TEXT GENERATION FOR STORY COMPLETION

DreamWeaver AI

BEN AND KHAL

What is DreamWeaver?

HOW DOES IT WORK?

AI Model

Context Management

Generated Text Evaluation

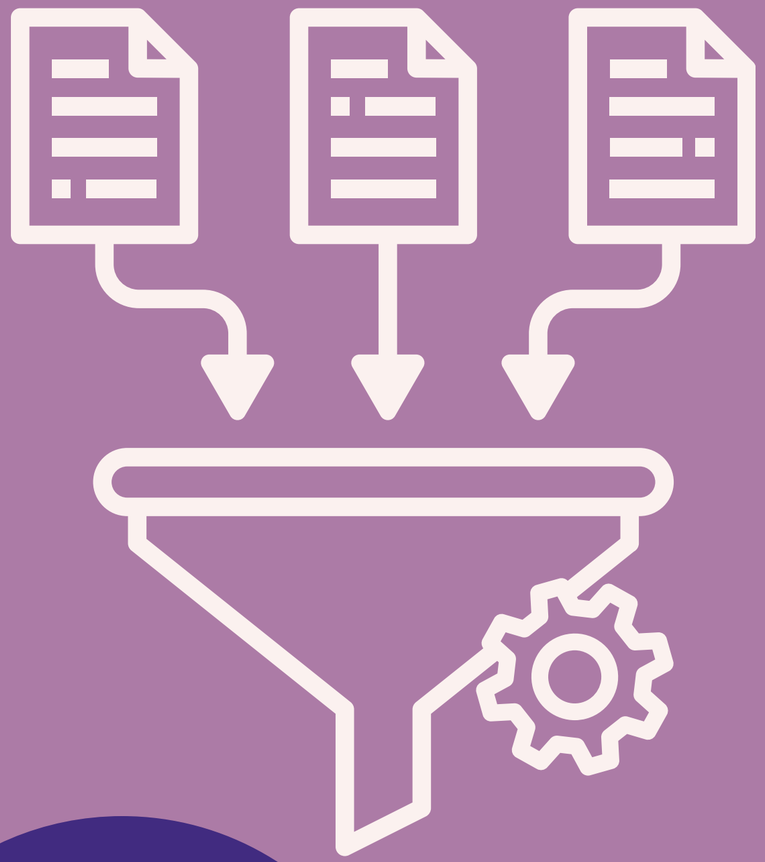
The Product

AI MODEL

Based from

Mistral-7B-Instruct

MISTRAL AI



Context Management

WHAT IS EXACTLY CONTEXT?

Context

/ˈkɒntɛkst/

noun

The circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.

Flowchart

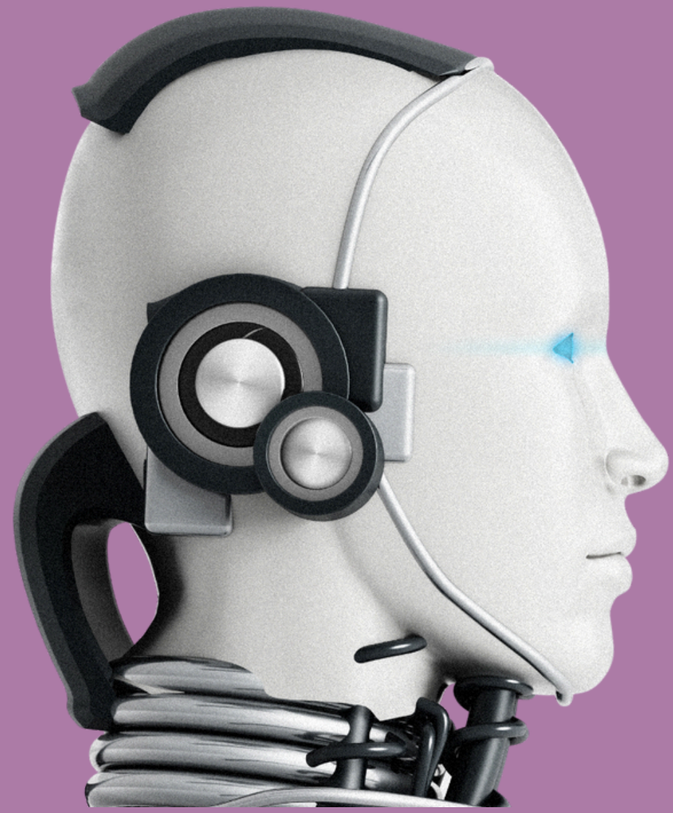
SIMPLE.

MISTRAL 7B

TROUBLED

Confused, mixed up, no context awareness.

GPT-4



NOT (QUITE) TROUBLED

Aware, not confused, has somewhat context awareness

GENERATED TEXT EVALUATION

Using NLTK, BLEU and SpaCy

NLTK

NATURAL LANGUAGE TOOLKIT

- **Tokenization**: cutting up a sentence or paragraph into smaller pieces. For example, you can break it into words or sentences so the computer can understand and work with them.
- **Stemming and Lemmatization**: Both of these are ways to simplify words by reducing them to their "core" or "base" form.
  - **Stemming**: It chops off the ends of words. It’s not always perfect but works fast.
  - **Lemmatization**: It’s smarter and uses a dictionary to find the proper root form of a word.
- **Text Classification**: This is like sorting text into categories or labels. For example, you could teach a program to decide if a piece of text is about sports, news, or entertainment.

BLEU

BILINGUAL EVALUATION UNDERSTUDY

Check how good a computer's generated text is compared to a human's text.

- N-gram Matching: Think of "N-grams" as small chunks of words.
  - A 1-gram is just one word, like "cat".
  - A 2-gram is two words in a row, like "the cat".
  - A 3-gram is three words in a row, like "the cat jumped".

BLEU checks how many chunks of words (like 2-grams or 3-grams) match between the computer's text and the human's text. Longer matches mean the computer's sentences are more natural.

For example:

Human: "The cat jumped over the wall."

Computer: "The cat jumped the fence."

BLEU compares pairs like "The cat" and "cat jumped" to measure similarity.

SpaCy

- Tokenization
- **Part-of-Speech (POS) Tagging:** This labels each word with its role in the sentence, like whether it’s a noun, verb, or adjective.
- **Named Entity Recognition (NER)**: It finds specific names in the text and figures out what they are, like a person, place, or company.
- **Dependency Parsing**: It analyzes how words are connected in a sentence, showing relationships like which word is the subject and which is the object.
- **Word Embeddings:** This is a way to represent words as numbers (vectors) in a mathematical space. Words with similar meanings are placed closer together, which helps the computer understand relationships between words.

Vectorization: Representing Text as Numbers

Converts text into numerical representations (vectors) for machine processing.

Techniques

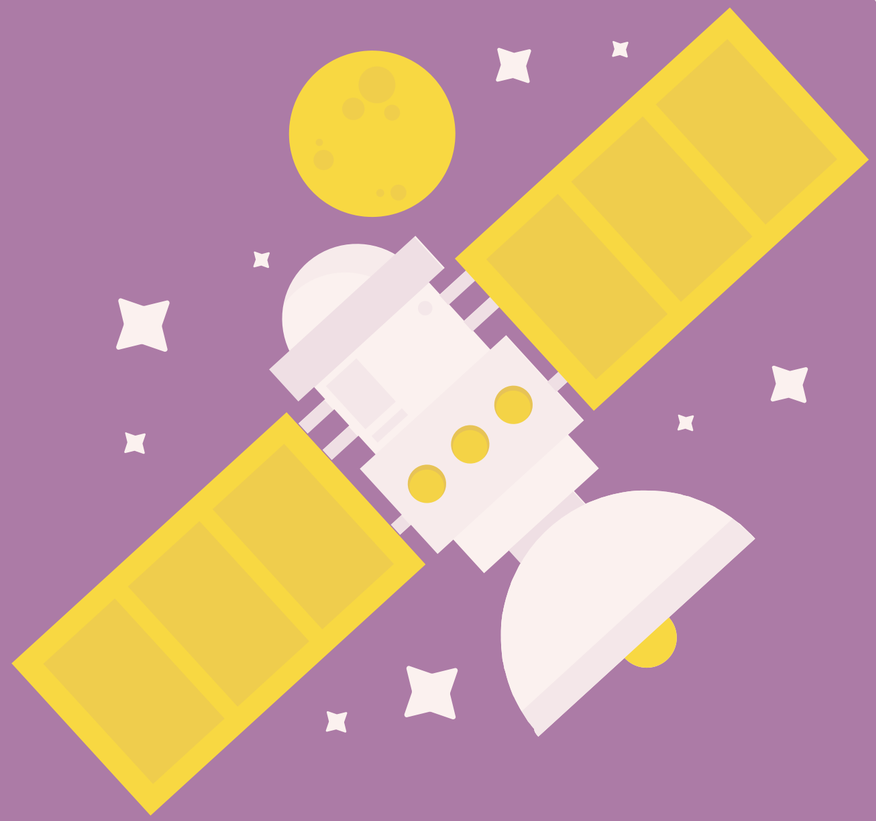
One-Hot Encoding: Binary representation for each word.

Word Embeddings: Dense vectors capturing word meanings (e.g., Word2Vec, GloVe, BERT).

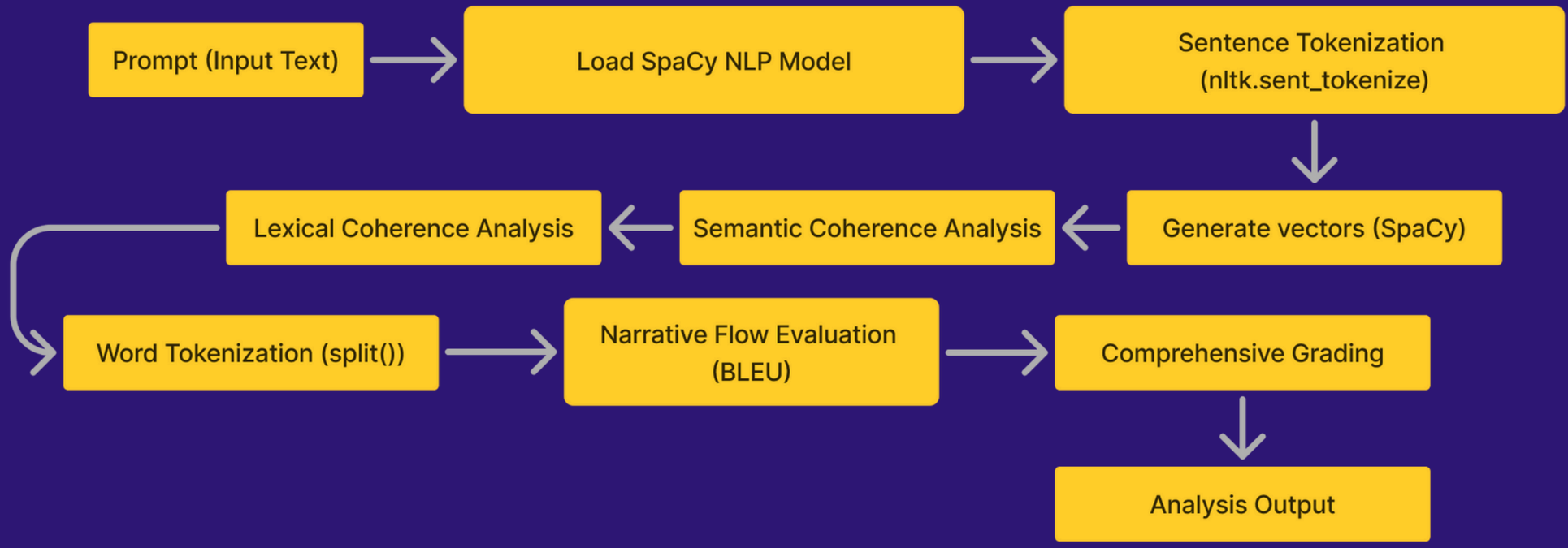
Why?

- **Machine Learning Compatibility**: Converts text for model training.
- **Similarity Measurement**: Compares text using vector distances.
- **Feature Engineering**: Generates features for models.

Let’s try it out !!



Comprehensive Analysis Flowchart



COHERENCE TEST

breaks down the story's coherence into different categories like meaning, word choice, and overall flow, then organizes and displays the results.

OUTPUT

- Overall Coherence Score: A number representing how coherent the story is overall.
- Semantic Coherence: Scores or details about the meaning connections in the story.
- Lexical Coherence: Scores for the consistency of words and phrasing.
- Narrative Flow: Scores for how logically the story unfolds.

BERT SCORE

Other than the makeshift comprehensive test, lets also apply BERT score testing to find out deeper in F1 level connections between prompt and the result.

OUTPUT

- Precision: This measures how accurate the model's predictions are. In the context of text generation, it tells you what proportion of the words predicted by the model are actually correct.
- Recall: This measures how complete the model's predictions are. It tells you what proportion of the correct words were actually predicted by the model.
- F1-Score: This is a harmonic mean of precision and recall. It provides a balance between the two, giving more weight to the lower-scoring metric.

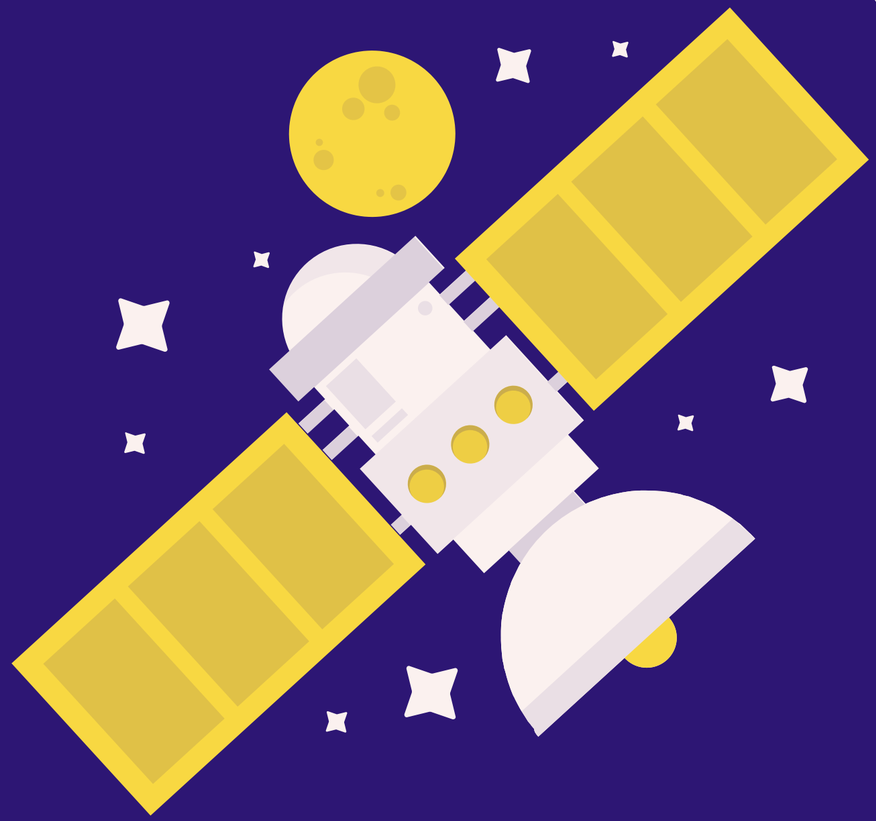
Final Product

HTTPS://DREAMWEAVERAI.STREAMLIT.APP

the difference

- instead of using the same LLM (mistral 7b) we are using google gemini's 1.5-flash model.
- it can generate a picture alongside the prompt that you feed it!

Let’s try it out !! (again)



REFERENCES

Marchenko, O. O., Radyvonenko, O. S., Ignatova, T. S., Titarchuk, P. V., & Zhelezniakov, D. V. (2020). Improving text generation through introducing coherence metrics. Cybernetics and Systems Analysis, 56(1), 13-21. https://doi.org/10.1007/s10559-020-00216-x

Zhao, W., Strube, M., & Eger, S. (2023). DiscoScore: Evaluating text generation with Bert and discourse coherence. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.278

Claude Sonnet 3.5

Gemini

Thank you for your time.