

Enhancing Connectome Gene Summaries with Generative AI



Jian Yun Zhuang, Vincent Lau, Nicholas Provart



ABSTRACT:

Mutwil's Lab at Nanyang Technological University developed a web-application called the Plant Connectome which processed **101341** publications with GPT's assistance to provide insights into **387777** relationships involving genes, molecules, compartments, stresses, organs, and other plant entities. This project uses Facebook's Llama 2, a powerful large language model, to convert complex and technical gene summaries into coherent, easily readable paragraphs. This approach addresses the challenge of comprehending the dense, note-like format of the original summaries, which often hinders quick understanding and accessibility. Furthermore, the General Agricultural Intelligent Agent (GAIA), developed at the University of Toronto, is designed to enhance research efficiency by aggregating data across many scientific resources and simplifying complex genetic data for plant biology. The anticipated incorporation of the Plant Connectome's capabilities into GAIA aims to significantly boost GAIA's functionality, helping with research efficiency and data accessibility in the field.

INTRODUCTION:

The General Agricultural Intelligent Agent (GAIA), developed in response to the plant biology community's demand for centralized data access, represents a groundbreaking approach. Hosted by the University of Toronto, GAIA aims to aggregate and interpret diverse scientific data through Natural Language Processing (NLP) and machine reading (MR) technologies, offering researchers streamlined summaries and answers to intricate queries.

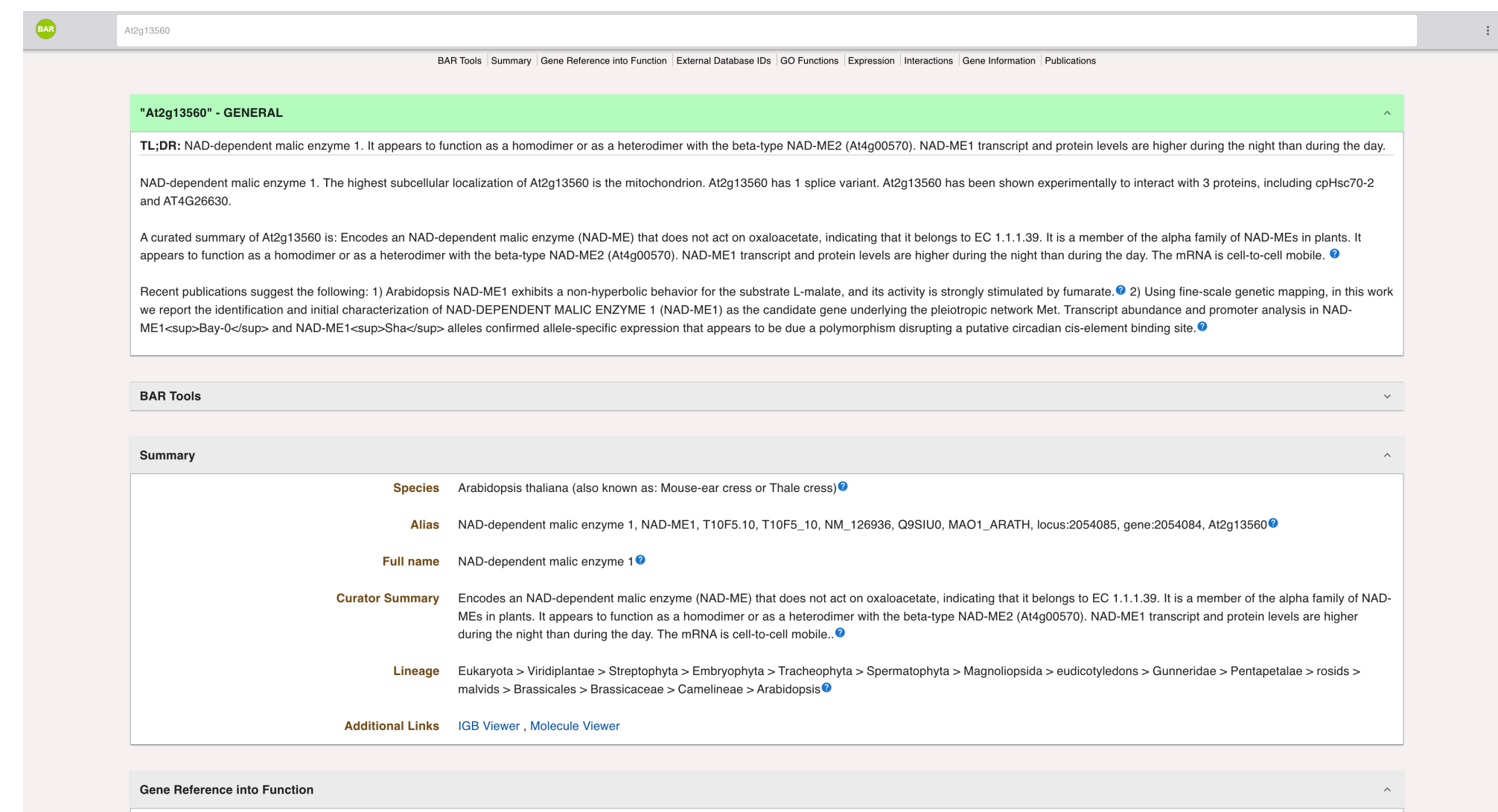


Figure 1: A representation of what the current GAIA BAR tool looks like when you search up for an AGI gene. The AGI Gene Identifier being searched here is AT3g24650'.

To enhance the GAIA platform's utility, we plan to introduce a feature that provides insights on publications, a capability not currently available within GAIA. This new addition aims to present users with concise, readable summaries derived from the Plant Connectome. However, the current text summary of Plant Connectome contains dense notes that pose significant comprehension challenges.

```
"text_summary": "NAD-ME1 AFFECTS PLANT METABOLISM (33523525). NAD-ME1 LOCALIZES TO MET.II.15 QTL (33523525). NAD-ME1 INTERACTS WITH CIRCADIAN CIS-ELEMENT BINDING SITE (33523525). NAD-ME1 ENHANCES TRICARBOXYLIC ACID CYCLE INTERMEDIATES (33523525). NAD-ME1 REPRESSIONS GSL BIOSYNTHESIS PATHWAY GENES (33523525). NAD-ME1 PRODUCES PRIMARY AND SECONDARY METABOLITES (33523525).",
```

Figure 2: A representation of what the Text Summary Section Looks like for AT2g13560 of the Plant Connectome Tool. The Text Summary section will be used to as an input for our Connectome Gene Summary Enhancer

The Connectome Gene Summary Enhancer, employing Facebook's Llama 2 LLM model, is designed to address this by converting complex summaries into understandable narratives.

```
Outout from llama: The gene AT2G13560, also known as NAD-ME1, plays a crucial role in plant metabolism. Studies have shown that NAD-ME1 localizes to the metabolic pathway II.15 QTL (PubMed ID: 33523525) and interacts with the circadian cis-element binding site (PubMed ID: 33523525). NAD-ME1 enhances the production of tricarboxylic acid cycle intermediates (PubMed ID: 33523525) and represses the expression of genes involved in the glycolysis and starch synthesis pathways (PubMed ID: 33523525). Additionally, NAD-ME1 produces primary and secondary metabolites, such as flavonoids and phenolic acids, which are important for plant defense and stress responses (PubMed ID: 33523525). Overall, the function of NAD-ME1 in plant metabolism is to regulate the balance between different metabolic pathways and to produce metabolites that are essential for plant growth and development.
```

Figure 3 Shows a Llama2-generated summary for AT2g13560, which achieved a 98% BERT entailment score, indicating high coherence and accuracy.

ACKNOWLEDGMENTS:

I would first like to express my deepest gratitude to Dr. Nicholas Provart, my project supervisor. Without his guidance and support, this independent research project would not have been possible. Special thanks to Vincent Lau, who offered me invaluable guidance on how to approach the project and assisted in debugging when setting up Llama 2 on Compute Canada. I am also thankful to Mutwil's Lab for developing the Plant Connectome and for providing us with instructions on how to access the endpoint.

Materials & Methods

Leveraged Facebook's Llama 2 large language model and Compute Canada's computing resources to interpret complex genetic data from the Plant Connectome. Employed Python to extract and structure over 8,000 gene summaries into JSON format, with shell scripts automating data retrieval, processing, and job management workflows. Conducted iterative prompt engineering to optimize LLaMA 2's summarization capabilities, significantly improving the accuracy and relevance of generated text summaries for detailed genetic information.

The **Llama 2 - Codellama 13b-Instruct Model** was selected due to its superior performance and ability to handle longer input sequences. The 13B in the model's name refers to the 13 billion parameters it contains, which generally correlates with increased model capacity and performance. Parameters are like knobs that make tiny adjustments to the connections in the neural network, allowing it to learn and generate human-like text. The numerical values that determine how the model processes language.

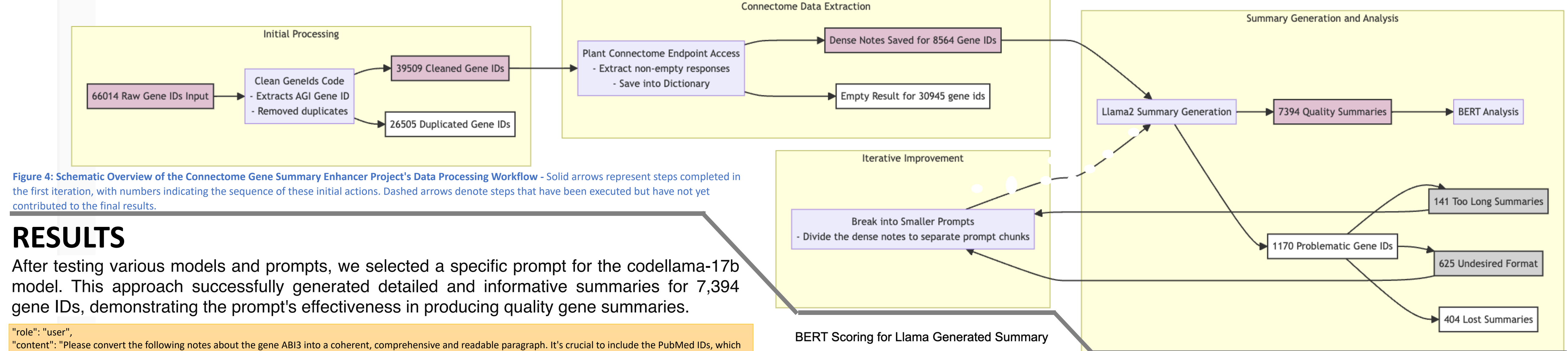


Figure 4: Schematic Overview of the Connectome Gene Summary Enhancer Project's Data Processing Workflow - Solid arrows represent steps completed in the first iteration, with numbers indicating the sequence of these initial actions. Dashed arrows denote steps that have been executed but have not yet contributed to the final results.

RESULTS

After testing various models and prompts, we selected a specific prompt for the codellama-17b model. This approach successfully generated detailed and informative summaries for 7,394 gene IDs, demonstrating the prompt's effectiveness in producing quality gene summaries.

```
"role": "user",
"content": "Please convert the following notes about the gene AB13 into a coherent, comprehensive and readable paragraph. It's crucial to include the PubMed IDs, which are the numbers in the brackets, within the paragraph. These IDs provide essential references and should not be omitted. For example, you can mention studies or findings followed by their respective PubMed ID. Here are the notes: AB13 MAINTAINS EMBRYO DEVELOPMENT (10743655), SSP ACCUMULATION (15695450), AT253 (15695463), CRC (15695463), PLANT EMBRYO DEVELOPMENT (17158584), MIR159 (17217461), H5FA9 (17220197), ABA-INDUCED ARREST (18278579), STORAGE PROTEIN SYNTHESIS (18701524), WRKY2 (19622176), TWO MAJOR STAGES IN EMBRYO MATURATION (19659659), LEA PROTEINS (24043848), SEED DEVELOPMENT (24388521, 29475938), SEED MATURATION (24473899, 28346448, 35318532), PROTEIN RESERVES (25840088), ABA SIGNALING (26496910)."
```

```
"role": "assistant",
"content": "The gene AB13 plays a pivotal role in maintaining embryo development, as evidenced by research documented in PubMed ID 10743655. It is also involved in the accumulation of SSP (PubMed ID 15695450) and influences various processes such as AT253 and CRC (PubMed IDs 15695463), plant embryo development (PubMed ID 17158584), and the regulation of MIR159 (PubMed ID 17217461). Further, AB13 is integral to H5FA9 mechanisms (PubMed ID 17220197), ABA-induced arrest (PubMed ID 18278579), and storage protein synthesis (PubMed ID 18701524). It interacts with WRKY2 (PubMed ID 19622176) and is crucial in two major stages of embryo maturation (PubMed ID 19659659), LEA protein production (PubMed ID 24043848), and seed development (PubMed IDs 24388521, 29475938). The gene's role extends to seed maturation (PubMed IDs 24473899, 28346448, 35318532), protein reserve synthesis (PubMed ID 25840088), and ABA signaling pathways (PubMed ID 26496910)."
```

```
"role": "user",
"content": "Please convert the following notes about the gene " + gene_id + " into a more readable and comprehensive paragraph like the given example. Remember to include all numbers in the brackets, as these PubMed IDs are essential for referencing the studies. Also include all given details. Here are the notes: " + text_summary
```

```
Figure 6: Final Prompt in Llama 2 for Gene Summaries - This figure highlights the prompt used in Llama 2's llama 17b model to convert gene notes into detailed paragraphs with PubMed IDs. Starting with an AB13 gene example, it shows the synthesis of concise notes into a narrative, linked with relevant PubMed IDs. The prompt then guides the creation of similar summaries for other gene_ids, ensuring integration of all details and PubMed IDs into coherent paragraphs.
```

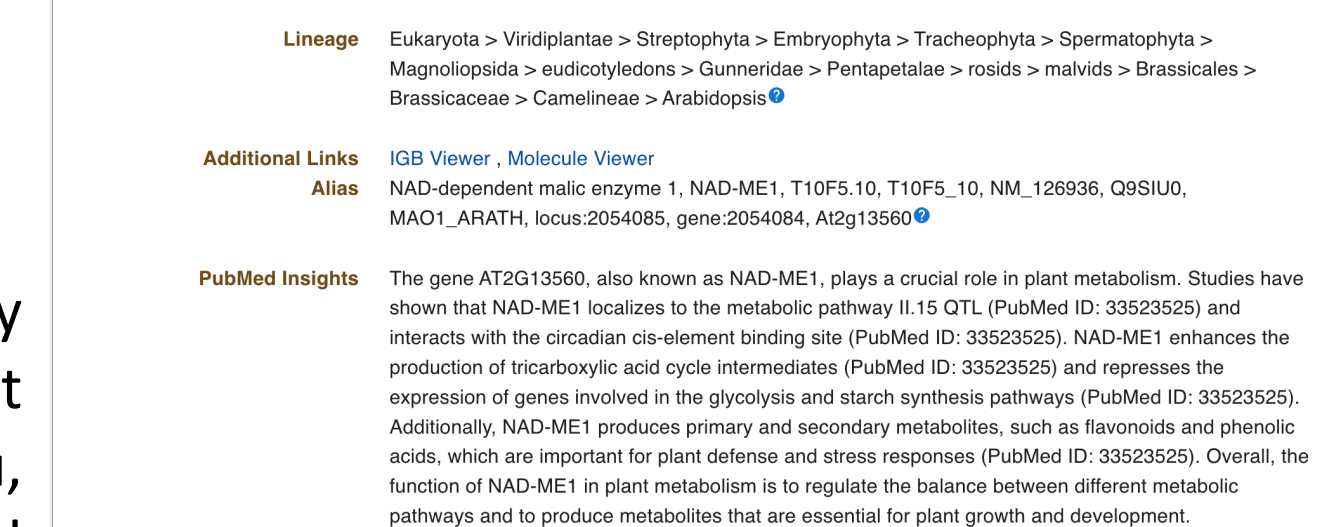


Figure 7: New GAIA BAR Tool Website Interface - This figure previews the design on the gene search Summary functionality for AT2g13560. It features a new section for "PubMed Insights", where outputs generated by Llama, utilizing Connectome data, is showcased.

DISCUSSION

Llama 2 Limitations

- Inconsistent Gene Naming:** Struggles with the variability in biological nomenclature, impacting summary accuracy.
- Repetitive Summaries:** Tendency to repeat sentences, reducing content diversity and utility.
- Extrapolated Information:** Llama 2 may generate additional information that expands on the given text summary. While this feature aims to enrich the content, it also presents a challenge in verifying the accuracy and reliability of these extrapolations.

Proposed Approach: LLM Reasoner for Enhanced Reasoning

- Specialize LLM models for logical inference within NLP through fine-tuning and architectural modifications.
 - Leverage Llama 2 generated paragraphs as a starting point for targeted domain adaptation.
 - Address gene naming inconsistencies and improve summary coherence encountered with Llama 2.
 - Produce more coherent and logically reasoned summaries, enhancing readability and accuracy.
- Fine-Tuning BioBERT for Biomedical Summarization
- Adapt BioBERT, a biomedical-focused BERT model, for better context recognition in gene summaries.
 - Fine-tune BioBERT on domain-specific datasets, utilizing Llama-generated paragraphs as a source.
 - Expected Outcome: More precise, informative summaries tailored to the complexities of biomedical texts.

REFERENCES:

Fo, K. and Mutwil, M. (2023) 'PlantConnectome: knowledge networks encompassing >100,000 plant article abstracts', bioRxiv (Cold Spring Harbor Laboratory). Available at: <https://doi.org/10.1101/2023.07.11.548541>.

Hao, S., Gu, Y., Ma, H., Hong, J., Wang, Z., Wang, D. and Hu, Z. (2023) 'Reasoning with Language Model is Planning with World Model'. Available at: <https://doi.org/10.18653/v1/2023.emnlp-main.507>.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J. (2019) 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', Bioinformatics, 36(4). Available at: <https://doi.org/10.1093/bioinformatics/btz262>.

Sullivan, A.J. (2021) GAIA: An 'Ecosystem' of Aggregated Information for Plant Biology. Available at: <https://hdl.handle.net/1807/10879>.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., et al. (2023) 'Llama 2: Open Foundation and Fine-Tuned Chat Models', arXiv.org, 19 July. Available at: <https://arxiv.org/abs/2307.09288>.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (No date) 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding', ACL Anthology. Available at: <https://aclanthology.org/N19-1423/> (Accessed: 26 March 2024).

Set Up & Installation

- ssh to your cedar.computecanada.ca account
- clone the github repo for llama2-codellama

```
git clone https://github.com/meta-llama/codellama.git
```
- Follow the instruction for Codellama, visit the Meta website <https://llama.meta.com/llama-downloads/> and register to download the **codellama 13b-Instruct model**
- Clone this repo, move the script folder parallel to the `example_chat_completion.py` (You don't need to worry about the dependencies for the llama model, the setup is already included in the script)

```
# Clone the repository
git clone https://github.com/emma925/Genes-Summary-Enhancer.git
# Navigate to the project's scripts directory
cd Genes-Summary-Enhancer/scripts
```

Figure 5: Steps on the install and setup for this project. The GitHub repo is at <https://github.com/emma925/Genes-Summary-Enhancer/blob/main/README.md>

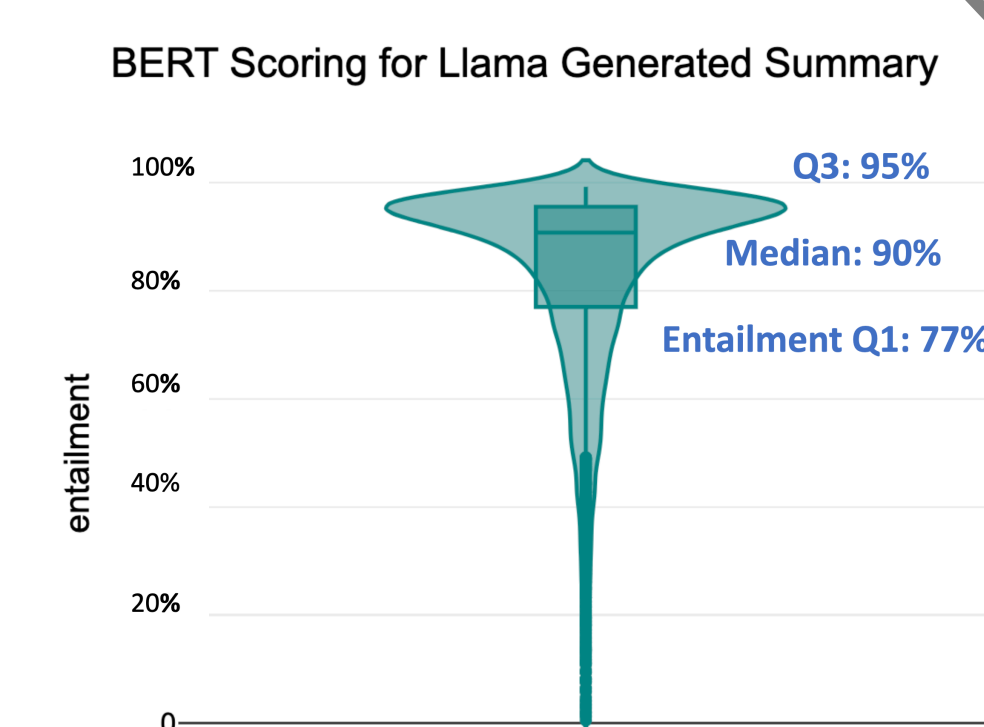


Figure 8: BERT Entailment Scores Violin Plot - This figure shows the distribution of BERT model entailment scores, comparing Connectome project's original summaries with those from Llama2. Entailment in NLP assesses if one text logically implies another.

BERT is an advanced NLP model that comprehends context in text bidirectionally, enhancing understanding and performance in tasks like sentiment analysis and question answering. In this context, entailment measures if the content of Llama2 summaries logically follows the original Connectome project summaries, assessing their coherence and accuracy.

- PLLaMA**
 - Collected all the articles in the S2ORC [Lo et al., 2020] dataset by either satisfying including 'Plant' in the article publisher or the publisher falling in the 750 journals. As a result, they obtained 1,676,389 plant science-related articles.
 - An extension of LLaMa-2, incorporating extensive plant science literature to bridge the knowledge gap in specialized domains.