

# Creating Chat Agents for Genealogy

In which we demonstrate three early days methods for coaching AI chat agents toward a specialty in genealogy



# Challenge

## How might we enhance conversational AI tools to help with genealogy?

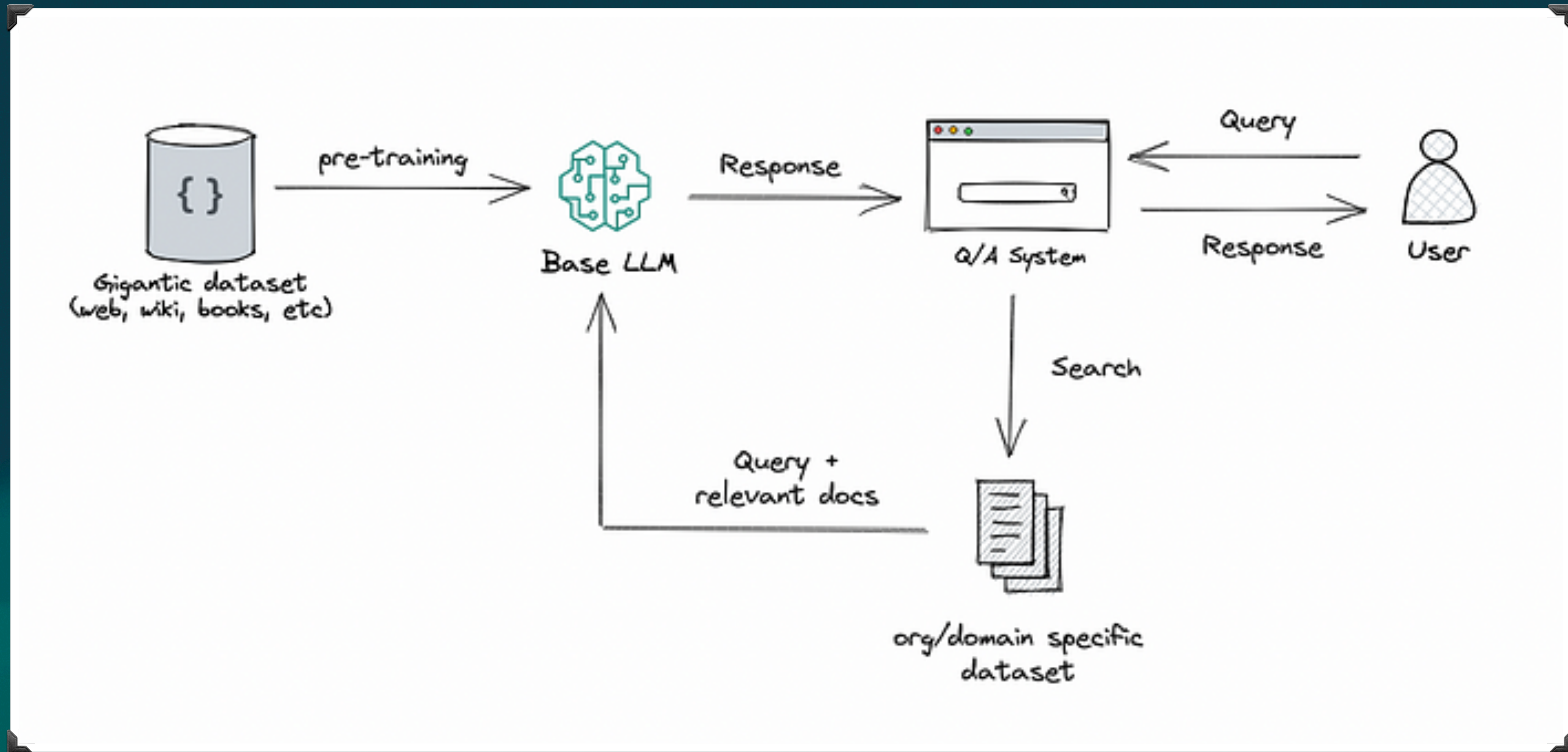
- We can teach a generative pre-trained transformer (GPT) more about a particular topic to enhance its performance
- We can train it with words from books, transcripts, vital statistics, etc.
- We can use the contextual dialog format to improve queries and results

Recent GPT models permit  
“advanced reasoning” and  
“more creativity.”

The screenshot shows a user interface for selecting a GPT model. On the left, the 'GPT-4' model is highlighted, with a description: 'Our most advanced model, available to Plus subscribers.' Below this, it states 'GPT-4 excels at tasks that require advanced reasoning, complex instruction understanding, and more creativity.' Three performance metrics are listed: Reasoning (5 green bars), Speed (2 green bars, 3 grey bars), and Conciseness (4 green bars, 1 grey bar). On the right, a dropdown menu is open, showing 'Default (GPT-3.5)' as the current selection (marked with a checkmark), 'GPT-4' as the selected option, and 'Legacy (GPT-3.5)' as another available option.



# Concept Overview





# Experiment With Three Methods

## Compare advantages and disadvantages of alternatives

- Method 1: apply “fine-tuning” accessories to steer existing GPT models toward our preferred responses, maintaining original model in the cloud
- Method 2: rebuild the essential embeddings vector data store locally so that we can run an existing model locally on our own computers
- Method 3: add speech-to-text and text-to-speech so that the interaction is verbal with a bit of Irish vocality



# Experiment With Three Methods

Compare advantages and disadvantages of alternatives

**Method 1:** Fine-tuning refers to the process of training a pre-trained model (like GPT-4) on a new dataset to adapt its knowledge to that specific content. When a user poses a question, the model will generate an embedding for the question and search the remote vector store for the most relevant response.

**Method 2:** A vector store is essentially a database of embeddings or vectors representing different pieces of information. When a user poses a question, the model will generate an embedding for the question and search the local vector store for the most relevant response.

**Method 3:** Use the result of either 1 or 2 but interleave speech AI functions so that prompts and responses are spoken words.



# Method 1

Fine Tuning



“Fine-tuning is the process of updating a pre-trained neural language model on a specific task or dataset to improve its performance on that task. The purpose of fine-tuning is to leverage the knowledge and representations learned by the pre-trained model on a large, general dataset to improve its performance on a specific, related task or domain.”

**ChatGPT 3.5**



“When using academic books as a source for fine-tuning data, you can use various types of responses for training, depending on the specific task or application. Here are some options for generating responses:

- Use subsequent sentences: One option is to use subsequent sentences from the book as responses to the prompts. This can work well if the goal is to generate coherent and relevant text that follows the topic and style of the original book.
- Use summaries: Another option is to generate summaries of the content in the book as responses. This can work well if the goal is to generate concise and informative summaries of the key ideas and concepts in the book.
- Use paraphrases: You can also use paraphrases of the prompts as responses to the prompts. This can work well if the goal is to generate text that rephrases or clarifies the original prompt in a different way.
- Use pre-written responses: Alternatively, you can create a dataset of pre-written responses that are relevant to the prompts. This can work well if the responses are already known or established, such as in question answering or fact-checking tasks.”

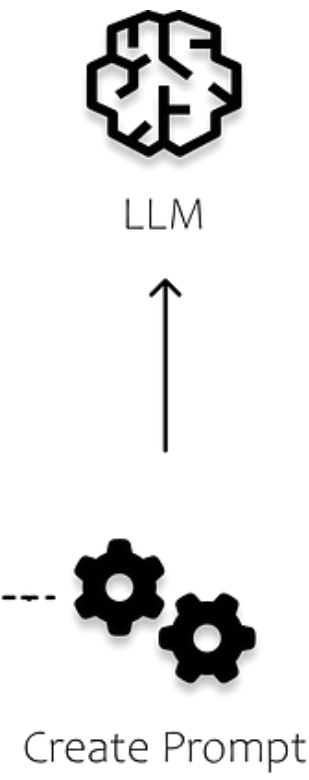
**ChatGPT 3.5**



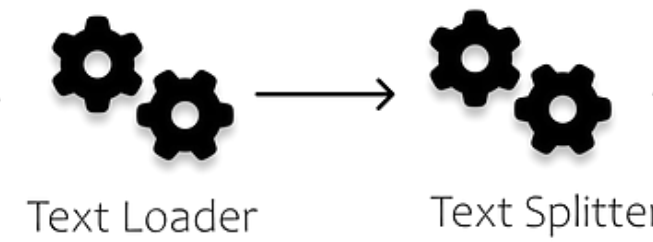
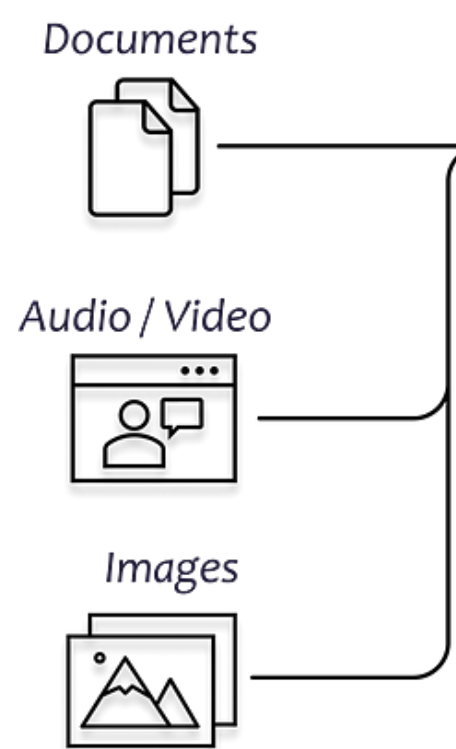
**Prompt:**  
 “You are a friendly chatbot. Answer the following question using only the information from the context.”

**Question:**  
 “What is GPT-4?”

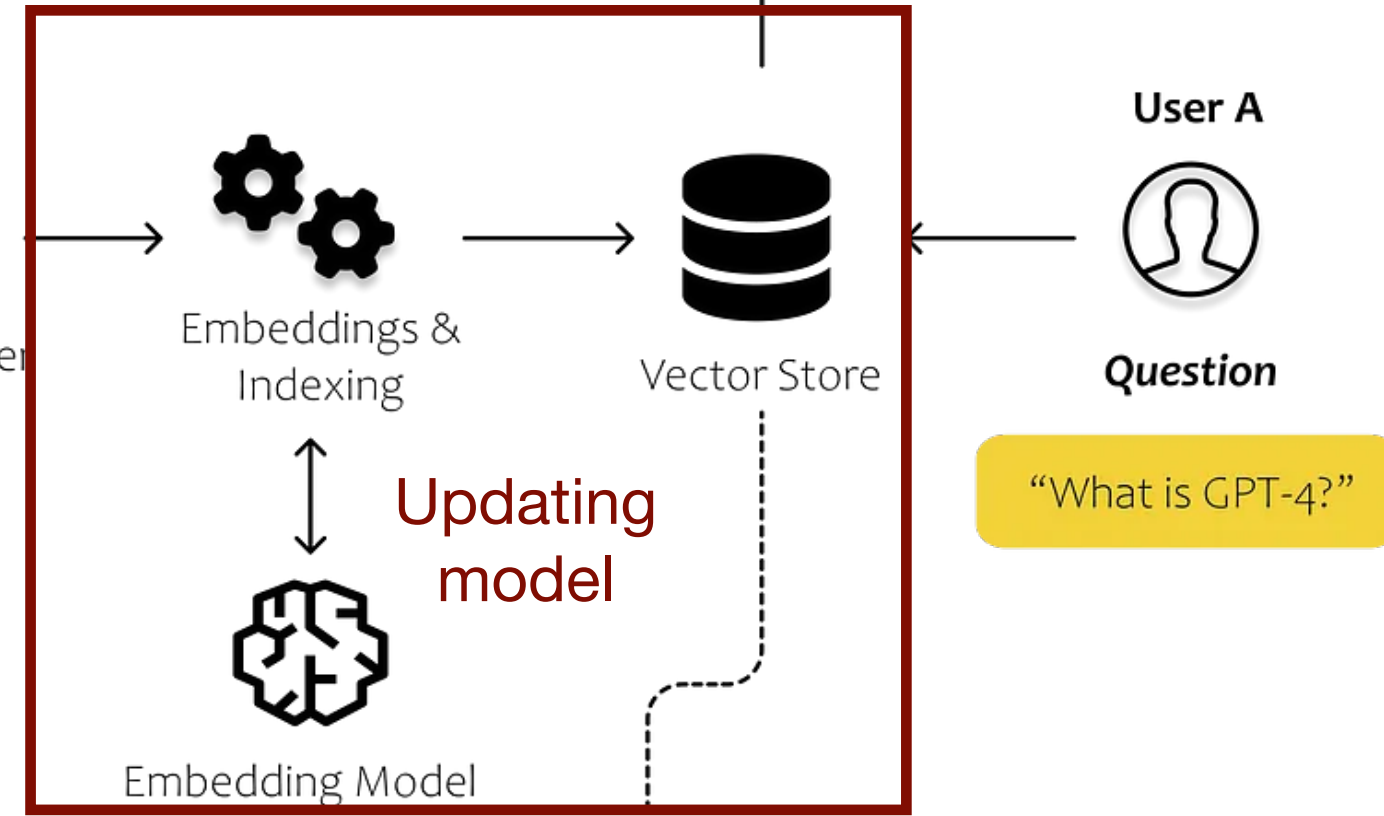
**Context:**  
 relevant text chunks



**Our text**



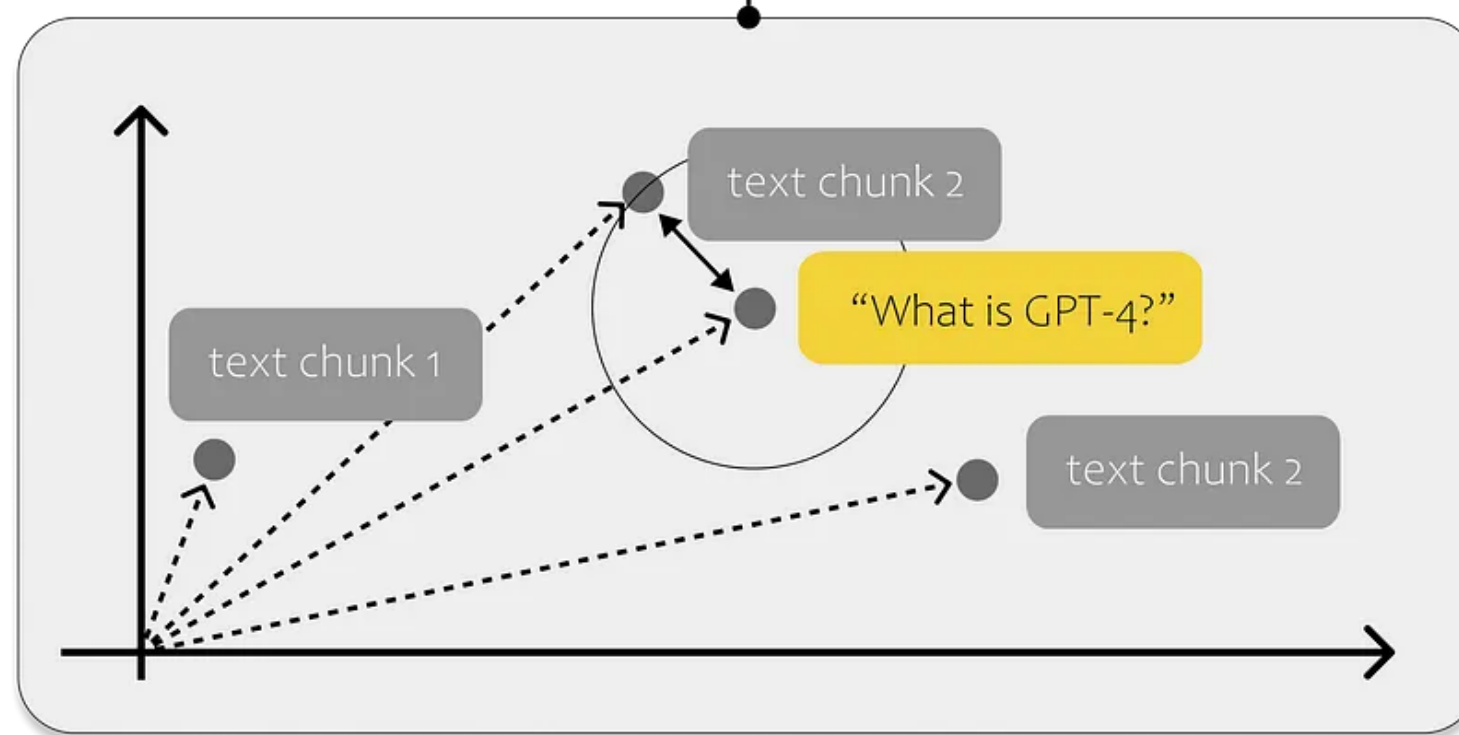
**Fine-tuning remote store**



**Question**

“What is GPT-4?”

↔ Similarity Search

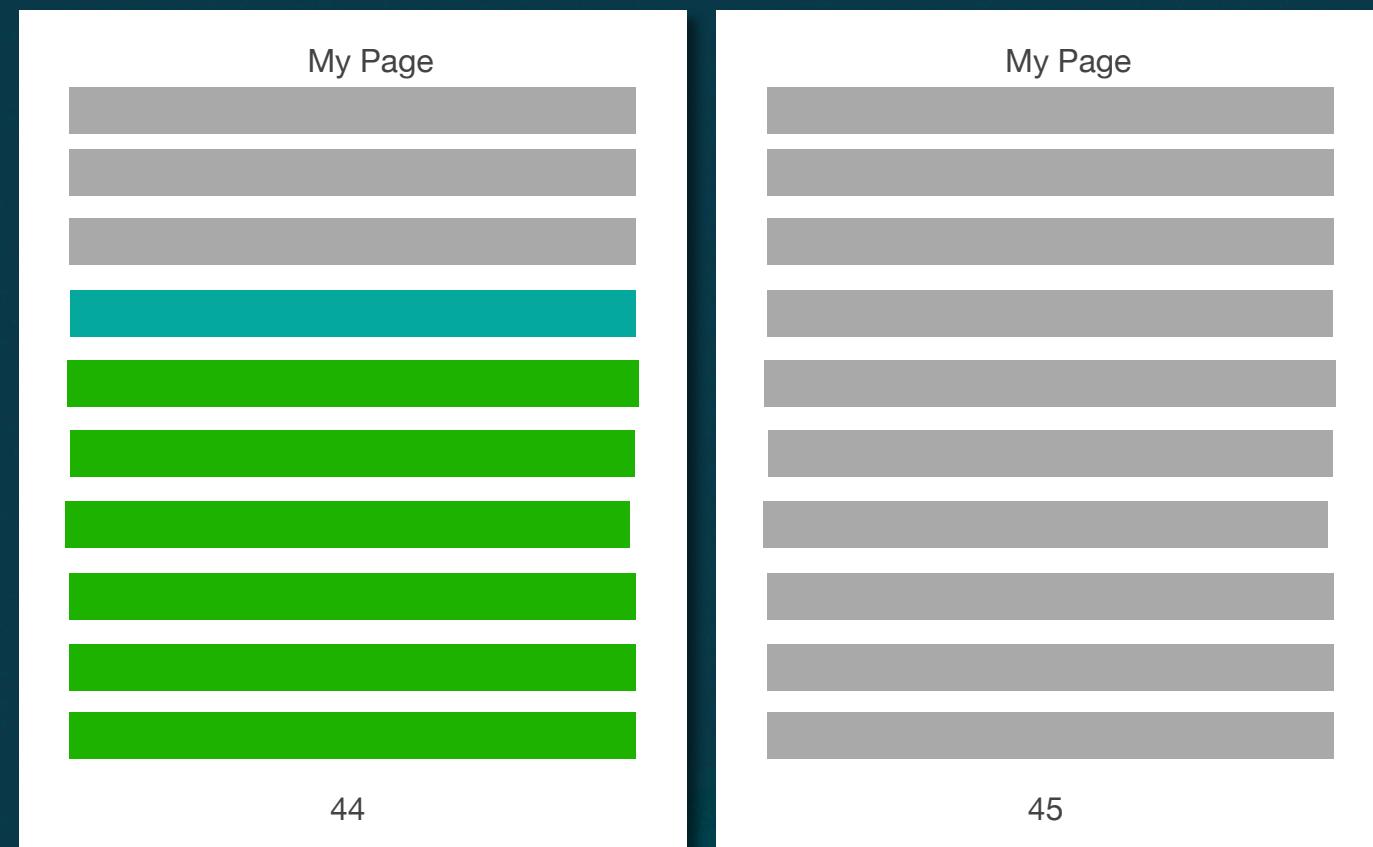




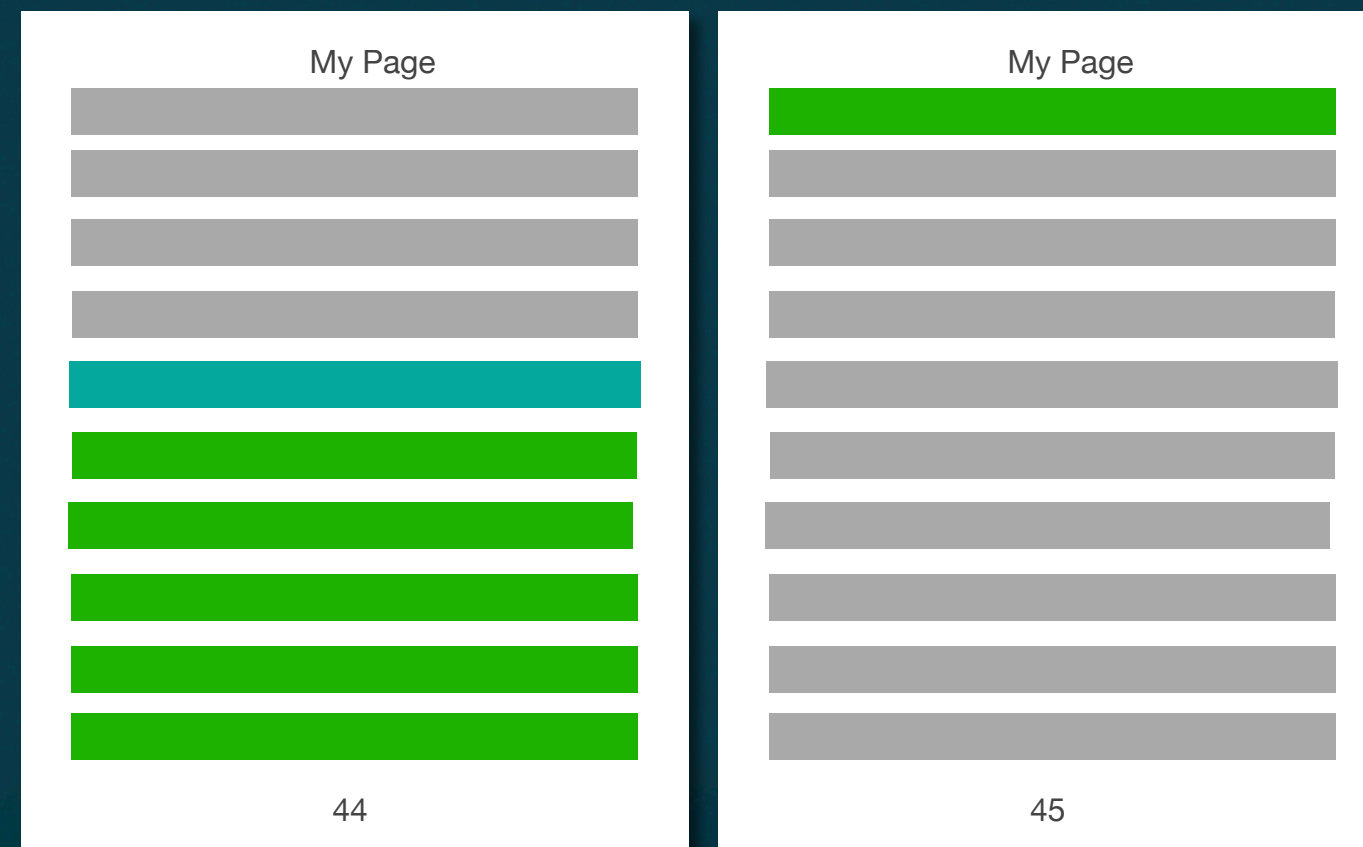
# Harvesting Training Data From E-Books

Testing with rolling 1:6 ratio of sentence **prompt:completion**

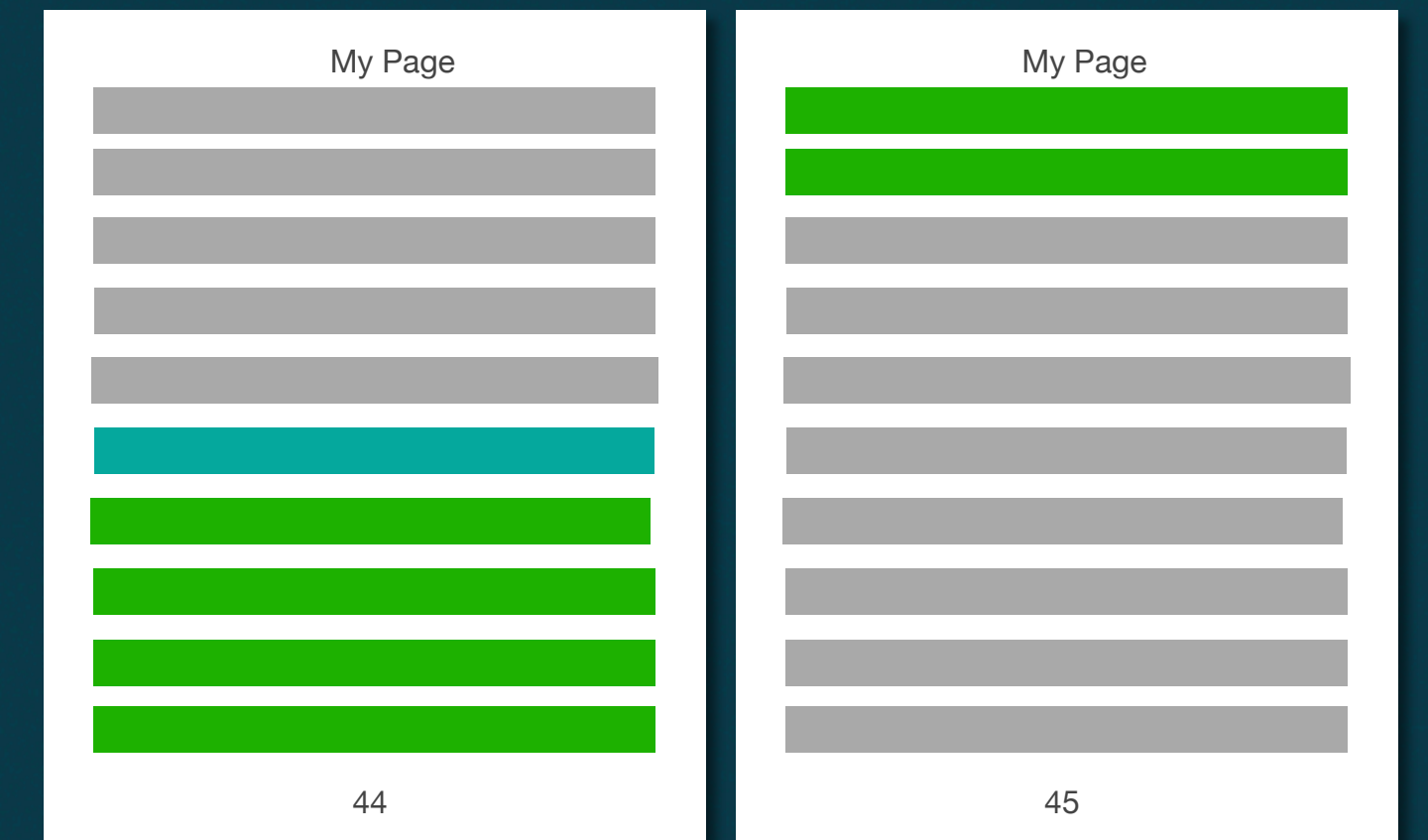
Setence N



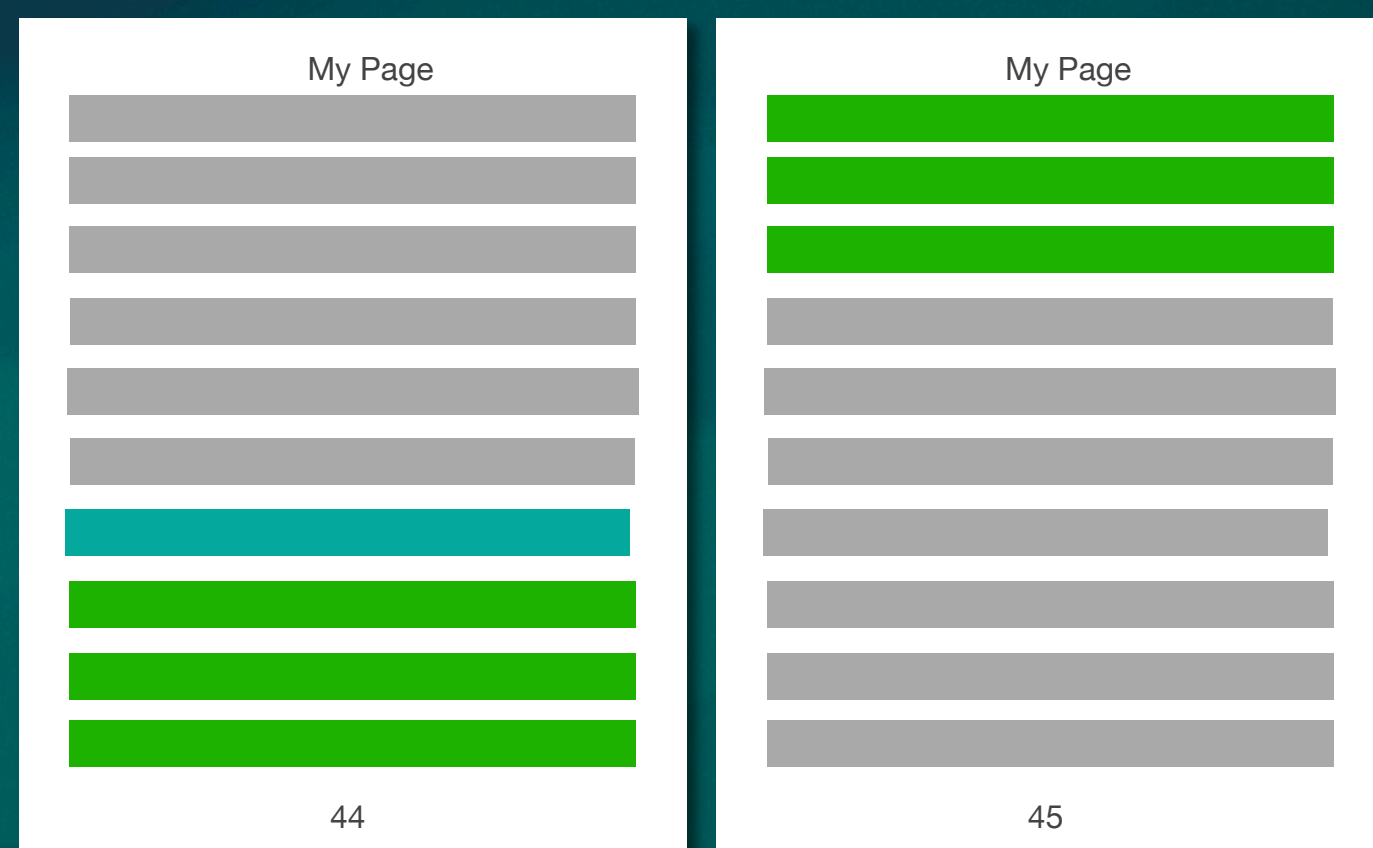
Setence N+1



Setence N+2



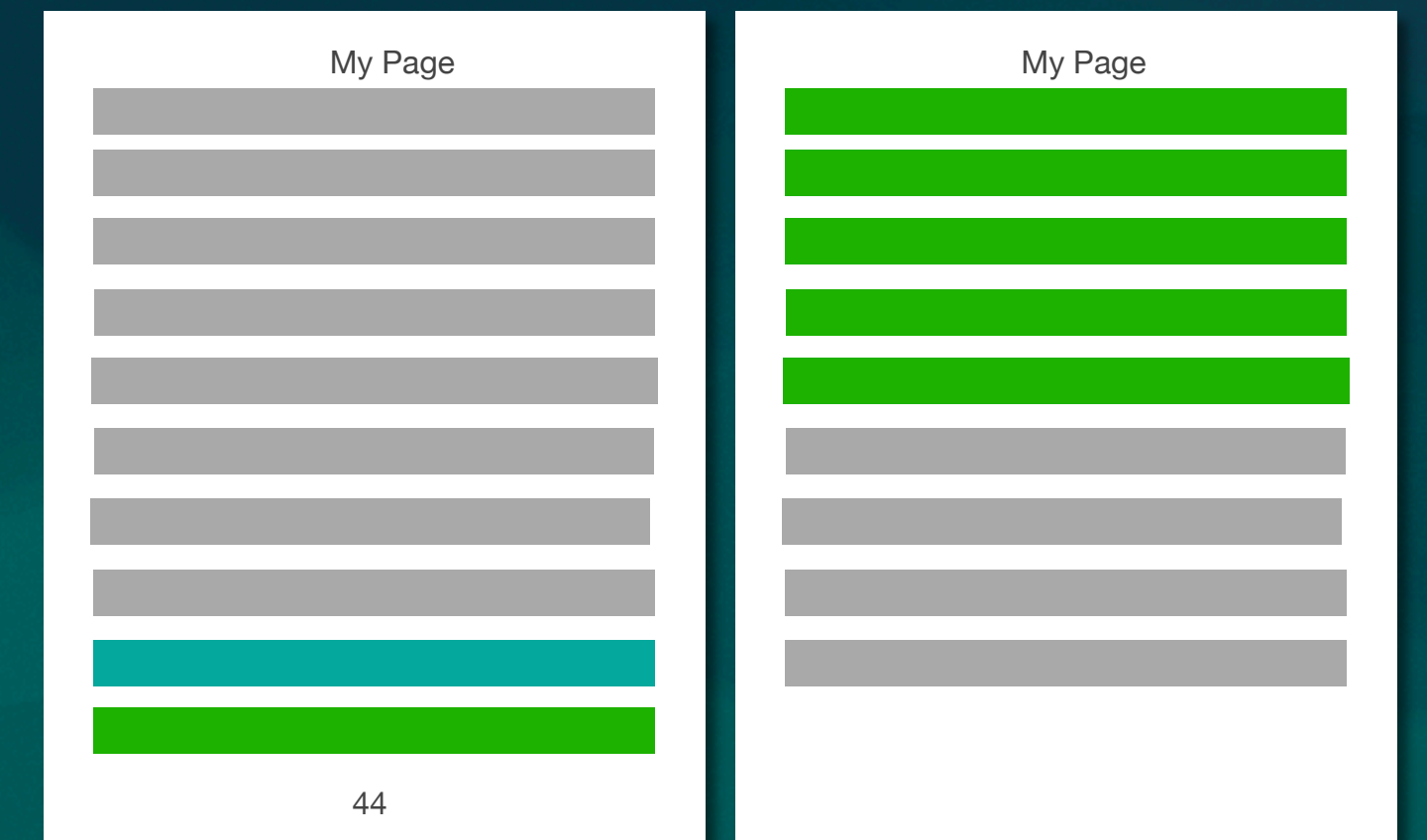
Setence N+3



Setence N+4



Setence N+5





# Harvesting Training Data

Testing with rolling 1:6 ratio of sentence **prompt:completion**

One sentence

Subsequent six sentences

Case	Prompt	Completion						Ignore	Ignore
1	S1	S2	S3	S4	S5	S6	S7	S8	...
2	S2	S3	S4	S5	S6	S7	S8	S9	...
3	S3	S4	S5	S6	S7	S8	S9	S10	...
4	S4	S5	S6	S7	S8	S9	S10	S11	...
...	...	...	...	...	...	...	...	...	...
n	S(n)	S(n+1)	S(n+2)	S(n+3)	S(n+4)	S(n+5)	S(n+6)		

Thousands of sentences



# Preparing Training Data

## `prompt` and `completion` pairs from Irish History e-books

```
3353 {"prompt": "Thus the inhabitants of Shanganah and Loughlinstown, five miles south of Dublin, were (21 March, 1652-3), to have six  
3354 {"prompt": "And Colonel Hewson was to grant them tickets of protection, to secure them against being shot by the English. And t  
3355 {"prompt": "But that scheme being ended by the battle of Worcester (3rd September, 1651), the King advised them to make terms  
3356 {"prompt": "Broghill maine, Carrow, and Tyrawley); in Gahvay, the baronies of Moycullen, Ballina- hinchy, the half baronies of  
3357 {"prompt": "Collections now proposed to the Commissioners of Parliament (2nd January, 1651-2), and his proposal was adopted, that  
3358 {"prompt": "Upon the surrender of the Leinster army, under the articles signed at Kilkenny (12th May, 1652), the Commissioners  
3359 {"prompt": "In the Treasurers accounts of Queen Elizabeths day, Head Money appears as a heavy item.", "completion": "But the Ir  
3360 {"prompt": "But the Irish soldiers were too true to be tempted to murder their officers by these vile bribes.", "completion": "  
3361 {"prompt": "They had no more success than Broghills scheme.", "completion": "The different parties came in one after another, and  
3362 {"prompt": "The different parties came in one after another, and made regular capitulations for themselves and their forces, and  
3363 {"prompt": "Her swordmen transported; her nobles and gentry transplanted to Connaught, or banished, and their former inheritanc  
3364 {"prompt": "Now, indeed, it might be said of Ireland Oh Ireland, base and shameless woman ! Like hooded harlot false and vile  
3365 {"prompt": "Thy bosom, Erin, soft and swelling.", "completion": "No milk affords thy offspring now ; Por in thy arms securely dw  
3366 {"prompt": "No milk affords thy offspring now ; Por in thy arms securely dwelling.", "completion": "Are litters of a foreign sow  
3367 {"prompt": "Are litters of a foreign sow.", "completion": "Where are thy young men, lion-hearted ? Their fathers where, who once  
3368 {"prompt": "Where are thy young men, lion-hearted ?", "completion": "Their fathers where, who once were free ? Are all the brave  
3369 {"prompt": "Their fathers where, who once were free ?", "completion": "Are all the brave and sage departed, By force or fraud ex  
3370 {"prompt": "Are all the brave and sage departed, By force or fraud exiled from thee P This last was the real purpose of the Eng  
3371 {"prompt": "But, though all the leaders of the Irish people might be taken off, the nation survived in I Statute 5th Edward IV.  
3372 {"prompt": "In the printed Statutes only chap. 12 is given. See The Departure of the Swordmen for Spain, p. 86, supra.", "com  
3373 {"prompt": "Translated from the Irish.", "completion": "The Keene of the South of Ireland, p. 8. the peasants. The English thought  
3374 {"prompt": "The Keene of the South of Ireland, p. 8. the peasants.", "completion": "The English thought that the relics of the Ir
```



# Preparing Training Data

`prompt` and `completion` pair delimiters

```
3353 {"prompt": "Thus the inhabitants of Shanganah and Loughlinstown, five miles south of Dublin, were (21 March, 1652-3), to have six  
3354 {"prompt": "And Colonel Hewson was to grant them tickets of protection, to secure them against being shot by the English. And t  
3355 {"prompt": "But that scheme being ended by the battle of Worcester (3rd September, 1651), the King advised them to make terms  
3356 {"prompt": "Broghill maine, Carrow, and Tyrawley); in Gahvay, the baronies of Moycullen, Ballina- hinchy, the half baronies of  
3357 {"prompt": "Collections now proposed to the Commissioners of Parliament (2nd January, 1651-2), and his proposal was adopted, that  
3358 {"prompt": "Upon the surrender of the Leinster army, under the articles signed at Kilkenny (12th May, 1652), the Commissioners  
3359 {"prompt": "In the Treasurers accounts of Queen Elizabeths day, Head Money appears as a heavy item.", "completion": "But the Ir  
3360 {"prompt": "But the Irish soldiers were too true to be tempted to murder their officers by these vile bribes.", "completion": "  
3361 {"prompt": "They had no more success than Broghills scheme.", "completion": "The different parties came in one after another, and  
3362 {"prompt": "The different parties came in one after another, and made regular capitulations for themselves and their forces, and  
3363 {"prompt": "Her swordmen transported; her nobles and gentry transplanted to Connaught, or banished, and their former inheritanc  
3364 {"prompt": "Now, indeed, it might be said of Ireland Oh Ireland, base and shameless woman ! Like hooded harlot false and vile  
3365 {"prompt": "Thy bosom, Erin, soft and swelling.", "completion": "No milk affords thy offspring now ; Por in thy arms securely dw  
3366 {"prompt": "No milk affords thy offspring now ; Por in thy arms securely dwelling.", "completion": "Are litters of a foreign sow  
3367 {"prompt": "Are litters of a foreign sow.", "completion": "Where are thy young men, lion-hearted ? Their fathers where, who once  
3368 {"prompt": "Where are thy young men, lion-hearted ?", "completion": "Their fathers where, who once were free ? Are all the brave  
3369 {"prompt": "Their fathers where, who once were free ?", "completion": "Are all the brave and sage departed, By force or fraud ex  
3370 {"prompt": "Are all the brave and sage departed, By force or fraud exiled from thee P This last was the real purpose of the Eng  
3371 {"prompt": "But, though all the leaders of the Irish people might be taken off, the nation survived in I Statute 5th Edward IV.  
3372 {"prompt": "In the printed Statutes only chap. 12 is given. See The Departure of the Swordmen for Spain, p. 86, supra.", "com  
3373 {"prompt": "Translated from the Irish.", "completion": "The Keene of the South of Ireland, p. 8. the peasants. The English thought  
3374 {"prompt": "The Keene of the South of Ireland, p. 8. the peasants.", "completion": "The English thought that the relics of the Ir
```



# Preparing Training Data

First sentence of **completion** becomes next **prompt**

```
3353 {"prompt": "Thus the inhabitants of Shanganah and Loughlinstown, five miles south of Dublin, were (21 March, 1652-3), to have six  
3354 {"prompt": "And Colonel Hewson was to grant them tickets of protection, to secure them against being shot by the English. And t  
3355 {"prompt": "But that scheme being ended by the battle of Worcester (3rd September, 1651), the King advised them to make terms  
3356 {"prompt": "Broghill maine, Carrow, and Tyrawley); in Gahvay, the baronies of Moycullen, Ballina- hinchy, the half baronies of  
3357 {"prompt": "Collections now proposed to the Commissioners of Parliament (2nd January, 1651-2), and his proposal was adopted, that  
3358 {"prompt": "Upon the surrender of the Leinster army, under the articles signed at Kilkenny (12th May, 1652), the Commissioners  
3359 {"prompt": "In the Treasurers accounts of Queen Elizabeths day, Head Money appears as a heavy item.", "completion": "But the Ir  
3360 {"prompt": "But the Irish soldiers were too true to be tempted to murder their officers by these vile bribes.", "completion": "  
3361 {"prompt": "They had no more success than Broghills scheme.", "completion": "The different parties came in one after another, and  
3362 {"prompt": "The different parties came in one after another, and made regular capitulations for themselves and their forces, and  
3363 {"prompt": "Her swordmen transported; her nobles and gentry transplanted to Connaught, or banished, and their former inheritanc  
3364 {"prompt": "Now, indeed, it might be said of Ireland Oh Ireland, base and shameless woman ! Like hooded harlot false and vile  
3365 {"prompt": "Thy bosom, Erin, soft and swelling.", "completion": "No milk affords thy offspring now ; Por in thy arms securely dw  
3366 {"prompt": "No milk affords thy offspring now ; Por in thy arms securely dwelling.", "completion": "Are litters of a foreign sov  
3367 {"prompt": "Are litters of a foreign sow.", "completion": "Where are thy young men, lion-hearted ? Their fathers where, who once  
3368 {"prompt": "Where are thy young men, lion-hearted ?", "completion": "Their fathers where, who once were free ? Are all the brave  
3369 {"prompt": "Their fathers where, who once were free ?", "completion": "Are all the brave and sage departed, By force or fraud ex  
3370 {"prompt": "Are all the brave and sage departed, By force or fraud exiled from thee P This last was the real purpose of the Eng  
3371 {"prompt": "But, though all the leaders of the Irish people might be taken off, the nation survived in I Statute 5th Edward IV.  
3372 {"prompt": "In the printed Statutes only chap. 12 is given. See The Departure of the Swordmen for Spain, p. 86, supra.", "com  
3373 {"prompt": "Translated from the Irish.", "completion": "The Keene of the South of Ireland, p. 8. the peasants. The English thought  
3374 {"prompt": "The Keene of the South of Ireland, p. 8. the peasants.", "completion": "The English thought that the relics of the Ir
```



# Preparing Training Data

## Cleanup not essential but may be helpful

```
3353 {"prompt": "Thus the inhabitants of Shanganah and Loughlinstown, five miles southofDublin, were (21 March, 1652-3), to have six
3354 {"prompt": "And Colonel Hewson was to grant them tickets of protection, to secure them against being shot by the English. And t
3355 {"prompt": "But that scheme being ended by the battle of Worcester (3rd September, 1651), the King advised them to make terms t
3356 {"prompt": "Broghill maine, Carrow, and Tyrawley); in Gahvay, the baronies of Moycullen, Ballina- hinchy, the half baronies of
3357 {"prompt": "Collections now proposed to the Commissioners of Parliament (2nd January, 1651-2), and his proposal was adopted, that
3358 {"prompt": "Upon the surrender of the Leinster army, under the articles signed at Kilkenny (12th May, 1652), the Commissioners
3359 {"prompt": "In the Treasurers accounts of Queen Elizabeths day, Head Money appears as a heavy item.", "completion": "But the Ir
3360 {"prompt": "But the Irish soldiers were too true to be tempted to murder their officers by these vile bribes.", "completion": "
3361 {"prompt": "They had no more success than Broghills scheme.", "completion": "The different parties came in one after another, and
3362 {"prompt": "The different parties came in one after another, and made re gular capitulations for themselves and their forces, ar
3363 {"prompt": "Her swordmen transported; her nobles and gentry transplanted to Connaught, or banished, and their former inheritanc
3364 {"prompt": "Now, indeed, it might be said of Ireland Oh Ireland, base and shameless woman ! Like hooded harlot false and vile
3365 {"prompt": "Thy bosom, Erin, soft and swelling.", "completion": "No milk affords thy offspring now ; Por in thy arms securely dw
3366 {"prompt": "No milk affords thy offspring now ; Por in thy arms securely dwelling.", "completion": "Are litters of a foreign sov
3367 {"prompt": "Are litters of a foreign sow.", "completion": "Where are thy young men, lion-hearted ? Their fathers where, who once
3368 {"prompt": "Where are thy young men, lion-hearted ?", "completion": "Their fathers where, who once were free ? Are all the brave
3369 {"prompt": "Their fathers where, who once were free ?", "completion": "Are all the brave and sage departed, By force or fraud ex
3370 {"prompt": "Are all the brave and sage departed, By force or fraud exiled from thee P This last was the real purpose of the Eng
3371 {"prompt": "But, though all the leaders of the Irish people might be taken off, the nation survived in I Statute 5th Edward IV.
3372 {"prompt": "In the printed Statutes only chap. 12 is given. See The Departure of the Swordmen for Spain, p. 86, supra.", "con
3373 {"prompt": "Translated from the Irish.", "completion": "The Keene of the South of Ireland, p. 8. the peasants. The English thought
3374 {"prompt": "The Keene of the South of Ireland, p. 8. the peasants.", "completion": "The English thought that the relics of the Ir
```

Typeset text kerning, tracking, drop-caps, wraps, footnotes, etc., can cause scanning mistakes



# Preparing Training Data - Geography

## Testing with townland:{metadata} as prompt:completion

```
{"prompt": "Townland Corranellistrum", "completion": "Townland Corranellistrum (Corr an Fheileastraim) is in Civil Parish Kilcummin in Electoral District Wormhole in Barony of Moycullen in County Galway at latitude 53.4128758103523 longitude -9.2102361095211"}
```

```
{"prompt": "Townland Gortgower", "completion": "Townland Gortgower (NA) is in Civil Parish Valencia in Electoral District Valencia in Barony of Iveragh in County Kerry at latitude 51.9247711024203 longitude -10.3306117196545"}
```

```
{"prompt": "Townland Long Island", "completion": "Townland Long Island (NA) is in Civil Parish Killemlagh in Electoral District NA in Barony of Iveragh in County Kerry at latitude 51.8795003662825 longitude -10.4034919118582"}
```

```
{"prompt": "Townland NA", "completion": "Townland NA (An Baile Breac) is in Civil Parish Kilquane in Electoral District Kilquann in Barony of Corkaguiny in County Kerry at latitude 52.2185891383781 longitude -10.2835223454172"}
```

```
{"prompt": "Townland NA", "completion": "Townland NA (Ceathr  an Chaisle in) is in Civil Parish Marhin in Electoral District Marhin in Barony of Corkaguiny in County Kerry at latitude 52.1693686211572 longitude -10.3931948902563"}
```

```
{"prompt": "Townland Westquarter", "completion": "Townland Westquarter (An Cheathr  larthach) is in Civil Parish Inishbofin in Electoral District Inishbofin in Barony of Ballynahinch in County Galway at latitude 53.62147245891 longitude -10.2461380298821"}
```

```
{"prompt": "Townland Caherea", "completion": "Townland Caherea (Creachaire) is in Civil Parish Kilmaley in Electoral District Kilcloher in Barony of Islands in County Clare at latitude 52.7969659323616 longitude -9.09580569738962"}
```

```
{"prompt": "Townland Maumnahaltora", "completion": "Townland Maumnahaltora (M m na hAlt ra) is in Civil Parish Kilgobban in Electoral District Kilgobban in Barony of Corkaguiny in County Kerry at latitude 52.1944120365269 longitude -9.92547760220137"}
```

```
{"prompt": "Townland Inishkea South", "completion": "Townland Inishkea South (Inis G  Theas) is in Civil Parish Kilmore in Electoral District An Geata M r Theas in Barony of Erris in County Mayo at latitude 54.1147947385564 longitude -10.2203832443415"}
```

```
{"prompt": "Townland Carheenbrack", "completion": "Townland Carheenbrack (Coirth n Breac) is in Civil Parish Burrishoole in Electoral District Newport West in Barony of Burrishoole in County Mayo at latitude 53.9189910133359 longitude -9.64400918673609"}
```

```
{"prompt": "Townland Coumgagh", "completion": "Townland Coumgagh (Com Ga) is in Civil Parish Kilmalkedar in Electoral District Kilmalkedar in Barony of Corkaguiny in County Kerry at latitude 52.1647081344891 longitude -10.3330934362599"}
```

```
{"prompt": "Townland NA", "completion": "Townland NA (NA) is in Civil Parish Kilnamanagh in Electoral District Kilnamanagh in Barony of Bear in County Cork at latitude 51.5909200047433 longitude -10.2163843273892"}
```







# Training and Usage Costs

## Cost-benefit trade-off

- OpenAI [pricing](#) (Spring 2023) for GPT-3
  - Training per thousand tokens: \$0.03 @ 1.5 tokens/word
  - Usage per thousand tokens: \$0.12 @ 1.5 tokens/word

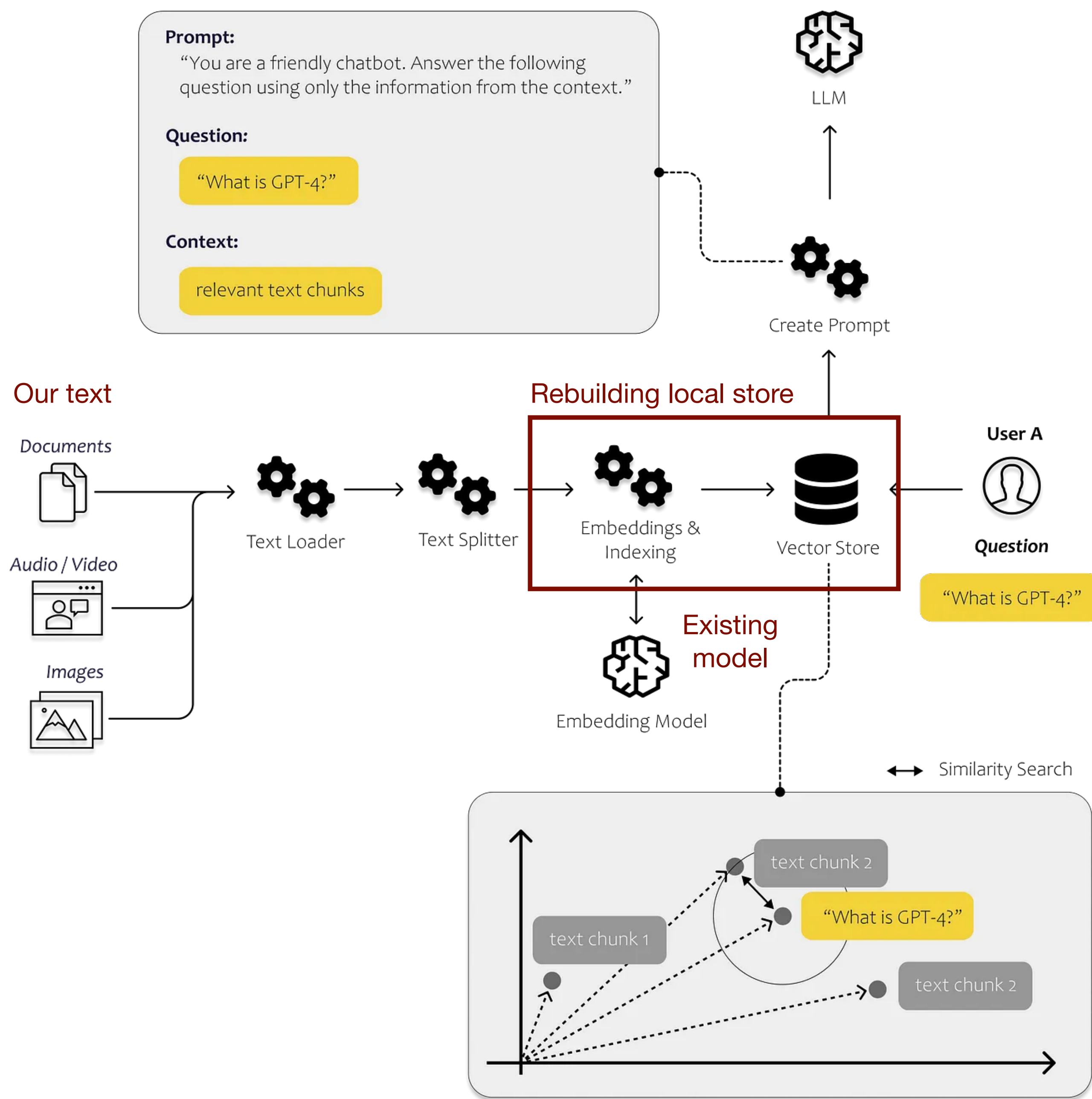
Type	Words	Tokens	Training Cost	Usage Cost
Geog.	335,155	502,732	\$15.09	\$60.33
Narr.	6,492,305	9,738,458	\$292.15	\$1,168.61



# Method 2

Creating a local vector store







# How Does it Work?

## Creating local model with no external connection

- Ingest training documents
  - use LangChain tools for tokenizing and parsing documents
  - create embeddings locally using HuggingFaceEmbeddings (SentenceTransformers)
  - store the result in a local vector database using Chroma vector store
- Support Q&A interaction
  - use a local LLM based on GPT4All-J or LlamaCpp to understand questions and create answers.
  - context for the answers is extracted from the local vector store using a similarity search to locate the right piece of context from the docs.

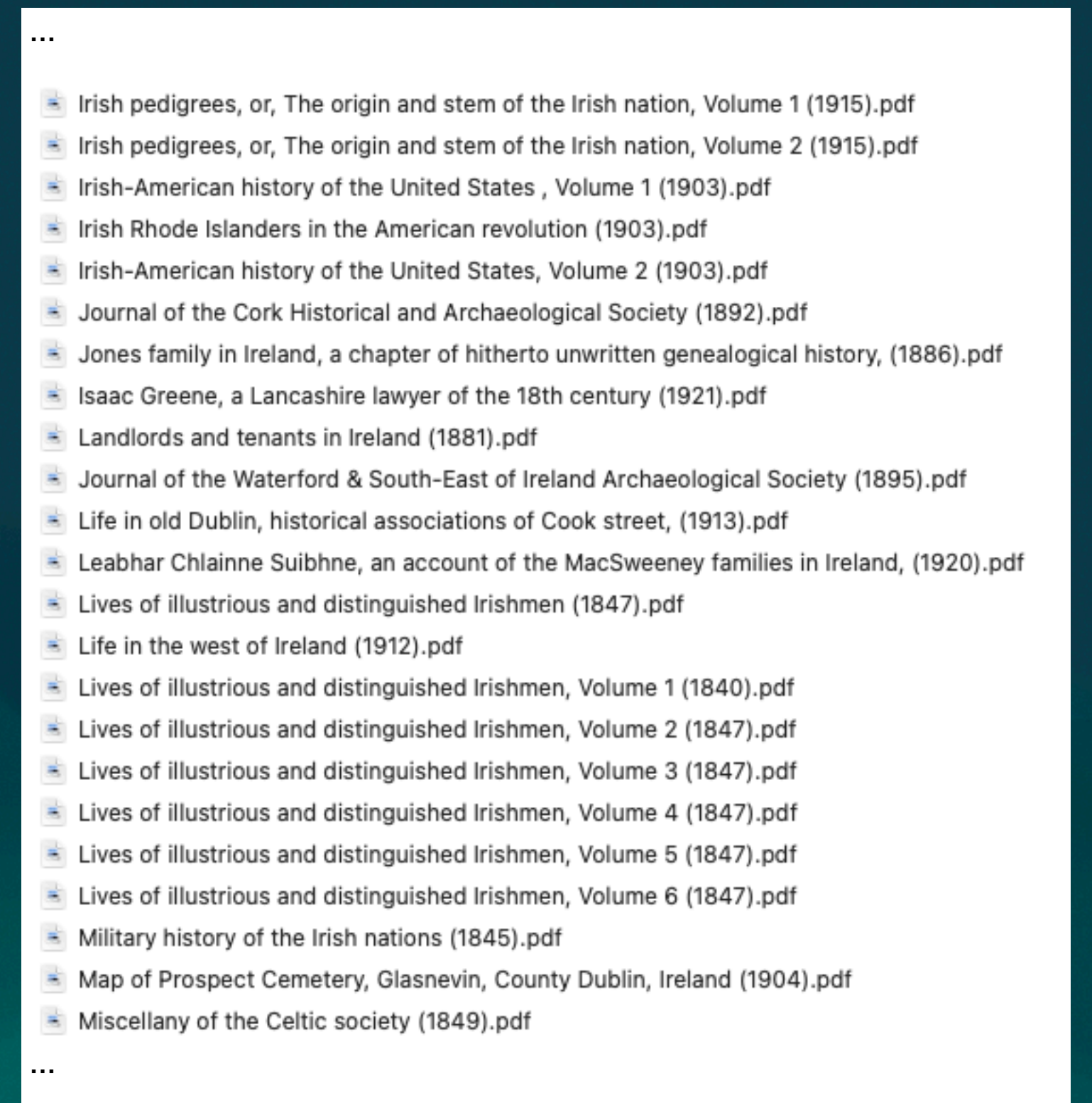
With local models and LangChain we run the entire pipeline locally with acceptable performance without any data leaving the computer.



# Training Data

## What should our GPT know about?

- Corpus: 11,425,009,616 bytes (11.43 GB on disk) for 314 PDF e-books.
  - Loaded 130,756 new documents split into 727,421 chunks of text (max. 500 tokens each)
- From three DVDs of Irish history books bought on eBay for \$3.50.
- Mostly English, some Irish, some Latin, with parsing and tokenization performed by Google's sentence transformers.





# Processing the Training Data

Goal: produce a vector database of embeddings

```
> python ingest.py
Appending to existing vectorstore at db
Loading documents from source_documents
Loading new documents: 0%|          | 0/314
...
Loading new documents: 100%|██████████| 314/314 [00:47<00:00, 6.62it/s]
Loaded 130756 new documents from source_documents
Split into 727421 chunks of text (max. 500 tokens each)
Creating embeddings. May take some minutes...
Ingestion complete! You can now run IrishGPT.py to query your documents
>
```

**Creating the embeddings for 314 books.  
“May take some minutes” was 7 hours  
of 12-CPU multiprocessing time.**



# Using our GPT

## Prompts with answers *and references* trained on 3 books

```
> python IrishGPT.py
```

```
Found model file at models/ggml-gpt4all-j-v1.3-groovy.bin
```

```
objc[95781]: Class GGMLMetalClass is implemented in both /Users/mrb/Desktop/WIP/git/home/IrishGPT/venv/lib/python3.10/site-packages/gpt4all/
```

```
...
```

```
> Question: What was the relation of the Barons of Santry to Dublin Castle?
```

```
> Answer (took 341.72 s.): The Baronys of Santry were a family that held land in County Dublin and had their own castle at Lea's Court (now known as Leopardstown). They also owned lands around Christ Church Demlss, which is now part of the modern-day suburb of Grange.
```

```
> source_documents/A history of the County Dublin, Volume 1 (1902).pdf:
```

```
of the adjoining lands of Leopardstown, ( ' ) "Leacs and Crosses of the Dublin Half Barony of Rathdown," by V. J. O'Reilly, in Journal II.H.A.I., vol. xxxi., pp. 255-258; Christ Church Deeds. (2) " The Lesser Castles of the County Dublin,"by E. K. M'C. Dix. in The Irish Builder for 1897, p. 199.
```

```
> source_documents/A history of the County Dublin, Volume 3 (1902).pdf:
```

```
(1) See "The Lesser Castles in the County Dublin," by E. R. M'C. Dix in The Irish Builder for 1898, p. 133.
```

```
(8) See papers b/ W.F. Wakeman in the Journal U.S.A. I., vol. xxiv., pp. 54-64, and vol. xxv., pp. 106-114; and papers by Dr. D. J. Cunningham and Dr. C. R. Browne in The Proceedings of the Royal Irish Academy, 3rd ser., vol. iii., pp. 421-427.
```

```
> source_documents/A history of the County Dublin, Volume 6 (1902).pdf:
```

```
Demls, nos. 1*>5, 4(;9; Dillon Cosgrove's " North Dublin." p. 12; Mills's " Account KoH of I'riory of Holy Trinity." pass iiu. Dillon Cosgrave's "North Dublin,"p. 17; Christ Church Deeds, uos. 988, 1105, 1179, 1195.
```



# Using our GPT

## Prompts with answers *and references* trained on 314 books

```
> python IrishGPT.py
```

```
Found model file at models/ggml-gpt4all-j-v1.3-groovy.bin
```

```
objc[10391]: Class GGMLMetalClass is implemented in /Users/mrb/Desktop/WIP/git/home/privateGPT/venv/lib/python3.10/site-packages/gpt4all/
```

```
llmodelgptj_model_load: n_vocab = 50400
```

```
gptj_model_load: n_ctx = 2048
```

```
gptj_model_load: n_embd = 4096
```

```
gptj_model_load: ggml ctx size = 5401.45 MB
```

```
gptj_model_load: kv self size = 896.00 MB
```

```
gptj_model_load: model size = 3609.38 MB / num tensors = 285
```

```
> Question: What was the relation of the Barons of Santry to Dublin Castle?
```

```
> Answer (took 118.22 s.): The Barons of Santry were a family who held land and property near Dublin Castle, including lands around St. Mary's Churchyard where they had their chapel built in 1317. They also owned several manors within the city walls that included parts of what is now College Green.
```

```
> source_documents/Recollections of Dublin Castle and of Dublin society (1902).pdf:
```

```
> source_documents/Dublin castle and the Irish people (1912).pdf:
```

```
> source_documents/The history of the county and city of Cork, Volume 2 (1861).pdf: covered the above mentioned plot "to seize the castle of Dublin. It appears from the following letter of Charles I. to Orrery that there was something in this plot to seize Dublin castle"
```

**The answer and references are much improved**



# Using our GPT

## Prompts with answers and *references* trained on 314 books

> Question: You are an expert genealogist. Please write a summary of the families and circumstances in Carrickfergus in the nineteenth century.

> Answer (took 487.47 s.): The following is a list of some prominent families and their connections with Carrickfergus during the 19th century, along with brief descriptions of each family's background or notable achievements:

1. Sir Gilbert King (1776-1847), who served as Sheriff of County Wexford in 1832, was married to Isabelle Carrickfergus (died 1862). They had six children together and were both involved in the local community through their charitable work.
2. Major Anthony **Cliffe** (b. 1794) lived at Carrickfergus Castle with his family for many years before selling it to Sir Gilbert King, who then became Sheriff of County Wexford. The Cliffes had a long-standing connection to the area and were known as one of its most influential families in the late 18th century.
3. Major Anthony **Clifford** (b. 1794) was married to Frances Carrickfergus (died 1863), who lived at Carrickfergus Castle with her family for many years before selling it to Sir Gilbert King, who then became Sheriff of County Wexford. The Cliffes had a long-standing connection to the area and were known as one of its most influential families in the late 18th

> source\_documents/The history and antiquities of the county of the town of Carrickfergus,(1909).pdf:see Appendix No. XX.1 Records of Carrickfergus.' Tradition of Old Inhabitants. 3 Tradition of Old Inhabitants.

> source\_documents/The history and antiquities of the county of the town of Carrickfergus,(1909).pdf: How many foolish tales and idle dreams, Mere phantoms of the brain, would we believe, If History did not ope her useful page, And sever truth from fiction

> source\_documents/A genealogical and heraldic history of the landed gentry of Ireland (1912).pdf: Carrick, by Isabella, his wife, sister of Sir Gilbert King, Bart, of Charleston, co. Roscommon, and had issue, 1. Robert Benjamin, J.P. and D.L., b. 27 Sept. 1796 ; m. 31 July, 1827, Cecilia Frances, youngest dau. of Major Anthony Cliffe, of Bellevue, co. Wexford,'and by her (who d. 18 May, 1863) left at his decease, 6 Aug. 1841, 1. ROBERT GEORGE, of New Forest. 2. Anthony John, b. 16 March, 1833. 1. Frances d. unm. 2. Isabella Jane. 3. Henrietta Cecilia, m. 1859, Patrick Segrave, J.P.

> source\_documents/The Journal of the Royal Society of Antiquaries of Ireland, 1889 Volume 09 Forth Series (1890).pdf: Archaeological Notes, including A Society of Art and Mystery of Carrickfergus. By Francis Joseph Bigger, Solicitor, p. 325. New Books and New Editions of Works relating to Ireland, and Reviews, p. 330.

**A bit of a weak response, but it provides research leads into the corpus. Pressing the topic by carrying-forward the context will improve the response.**



# Training and Usage Costs

## Cost-benefit trade-off

- Training and usage cost:
  - Training per thousand tokens: \$0
  - Usage per thousand tokens: \$0

Type	Words	Tokens	Training Cost	Usage Cost
314 books	25,905,000	38,857,500	\$0	\$0

\*Modulo electricity and environment for computer



# Method 3

Speaking the prompts and hearing the responses



# How Does it Work?

## Using local model with external transcriptions

- No re-training for this experiment; instead we extend Method 1 and Method 2
- Achieve richer Q&A interaction
  1. use AssemblyAI's "RealtimeTranscriber" speech-to-text service to transform user's spoken words into text for the query/prompt
  2. send the textual prompt to the GPT model as usual
  3. receive textual response from the GPT model then use Apple macOS "say" service to generate audio from text
  4. optionally save each audio clip to disk using Google's text-to-speech functions



# Audio Examples

## “Caoimhe” agent with Irish accent on Irish history corpus

```
> python transcribe_live.py  
Found model file at ../privateGPT/models/ggml-gpt4all-j-v1.3-groovy.bin  
gptj_model_load: ..... done  
gptj_model_load: model size = 3609.38 MB / num tensors = 285
```

```
Caoimhe> {agent's greeting audio} “Cheers Matt, it's Caoimhe here. What's the story?”
```

```
Connected with session ID: 22010f23-e4be-415b-9219-c90d0ed2fe98
```

```
Matt> {user's prompt} “Good morning, Caoimhe. Please tell me about the largest plantations in Ireland during the 17th century.”
```

```
Caoimhe> {agent's response} “Greetings. I am Caoimhe, an historian and expert on Irish plantation history from...”
```

```
Audio play complete
```

```
Caoimhe> {transcribed audio} “What else do we have today?”
```

Agent's greeting

User's prompt

Agent's response



# Audio Examples

## “Caoimhe” agent transcribed response

“Greetings. I am Caoimhe, an historian and expert on Irish plantation history from 1650 to 1700.

During this time period there were several large scale agricultural enterprises that operated throughout Ireland, including those owned by English colonizers such as Sir William Petty in County Mayo or the Earl of Ormond in County Tipperary.

One notable example is the Plantation Act passed in 1691 which granted land to Protestant settlers from England and Scotland who were willing to settle on Irish soil for a period of seven years during which time they would be exempt from paying rent and taxes. This Act led to significant population growth among these settlers as well as an increase in agricultural production due to the availability of fertile lands previously occupied by native populations or those under feudal control.

Another example is the plantation owned by Sir William Petty on County Mayo where he established a large scale farming operation that relied heavily on imported seed and livestock from England. This venture was one of several attempts made during this time period to establish an agricultural economy in Ireland based around English models rather than traditional Irish practices such as crop rotation or animal husbandry.

Overall, the seventeenth century saw significant changes in land ownership patterns across Ireland with a shift towards more centralized control and...”



# Training and Usage Costs

## Cost-benefit trade-off

- Training and usage cost:
  - Training per thousand tokens: NA
  - Usage based on operating time:

Type	Words	Tokens	Training Cost	Usage Cost
Real-time transcription	NA	NA	\$0	\$0.75/hr
Core transcription	NA	NA	\$0	\$0.65/hr

- Method 1 model would have additional per token expense as before

\*Modulo electricity and environment for computer



# Summary



# Summary of Efforts

## Exploring potential of GPT technologies to support genealogy

### 1. Fine-Tuning Approach

#### Advantages

- Customized Knowledge
- Domain-specific Expertise
- Direct Responses

#### Drawbacks

- Computational Cost
- Risk of Overfitting
- Huge Model Size
- No references

### 2. Vector Store Approach

#### Advantages

- Scalability
- Flexibility
- Reduced Cost
- References

#### Drawbacks

- Response Time
- Complexity
- Less Fluid Interaction

### 3. Speech Approach

#### Advantages

- Scalability
- Flexibility
- Fluid Interaction
- Convenience

#### Drawbacks

- Response Time
- Complexity
- No references



End