

Menu:

Upload your PDF Files and Click on the Submit & Process Button

Drag and drop files here

Limit 200MB per file

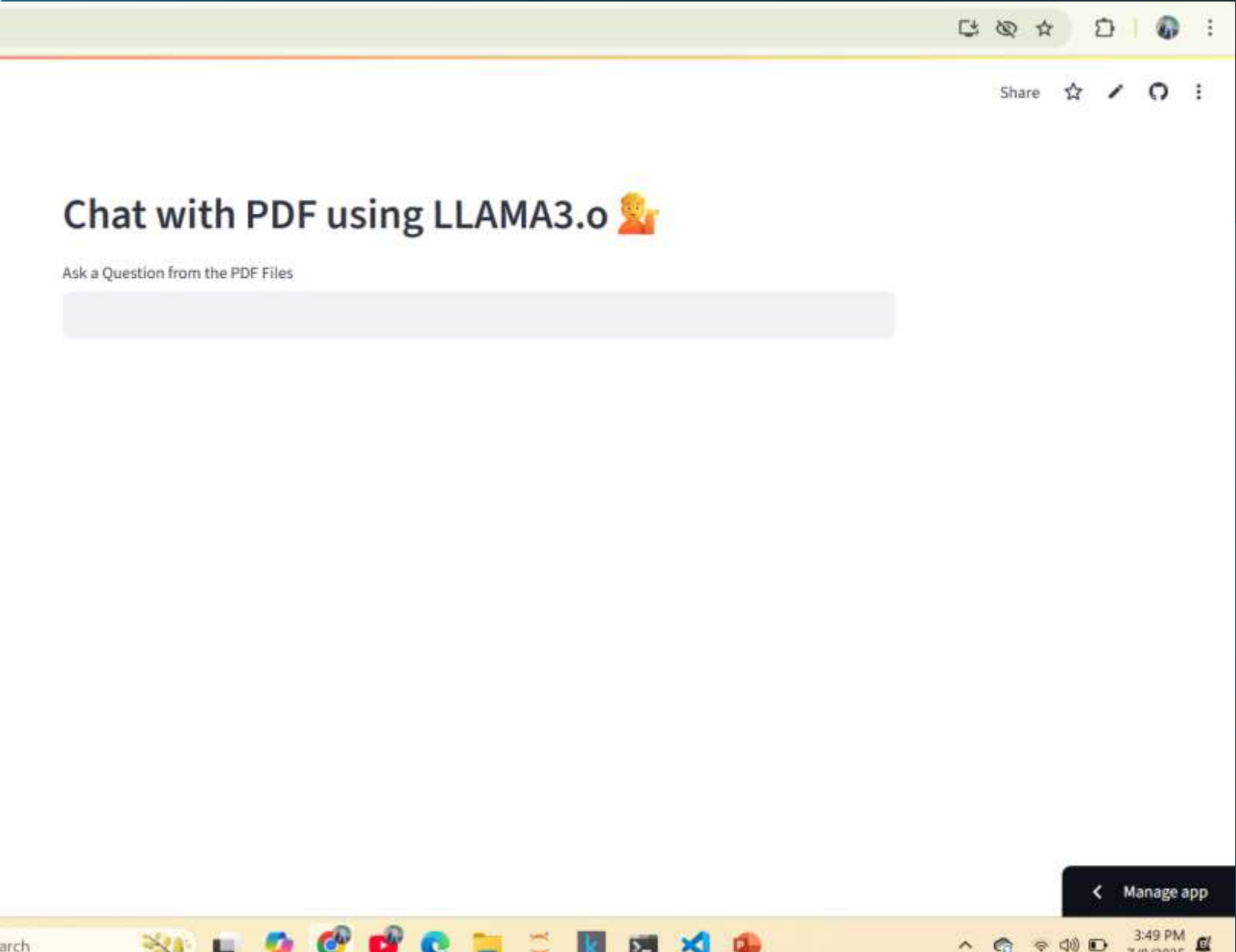
Browse files

Submit & Process

Chat with PDF using LLAMA3.o 🧑

Ask a Question from the PDF Files

< Manage app



Ask Question to know from PDF

Menu:

Upload your PDF Files and Click on the Submit & Process Button

Drag and drop files here
Limit 200MB per file

Browse files

attention_is...
2.1MB

Submit & Process

Done

Chat with PDF using LLAMA3.o

Ask a Question from the PDF Files

what is the formula for scalled dot product?

Press Enter to apply

Manage app

92°F
Weather.com

Search

3:51 PM
7/8/2025

Response Generated

← → ↻ 📄 chatbotpdf.streamlit.app

Share ☆ ✎ ↺ ⋮

Menu:

Upload your PDF Files and Click on the Submit & Process Button

Drag and drop files here
Limit 200MB per file

Browse files

📄 attention_is...
2.1MB

×

Submit & Process

Chat with PDF using LLAMA3.o 🧑🏻

Ask a Question from the PDF Files

what is the formula for scalled dot product?

Reply: $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$

Where Q = queries K = keys V = values d_k = dimension of K

In scaled dot product attention,

$QK^T / \sqrt{d_k}$ is computed and then a softmax is applied to get weights, then multiply these weights with V .

Manage app

92°F
Monday, August 14

Search

3:52 PM
7/8/2025

Sample Question 2

← → ↺ chatbotpdf.streamlit.app

Share ☆ ✎ ↺ ⋮

Menu:

Upload your PDF Files and Click on the Submit & Process Button

Drag and drop files here
Limit 200MB per file

Browse files

attention_is...
2.1MB

Submit & Process

Chat with PDF using LLAMA3.o 🧑

Ask a Question from the PDF Files

what are the llm models used to compare in the pdf

Reply: The models used for comparison are not explicitly stated as "llm" in the provided context. However, based on the information given, several models architectures and their results are compared:

1. **Recurrent Neural Network Grammar (RNN Grammar)** by Dyer et al. (2016): This model achieved 91.7 on WSJ23 F1.
2. **Vinyals & Kaiser et al. (2014)**: Their model achieved 88.3 on WSJ23 F1 for WSJ only setting and a discriminative approach.
3. **Petrov et al. (2006)**: This model achieved 90.4 on WSJ23 F1 for WSJ only setting.
4. **Zhu et al. (2013)**: Their model achieved 90.4 on WSJ23 F1 for the WSJ only setting.
5. **McClosky et al. (2006)**: They achieved 92.1 on a semi-supervised setting.
6. **The Transformer (4 layers) - WSJ only:** Achieved 91.3.
7. **The Transformer (4 layers) - semi-supervised:** Achieved 92.7.
8. **Luong et al. (2015) - Multi-task:** Achieved 93.0.
9. **Dyer et al. (2016) - Generative:** Achieved 93.3.

To directly answer your question: There are no specific "llm models" mentioned to compare; however, multiple models architectures and results comparisons are discussed throughout the document.

Manage app

92°F

Search

3:54 PM