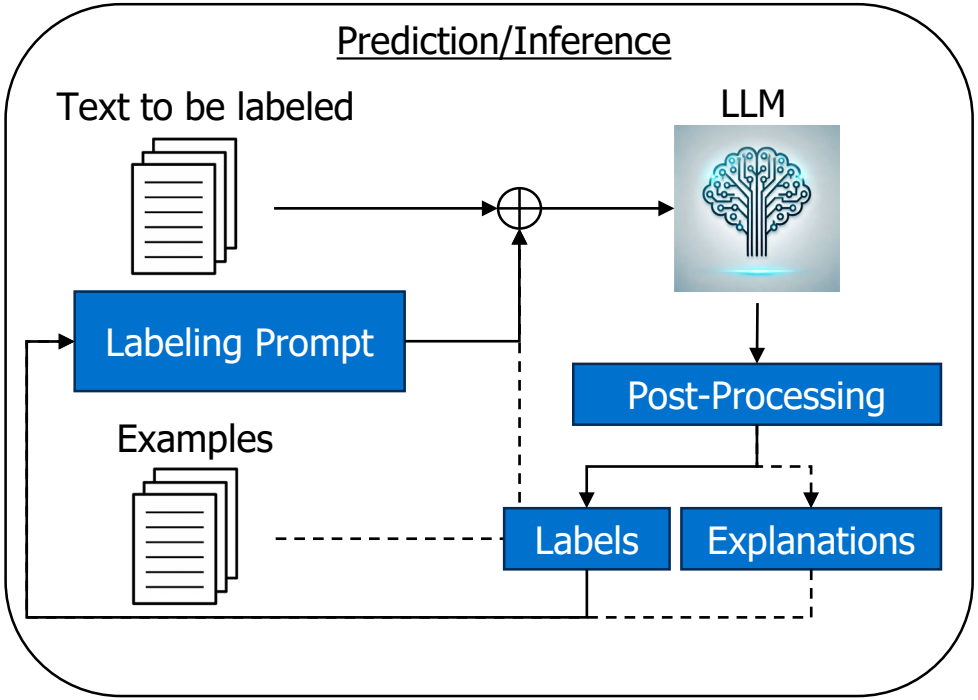


- Labeling by Generative AI:
  1. Develop prompting scheme.
  2. Add text to be labeled to prompt and send it to the model.
  3. Post-Process output.
  4. Refine the prompt based on some labeled examples.
- Consider few-shot examples and batch prompting, if applicable.

How does it work?

Text	Labeling Prompt	LLM	Labels (Sentiment)
"The product arrived two days late, but it works perfectly."	*The following is a product review, please classify the sentiment.* "The product arrived two days late, but it works perfectly." *Sentiment:*		"mixed"

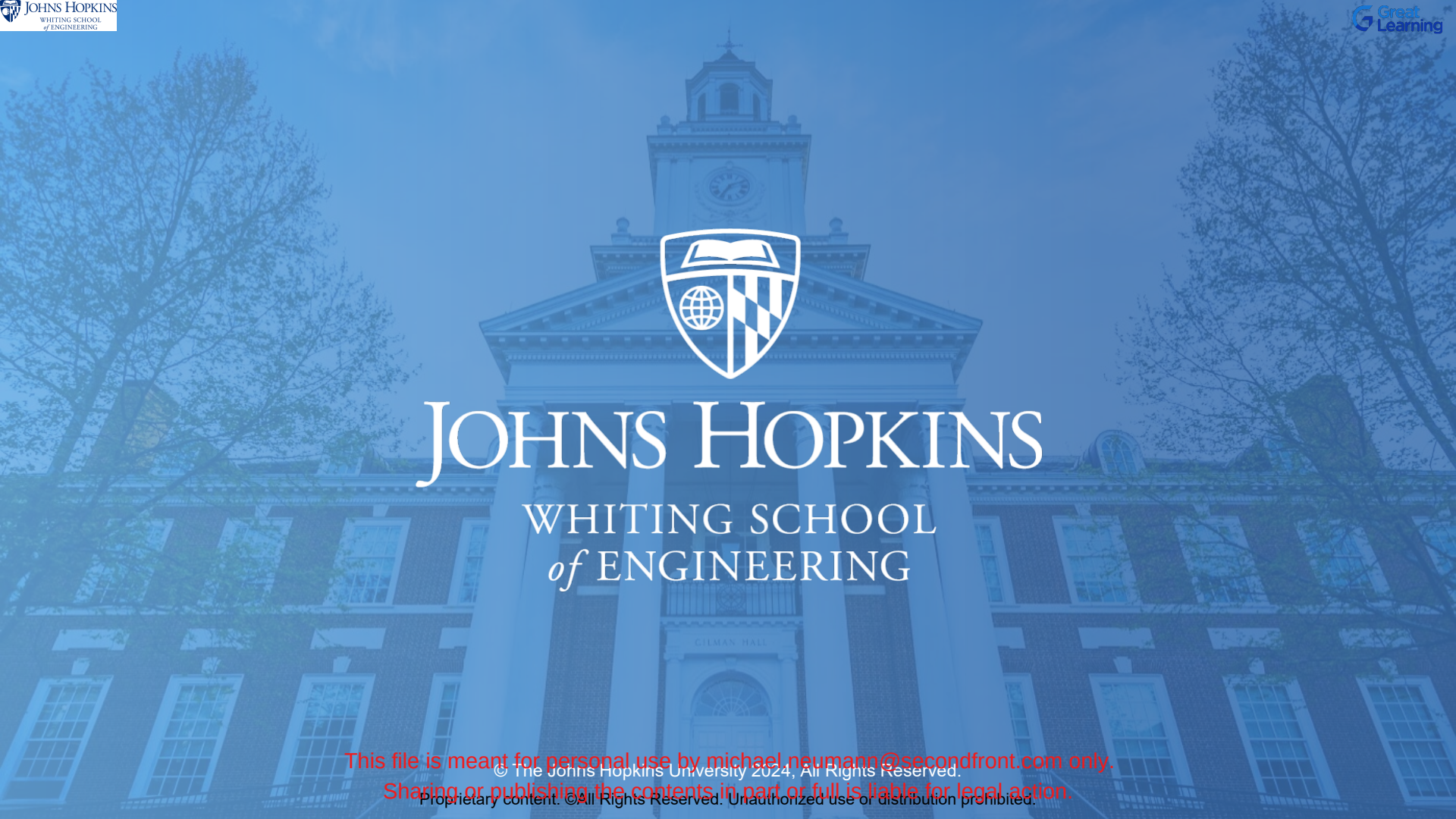
Challenges and Mitigation Strategies



- Challenges:
  - Prompt sensitivity: Results may vary based on prompt phrasing.
  - Lack of explainability in some cases.
- Mitigation:
  - Iterative prompt refinement.
  - Use explainable AI techniques or combine outputs with traditional models for critical tasks.

Key Takeaways

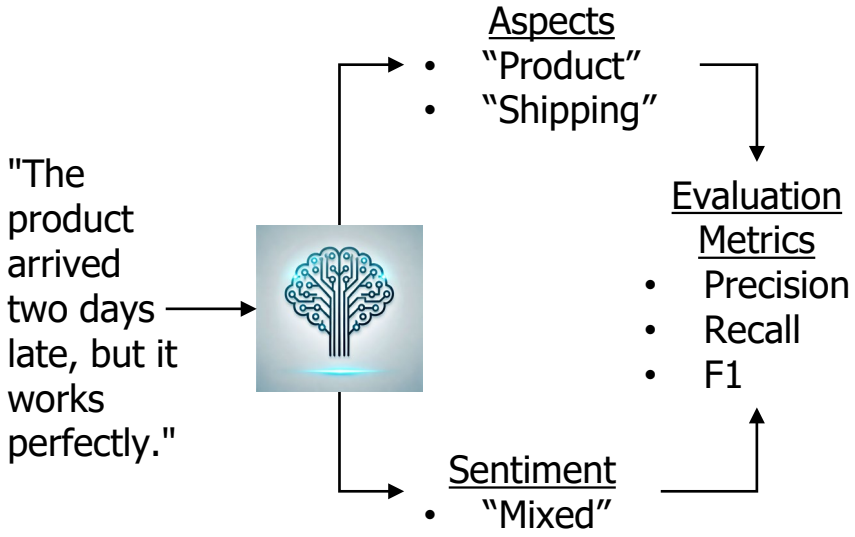
- Traditional methods struggle with complexity, while Generative AI offers flexibility.
  - LLMs are capable of understanding wide range of tasks and language.
- Prompt engineering is crucial to the success of labeling workflows.
- Generative AI simplifies workflows, improves performance, and reduces labeling costs.
  - It can often work in zero or few-shot settings.



Applied Generative AI

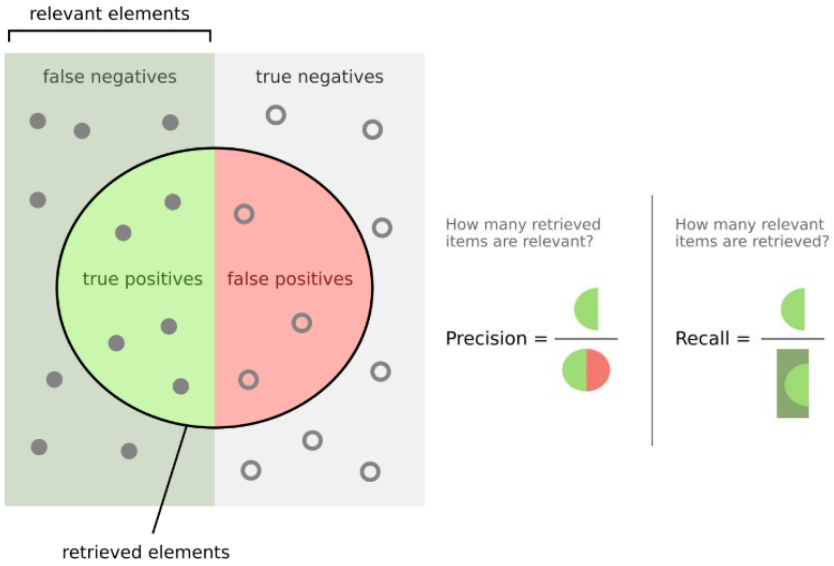
Using Generative AI for Classification: Evaluation Metrics

Introduction to Evaluation Metrics



- Purpose of evaluation metrics is to assess the quality of model predictions:
  - Identify strengths and weaknesses in the classification process.
  - Optimize performance and refine workflows.
- Multiple metrics can be used to assess different aspects of the output.

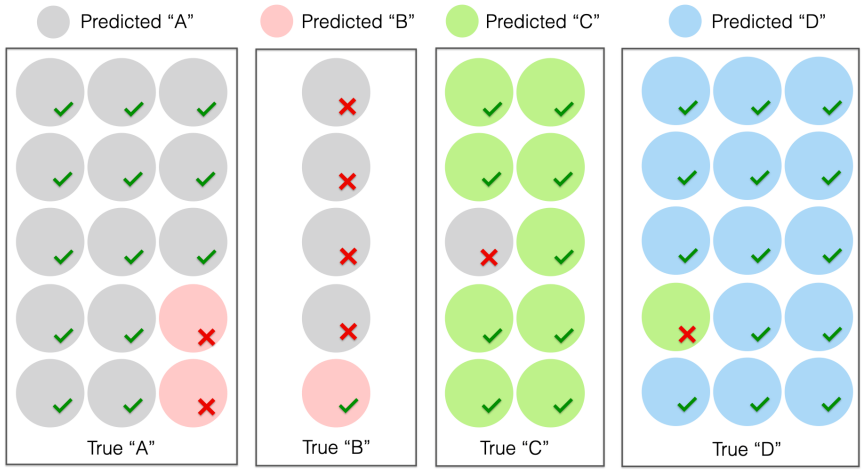
Key Metrics for Classification



- Precision: Measures how many of the predicted positive cases were correct.
- Recall (Sensitivity): Measures how many actual positive cases were correctly predicted.
- F1 Score: Harmonic mean of Precision and Recall; balances the trade-off between the two.

Precision and Recall. *Wikimedia Commons*. https://commons.wikimedia.org/wiki/Category:Precision_and_recall

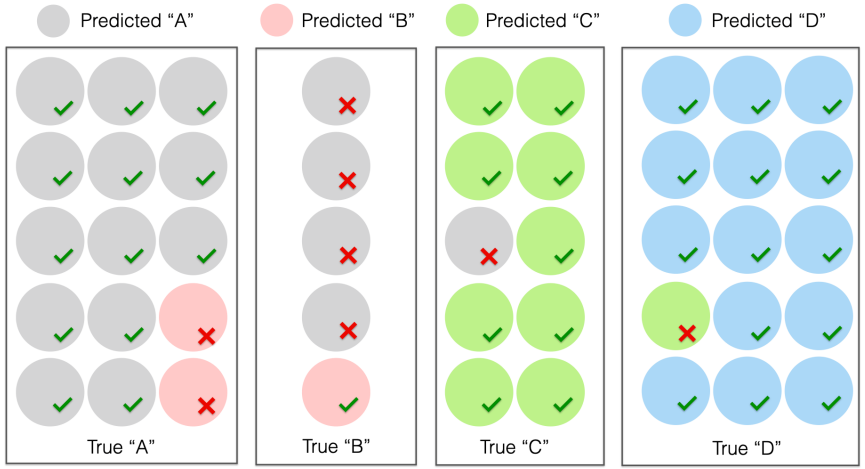
Metrics for Multi-Class and Multi-Label Classification



- Macro-Averaged Metrics:
  - Calculates metrics independently for each class and takes the average.
  - Useful when all classes/aspects are equally important.
- Weighted-Averaged Metrics:
  - Considers the frequency of each class/aspect in the dataset while averaging metrics.
  - Useful for imbalanced datasets.

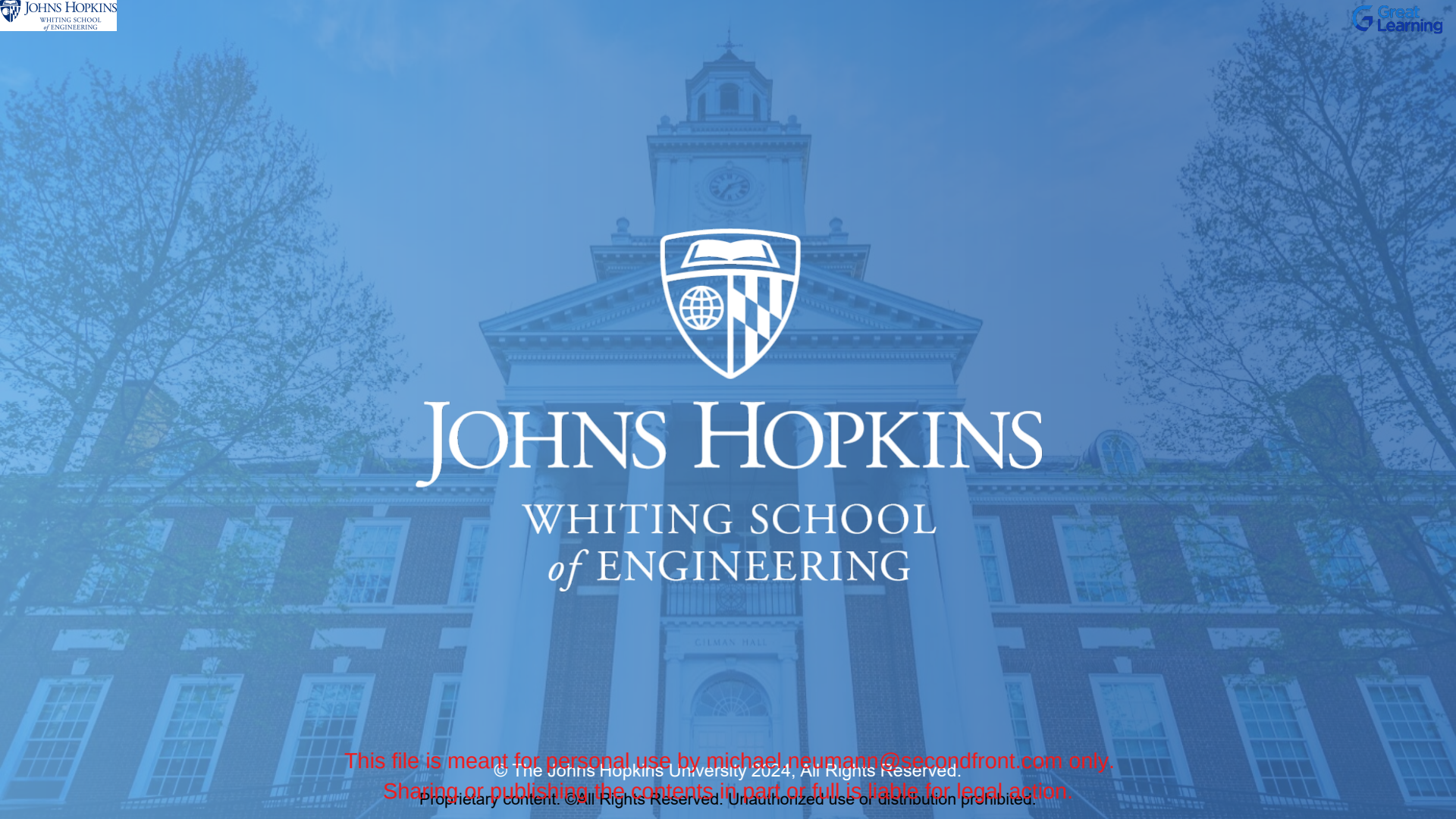
Accuracy, precision, and recall in multi-class classification. *Evidetly AI*. https://www.evidentlyai.com/classification-metrics/multi-class-metrics

Specific Challenges with Generative AI Evaluation



- Stochasticity in output:
  - “mixed” versus “Mixed” versus “the sentiment is mixed”.
- Ambiguity in output:
  - “camera is good” versus “the product is good”.
- Human review complements automated evaluation for edge cases.

Accuracy, precision, and recall in multi-class classification. *Evidetly AI*. https://www.evidentlyai.com/classification-metrics/multi-class-metrics



Applied Generative AI

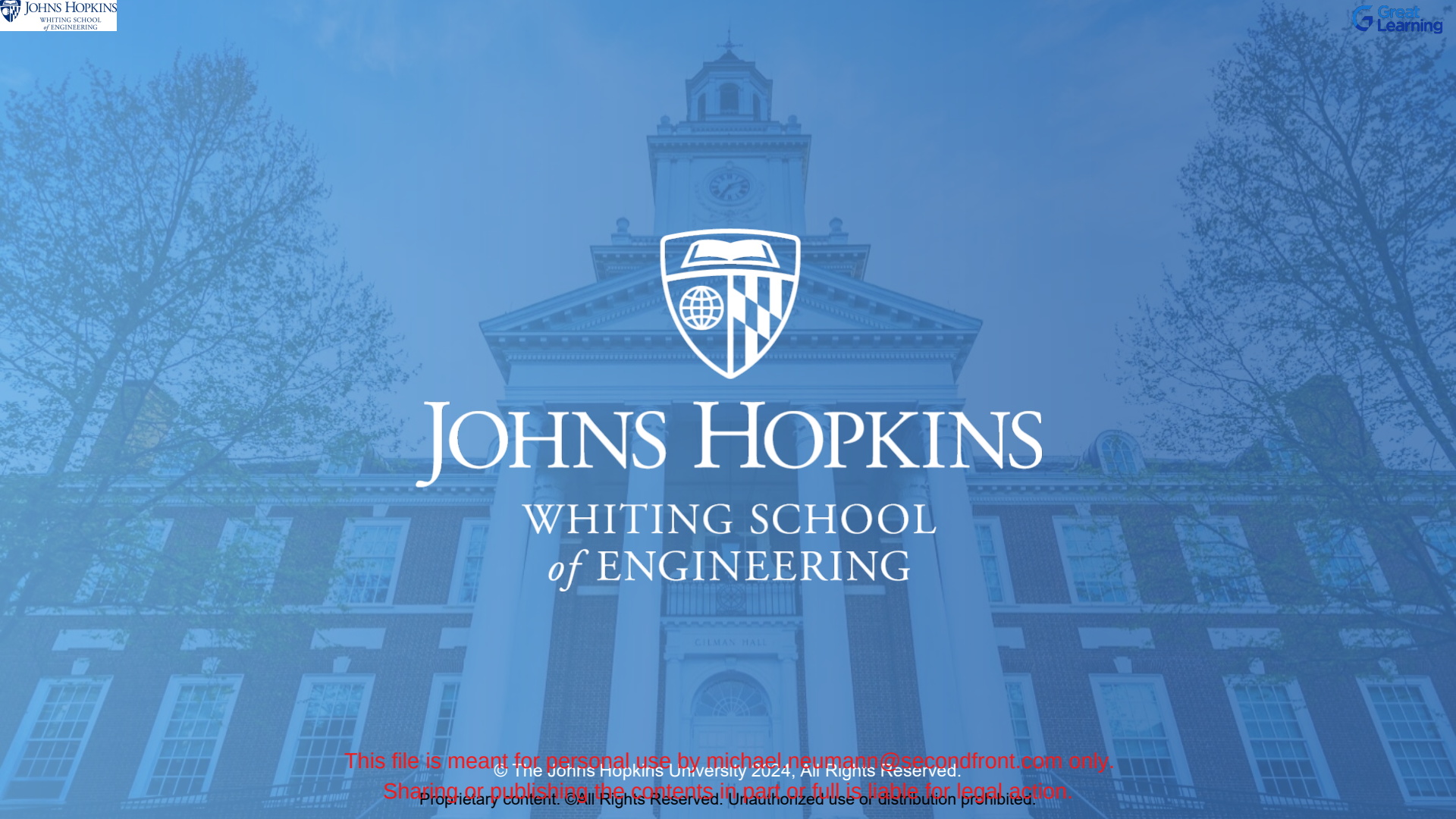
Aspect-Based Sentiment Classification Walkthrough

What is Aspect-Based Sentiment Classification?

- ABSC identifies the sentiment expressed about specific aspects or attributes of a product, service, or topic within a piece of text.
- Example: *"The battery life of this phone is amazing, but the camera is mediocre."*
  - Sentiment towards "battery life" → Positive
  - Sentiment towards "camera" → Negative
- Key Features of ABSC:
  - Goes beyond overall sentiment to analyze individual components.
  - Crucial for customer feedback, social media analysis, and product improvement.

Key Takeaways

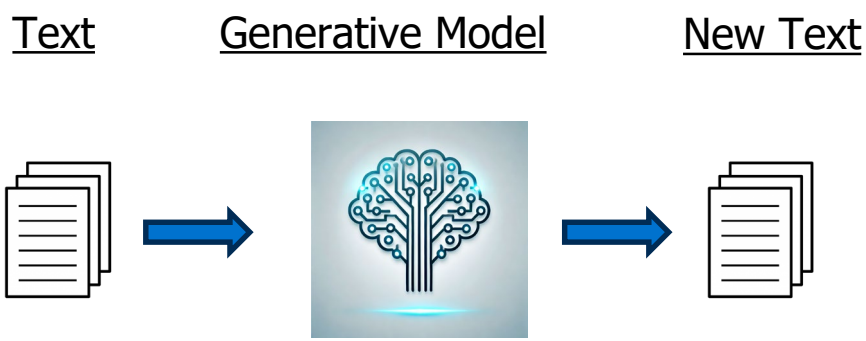
- ABSC provides nuanced, granular understanding by targeting specific aspects.
- Leveraging ABSC in AI pipelines enhances customer experience analysis and decision-making.
- Large Language Models (LLMs) streamline ABSC tasks by combining natural language understanding and classification accuracy.



Applied Generative AI

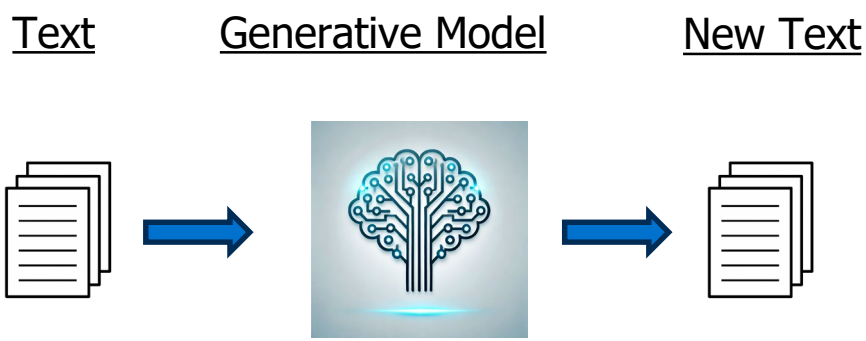
Text-to-Text Tasks: Flow and Use Cases

What are Text-to-Text tasks?



- Tasks where both the *input and output* are text.
- Examples:
  - Summarization
  - Translation
  - Question Answering
  - Text Style Transfer

What are Text-to-Text tasks?



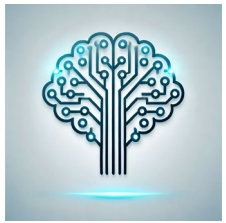
- Generative AI models are trained to generate new content based on input content.
- Generative AI models (e.g., LLMs) are inherently well-suited for these tasks.

How does it work?

Text	Tokenize and Embed	Transformer Model	Decode	Output

Translate the following into French: "Hello World!"

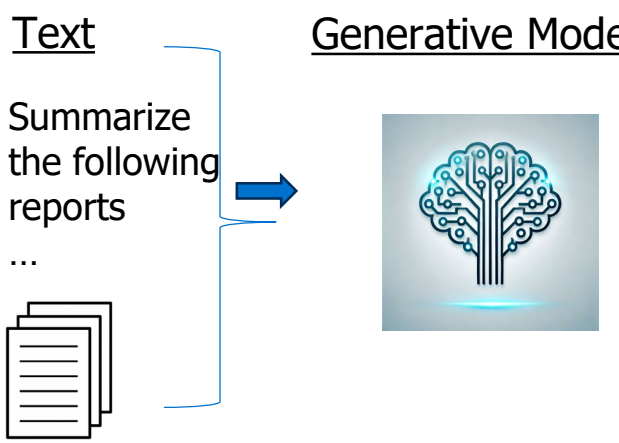
Translate	<vec>
the	<vec>
following	<vec>
into	<vec>
…	<vec>



<vec>	Bonjour
<vec>	le
<vec>	monde
<vec>	!

"Bonjour le monde!"

Use Cases of Text-to-Text Tasks

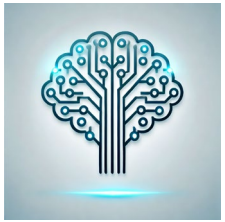


- Summarization
- **Use Case**: Creating concise summaries of lengthy reports.
- **Impact**: Saves time and aids decision-making.

Use Cases of Text-to-Text Tasks

Text → Generative Model → New Text

Translate the following …



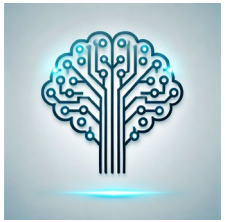
Bonjour les collègues …

- Translation
- **Use Case**: Bridging language barriers. Writing Code.
- **Impact**: Enables global collaboration and accessibility.

Use Cases of Text-to-Text Tasks

Text → Generative Model → New Text

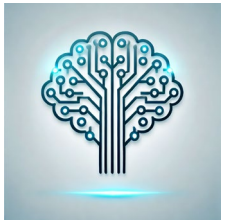
How do I return my package?



To return a package …

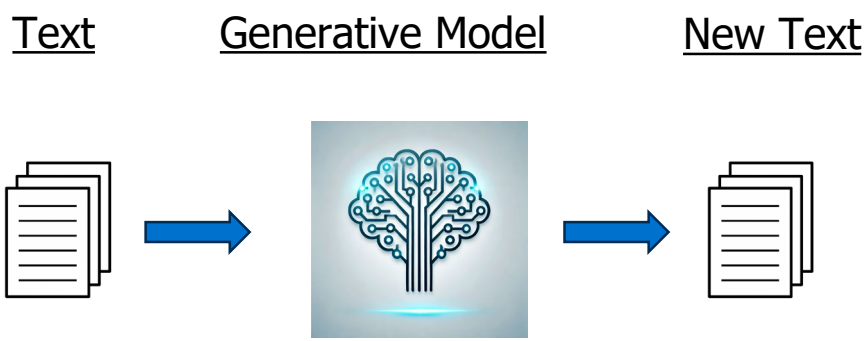
- Question Answering
- **Use Case**: Automates customer support queries.
- **Impact**: Enhances customer experience and reduces workload.

Use Cases of Text-to-Text Tasks

Text	Generative Model	New Text
Edit this e-mail to make it have a more professional tone: Hey Guys, ….		Dear Colleagues, ….

- Text Style Transfer
- **Use Case**: Adapts tone for business communication.
- **Impact**: Ensures consistency and professionalism.

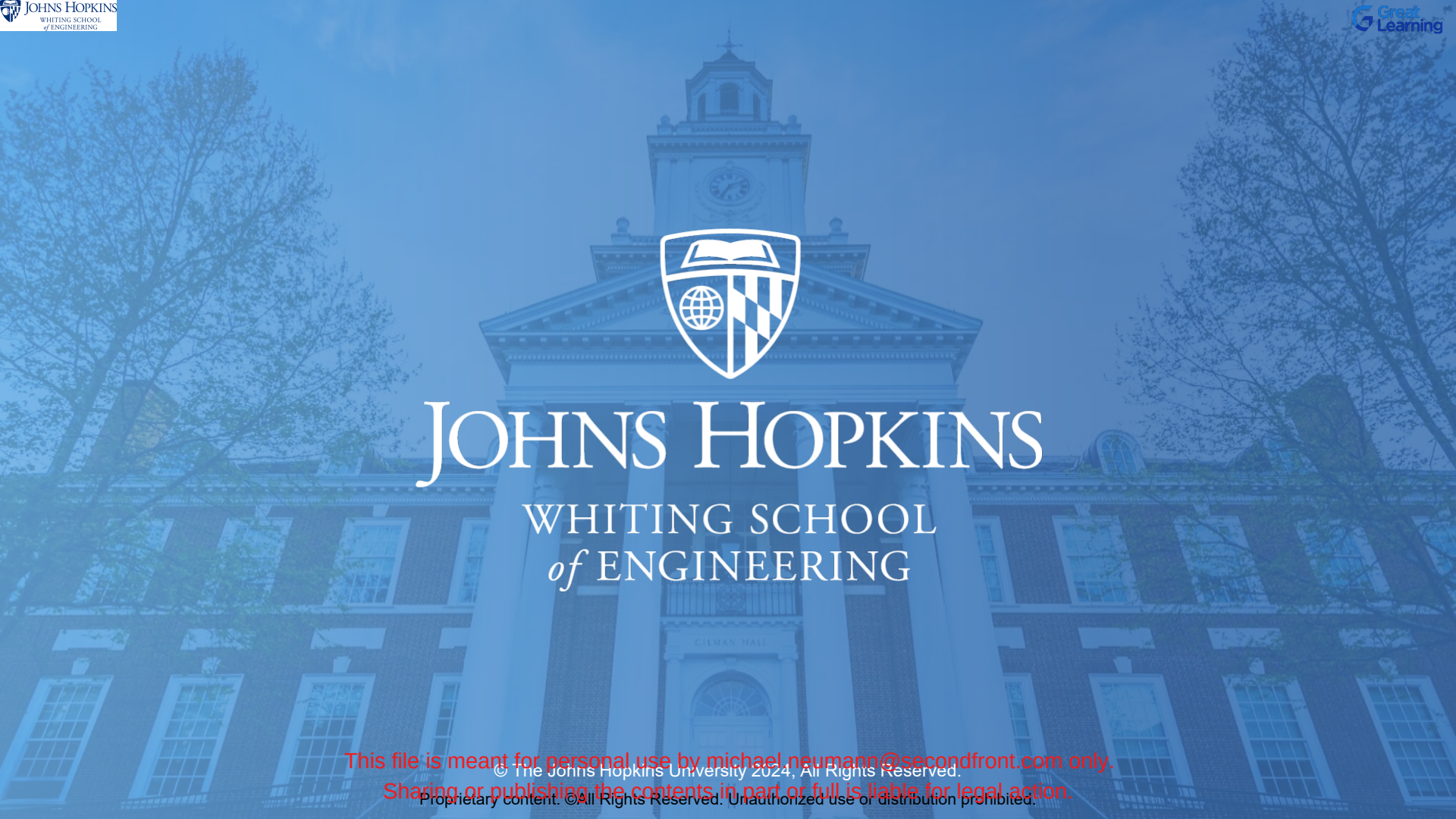
Why Use Generative AI for Text-to-Text Tasks?



- **Flexibility**: Can adapt to various tasks with simple prompt changes.
- **Context Understanding**: Handles nuanced requests like sarcasm, idioms, and ambiguous queries.
- **Efficiency**: Reduces time and effort compared to manual processes.

Key Takeaways

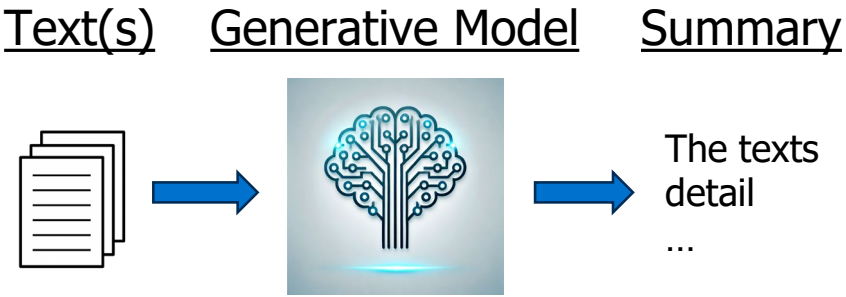
- Text-to-text tasks are foundational for many real-world AI applications.
- Generative AI excels in handling diverse tasks with high-quality results in the text domain.
- Understanding task flow and effective prompting are critical for effective use.



Applied Generative AI

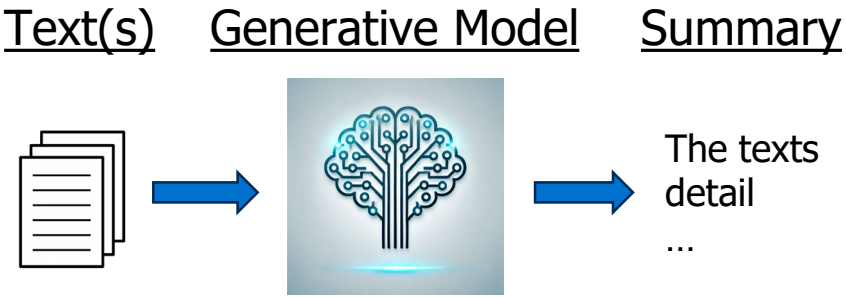
Summarization: Abstractive vs. Extractive Techniques

What is Summarization?



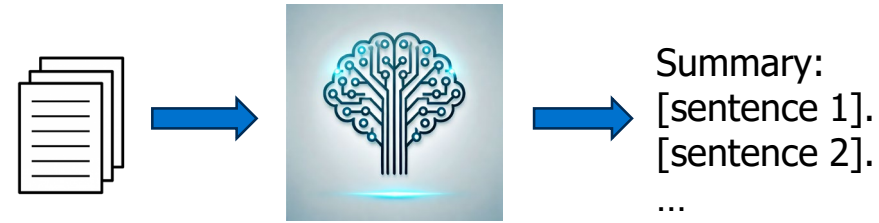
- **Definition**: The process of condensing a long text into a shorter version while retaining the core meaning.
- **Importance**: Enhances readability for lengthy documents like:
  - Research papers
  - Reports
  - News articles

What is Summarization?



- **Types of Summarization**:
  - **Extractive Summarization**: Selects key phrases or sentences directly from the original text.
  - **Abstractive Summarization**: Generates new sentences that summarize the text, mimicking human understanding.

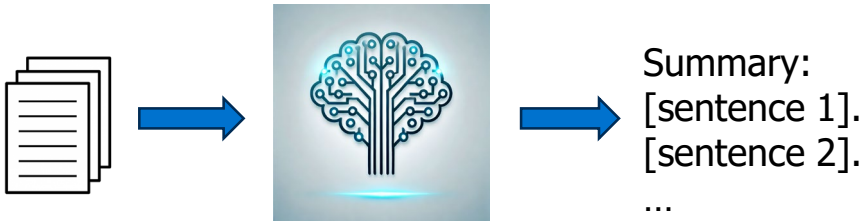
Extractive Summarization



- Directly identifies and pulls out the most important sentences or phrases from the text.
- **Key Techniques**:
  - TextRank algorithm
  - Keyword matching and heuristics

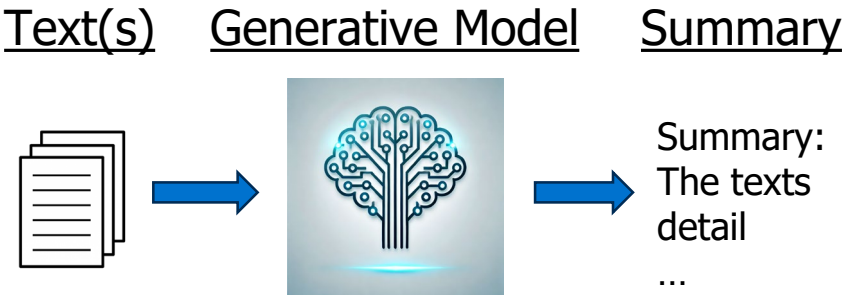
Extractive Summarization

Text(s) Generative Model Summary



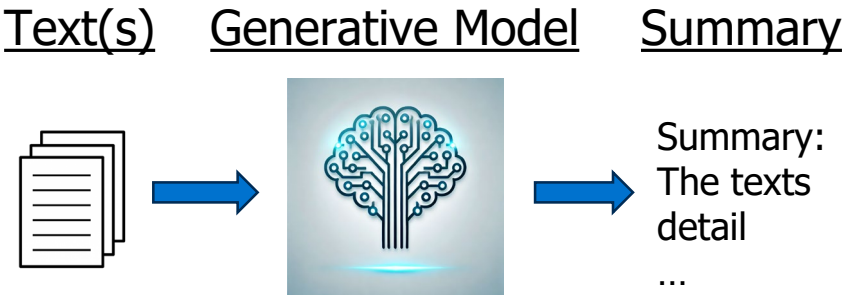
- **Advantages**:
  - Simple and computationally efficient.
  - Retains original text quality without paraphrasing errors.
- **Limitations**:
  - Lacks coherence when sentences are stitched together.
  - Misses the ability to rephrase or abstract ideas.

Abstractive Summarization



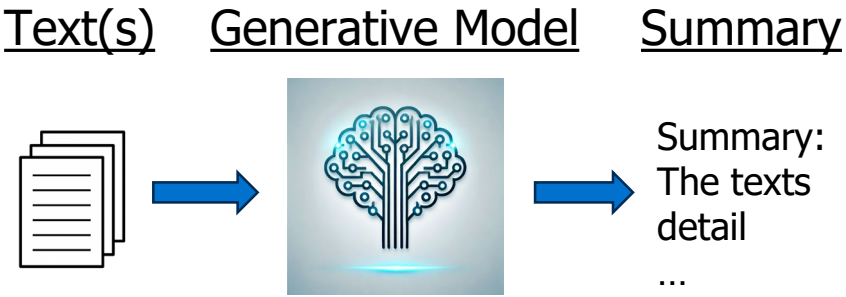
- Generates new sentences that encapsulate the core ideas of the original text.
- Leverages Generative AI models (e.g., GPT, T5, etc.) trained on summarization tasks.

Abstractive Summarization



- **Key Steps**:
  1. Selects a dedicated generative model for summarization or creates a summarization prompt.
  2. Encodes the input text into a dense representation.
  3. Decodes it into a coherent summary using generative mechanisms.

Abstractive Summarization



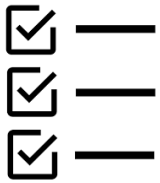
- **Advantages**:
  - Produces human-like summaries.
  - Captures ideas spanning across multiple sentences.
- **Limitations**:
  - Requires more computational resources.
  - Risk of generating inaccurate or hallucinated content.

Comparing Extractive and Abstractive Summarization

Original Text

"The company's profits increased significantly this quarter, owing to a rise in product sales and effective cost management."

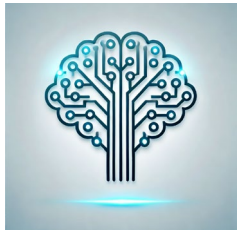
Heuristic Rules



Extractive Summary

"The company's profits increased significantly this quarter."

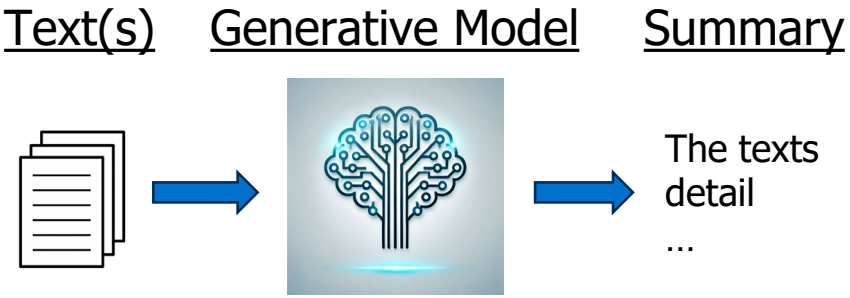
Generative Model



Abstractive Summary

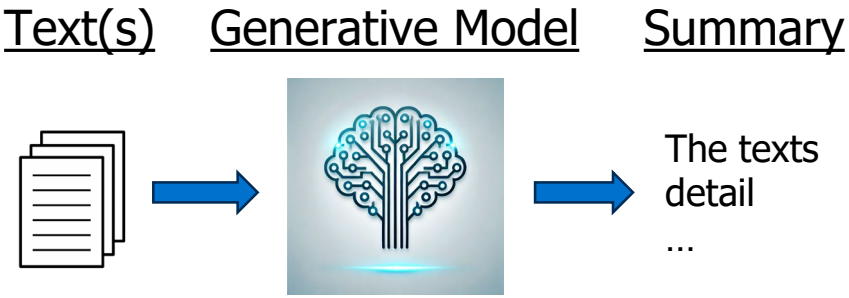
"The company saw a notable profit increase due to higher sales and better cost control."

Applications of Summarization



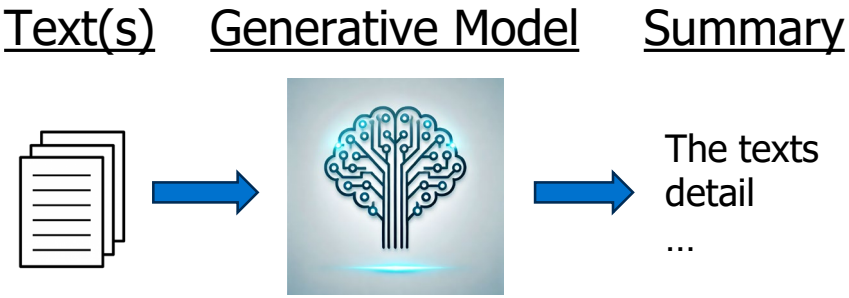
- Extractive summarization use cases:
  - Highlighting key points in legal documents.
  - Extracting headlines from news articles.
  - Extracting exact quotes or key phrases.

Applications of Summarization



- Abstractive summarization use cases:
  - Summarizing customer feedback to generate actionable insights.
  - Creating executive summaries for business reports.
  - News article aggregate summaries that can describe themes and narratives.

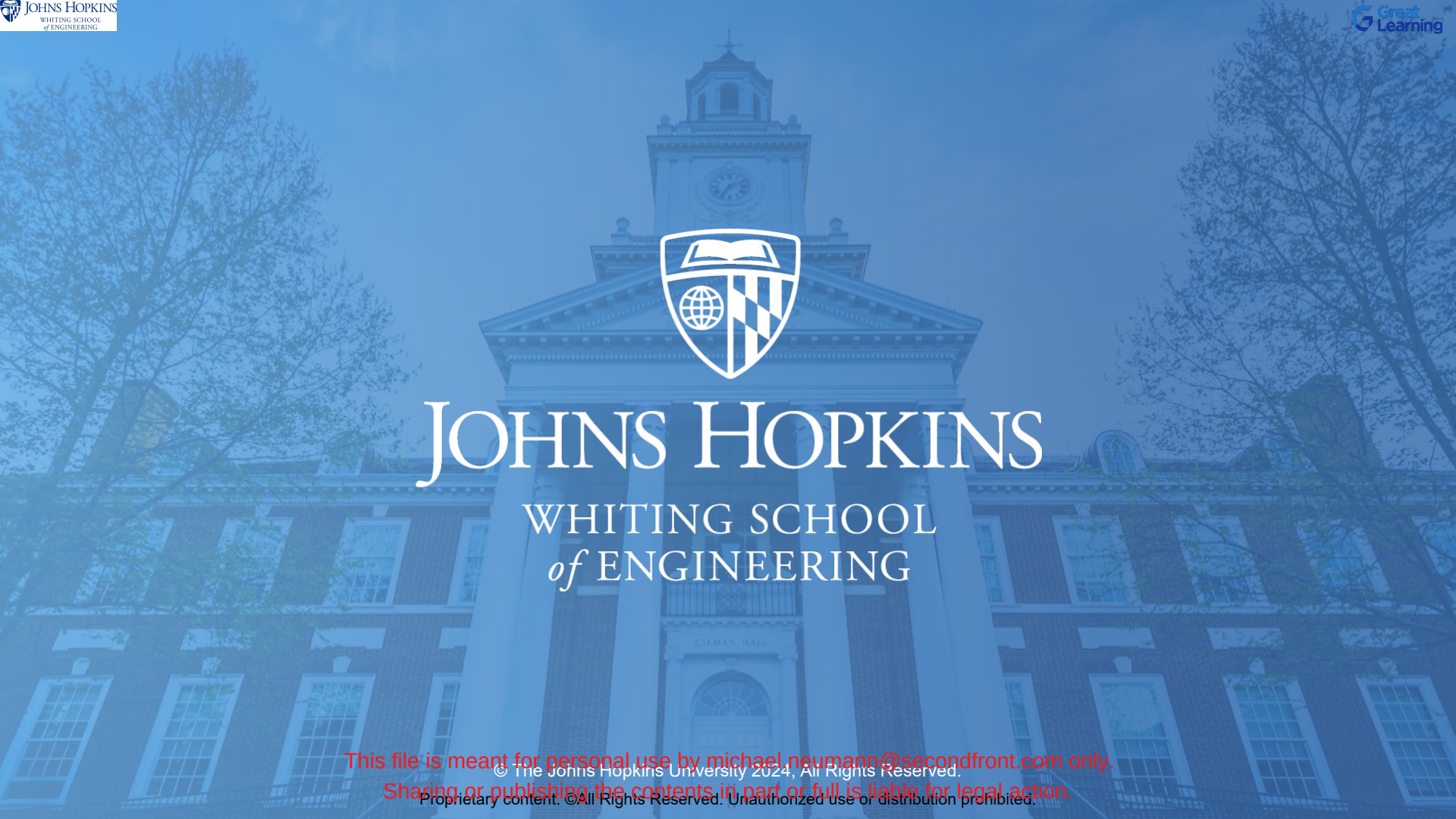
Applications of Summarization



- Abstractive and extractive summarization can be hybridized:
  - They extract parts of texts for certain key phrases or words, and then analyze and summarize that content.
  - They are useful for things like analyzing social media data.

Key Takeaways

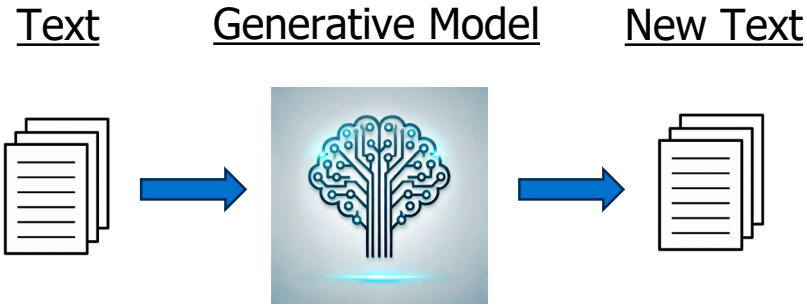
- Summarization is an important task for Generative AI, as it can often greatly enhance the analysis of large corpora of text and save significant time.
- Extractive summarization pulls things directly from the text.
- Abstractive summarization generates new text based on the input text.



Applied Generative AI

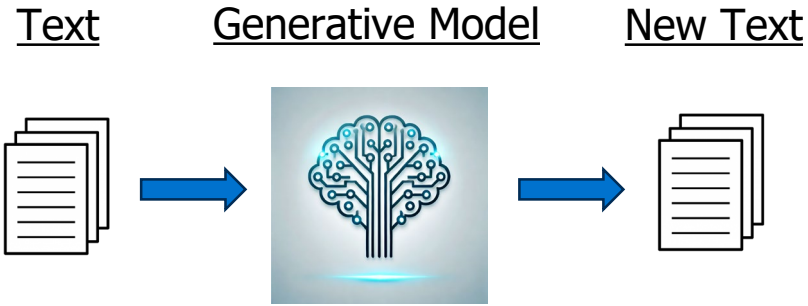
Evaluation Metrics for Text-to-Text Tasks

Why evaluate Text-to-Text outputs?



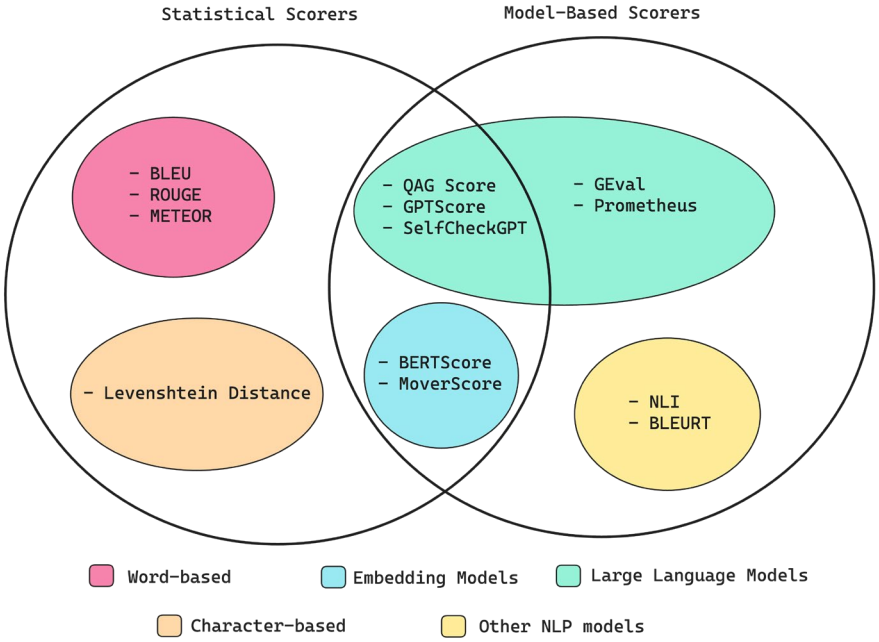
- Gauges how "good" our model's text outputs are.
- **Importance of Evaluation Metrics**:
  - Assess the quality of generated text against expected outputs.
  - Ensure models produce accurate, relevant, and fluent outputs.

Why evaluate Text-to-Text outputs?



- **Challenges in Evaluation**:
  - Text generation is inherently subjective.
  - Multiple valid outputs make defining "correctness" difficult.
  - Many different aspects of text that could be evaluated.

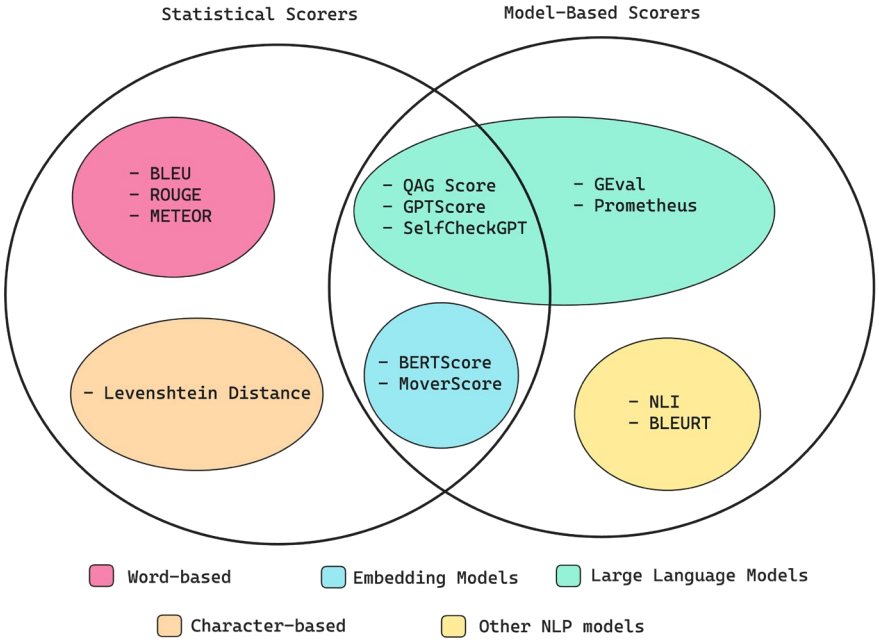
ROUGE: A Key Metric for Summarization



- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
- Measures overlap between generated text and reference text based on n-grams, longest common subsequences, or word overlaps.
- **Variants**:
  - ROUGE-N
  - ROUGE-L

LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

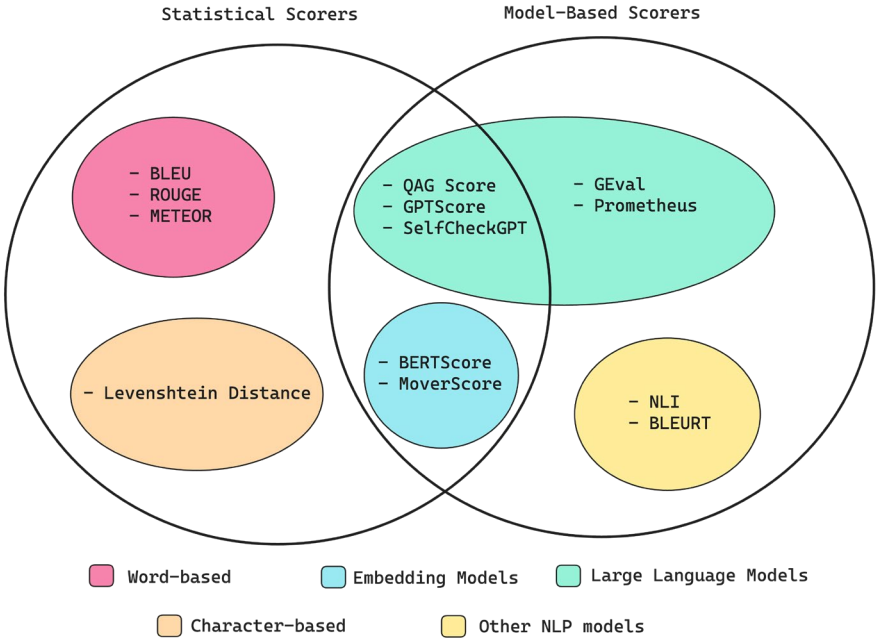
ROUGE: A Key Metric for Summarization



- **Use Cases**: Primarily used for summarization tasks.
- **Strengths**: Captures lexical similarity effectively.
- **Limitations**: Doesn't account for semantic meaning or paraphrasing.

LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

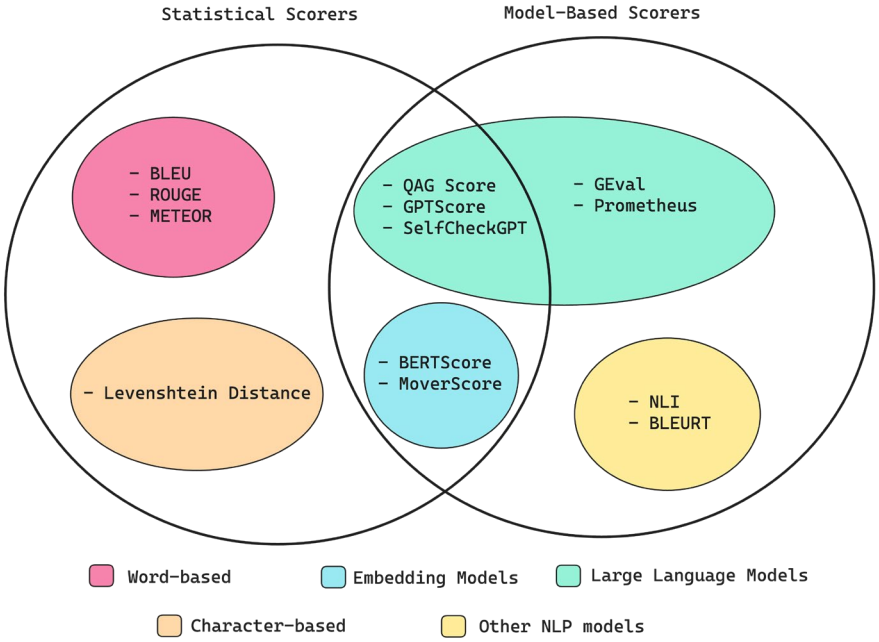
BLEU: Evaluating Translation Precision



- BLEU (Bilingual Evaluation Understudy)
- Measures n-gram precision between generated text and reference text, with a brevity penalty to prevent overgeneration.
- **Variants**:
  - BLEU-1 (unigram), BLEU-2 (bigram), etc.

LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

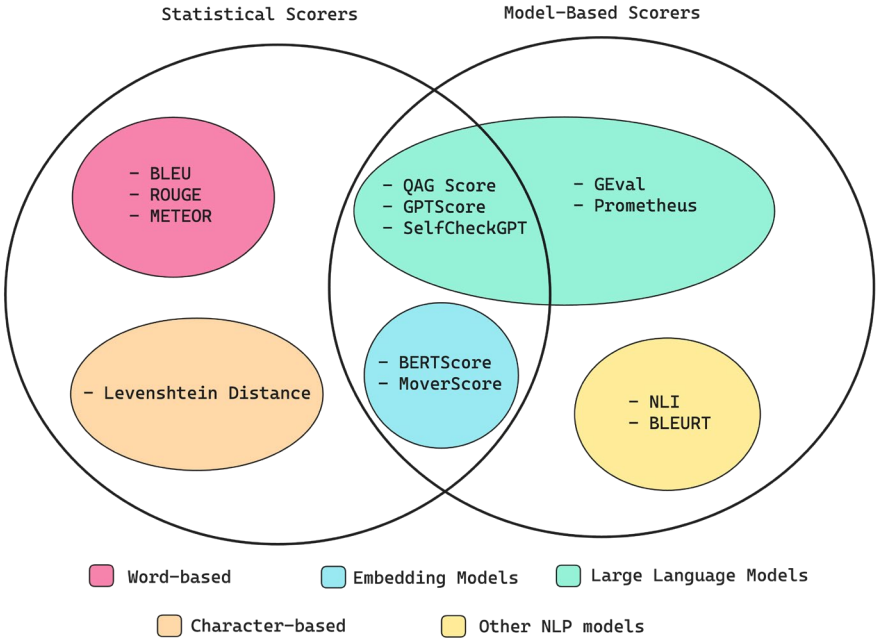
BLEU: Evaluating Translation Precision



- **Use Cases**: Commonly used for machine translation.
- **Strengths**: Simple and computationally efficient.
- **Limitations**: Focuses on precision, ignores recall, and struggles with synonyms or rephrased content.

LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

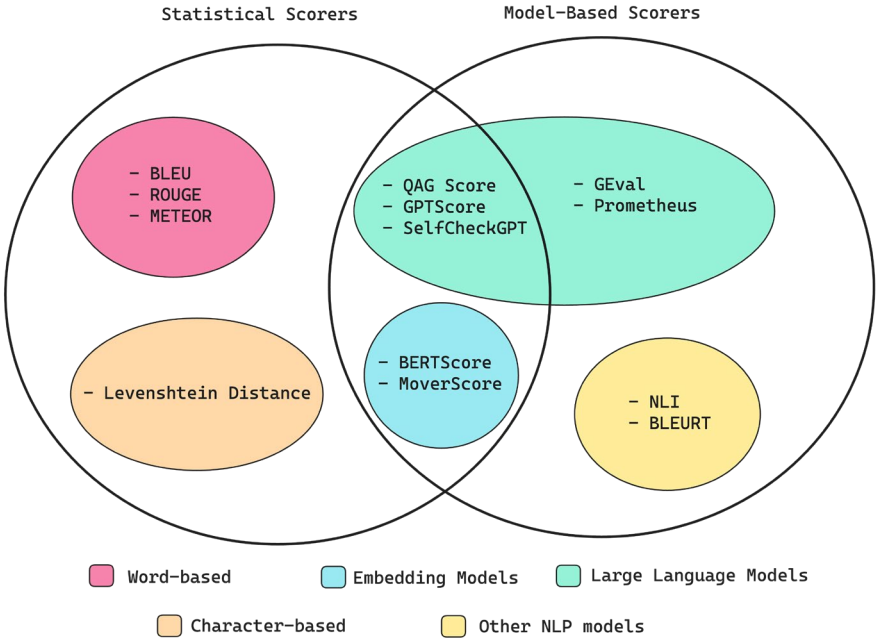
BERTScore: Embedding-Based Evaluation



- Leverages pre-trained language models (e.g., BERT) to compute similarity between embeddings of generated and reference texts.
- Variants by different embedding models.
- Computes cosine similarity of token embeddings instead of exact matches.

LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

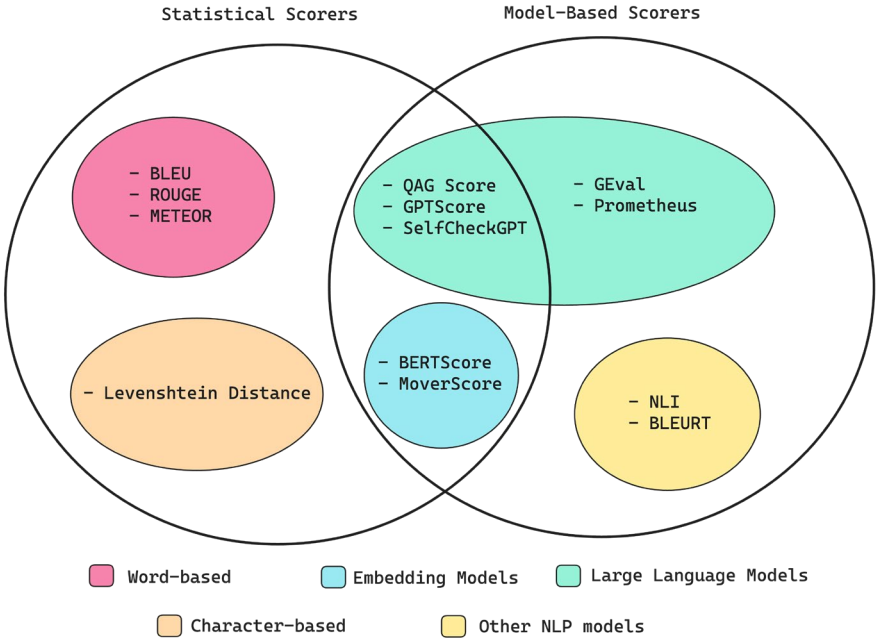
BERTScore: Embedding-Based Evaluation



- **Use Cases**: Effective for tasks requiring semantic understanding, such as abstractive summarization.
- **Strengths**: Accounts for paraphrasing and synonyms.
- **Limitations**: Computationally intensive and relies on quality embedding space.

LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

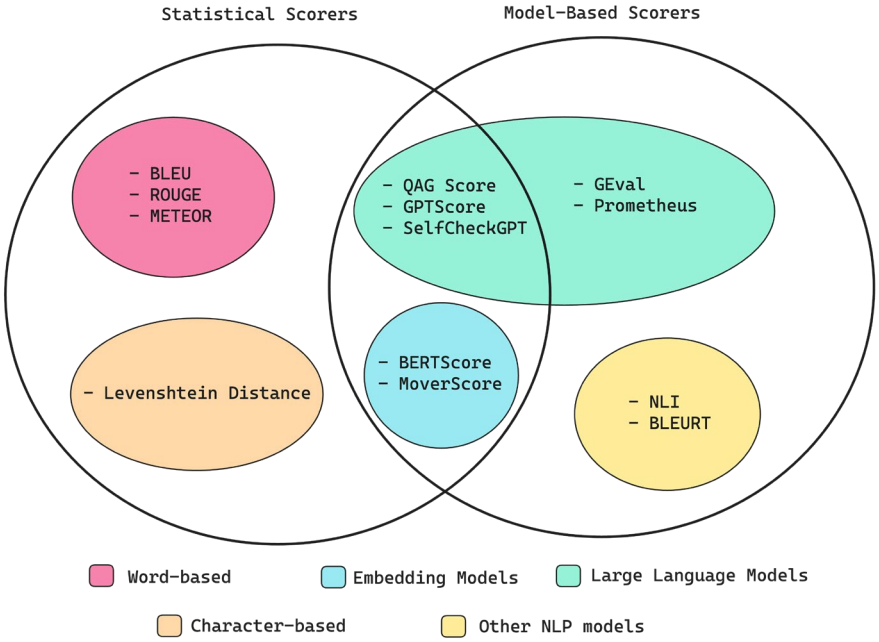
Using LLMs for Subjective Evaluation



- Uses generative models to evaluate outputs based on criteria like fluency, coherence, and relevance.
- Prompts the LLM with evaluation tasks. Example: Rate this summary on a scale of 1-5 for accuracy.
- Can be numerical or qualitative.

LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

Using LLMs for Subjective Evaluation



- **Use Cases**: Used for difficult to measure and subjective textual concepts.
- **Strengths**: Can assess subjective qualities like creativity and contextual relevance.
- **Limitations**: Results may be inconsistent or biased.

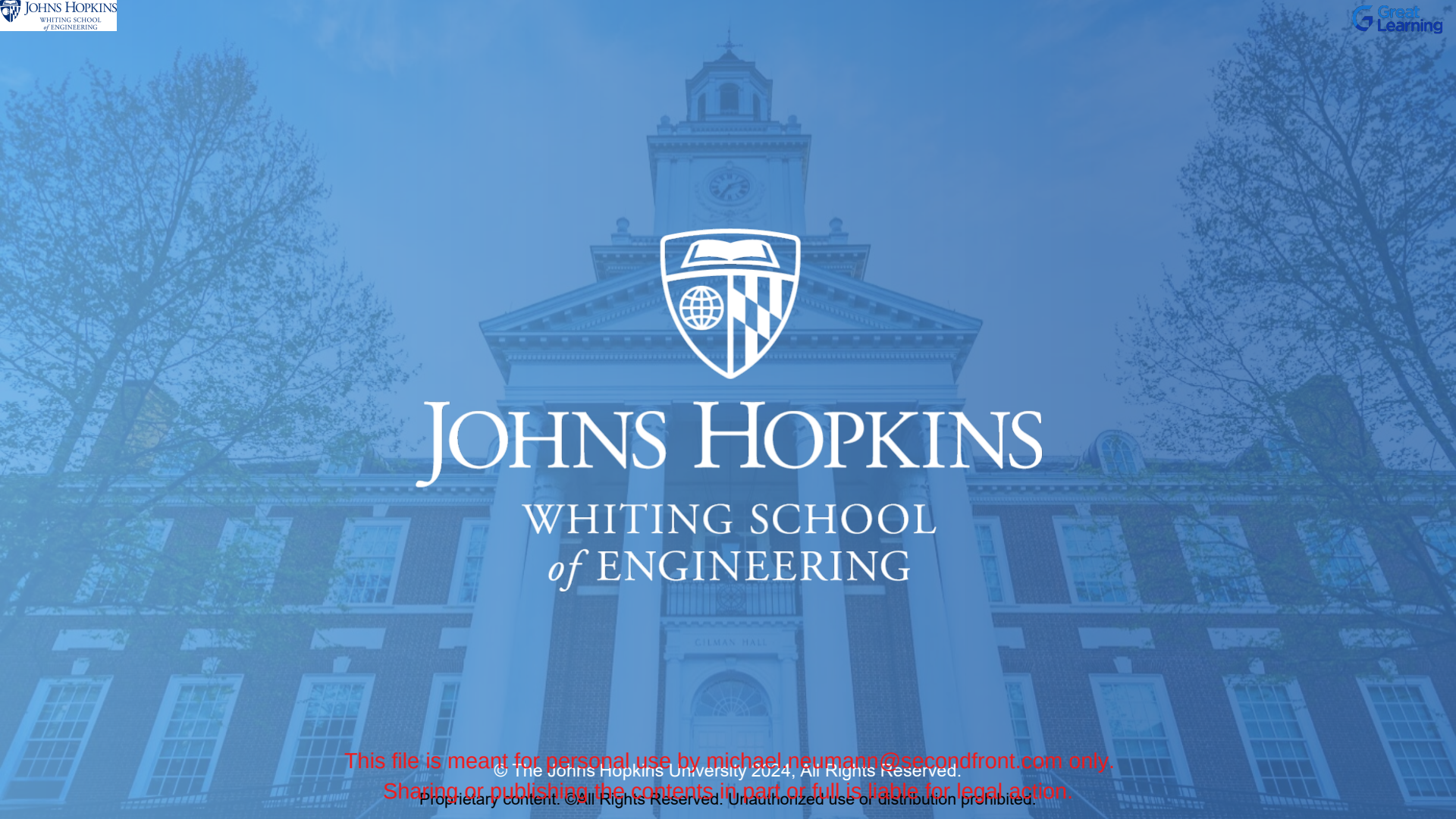
LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. *Confident AI Blog.* https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

Choosing the Right Metrics

- **Metric Suitability for Tasks**:
  - **ROUGE**: Best for extractive summarization.
  - **BLEU**: Ideal for translation and structured tasks.
  - **BERTScore**: Excellent for abstractive summarization and paraphrased outputs.
  - **LLM Evaluators**: Useful for high-level subjective assessments.
- **Best Practices**:
  - Use a combination of metrics to ensure balanced evaluation.
  - Complement automated metrics with human evaluation for subjective tasks.

Key Takeaways

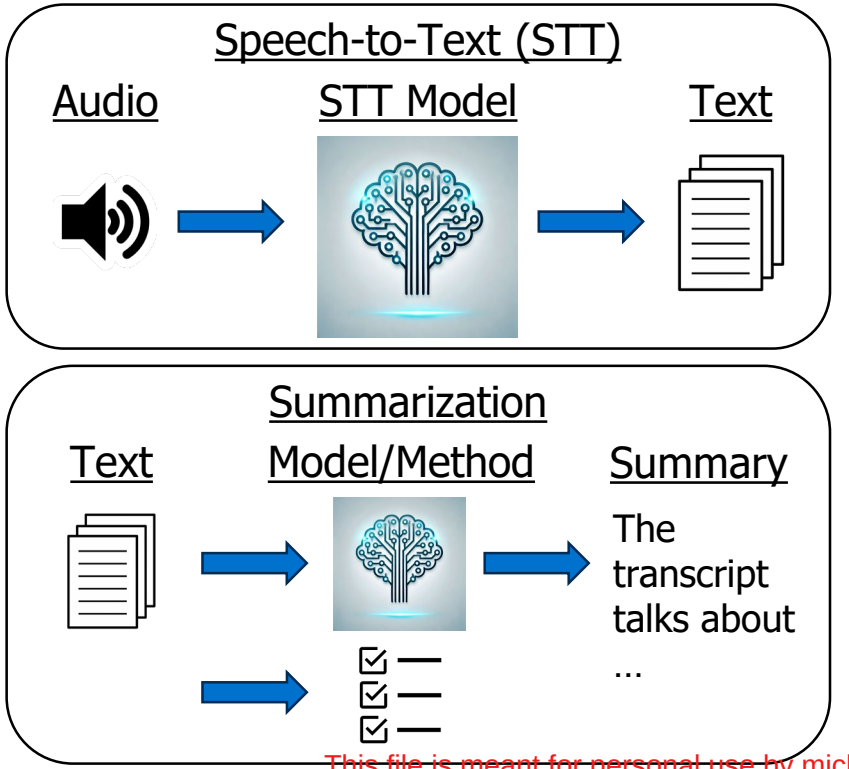
- Much like in text-to-label tasks, evaluating performance is also critical in text-to-text tasks.
- Evaluating output in text-to-text tasks is more difficult due to the nature of text output.
- It is often suitable to adopt a number of evaluation metrics to evaluate text outputs.



Applied Generative AI

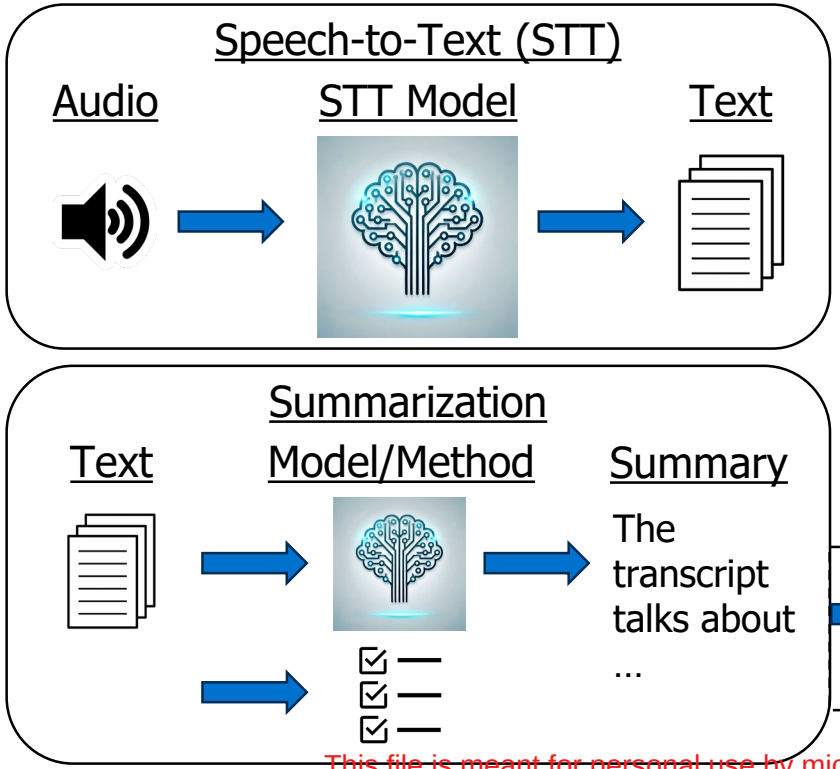
Speech-to-Text and Summarization Workflow

Combining Speech-to-Text and Summarization



- There are many examples of real-world workflows where we want to combine Speech-to-Text with text summarization for:
  - Creating meeting notes from discussions.
  - Generating insights from customer support calls or logs.

End-to-End Process for Speech-to-Text and Summarization



1. Speech-to-Text (STT) Conversion
2. Text Preprocessing:
   - Clean and structure transcriptions for consistency
3. Summarization:
   - Abstractive and/or extractive
4. Evaluation of Summarization Output

Key Takeaways

- Combining Speech-to-Text (STT) and summarization boosts accessibility and productivity.
- Practical applications span business, education, and content creation.
- As with any other task using Generative AI, it is usually a good idea to have some evaluation data points and engage in iterative refinement with things like prompts.