RAG for school literature - PoC

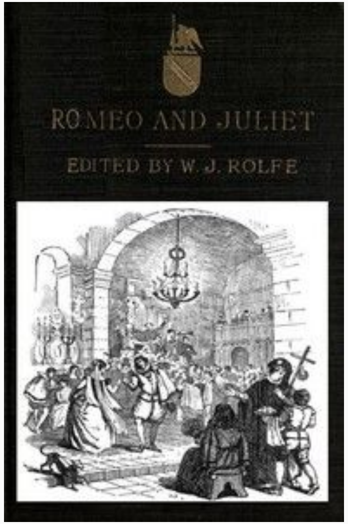
Maja Andrzejczuk, Piotr Bielecki, Paweł Gelar, Jakub Kasprzak, Maciej Orsłowski

Introduction to our project

Developing a RAG system for question answering.

Combining retrieval methods and generative models.

Focus on **famous classic books** (e.g., Romeo and Juliet, Hamlet, Around the World in 80 Days).



[Shakespeare's Tragedy of Romeo and Juliet by William Shakespeare | Project Gutenberg]

RAG (Retrieval-Augmented Generation)

Hybrid model combining document retrieval and text generation.

- Document retrieval: Retrieve relevant documents or passages from a large corpus.
- Text generation: Use a language model to generate answers based on the retrieved information.

Methodology and Framework Selection

1. Data Loading and Preprocessing:

   UnstructuredFileLoader: Loads and preprocesses text documents (PDF, TXT).

   RecursiveCharacterTextSplitter: Splits text into 2000-character chunks with overlap.

2. Vector Embeddings and Search:

   Embedding Model: all-MiniLM-L6-v2 converts passages into vector embeddings.

   FAISS Vector Database: Fast search tool for indexing and retrieval of document passages.

Methodology and Framework Selection

3. Cross-Encoder Reranker Model:

   bge-reranker-large: Re-ranks passages to enhance relevance.

4. LLM (Large Language Model):

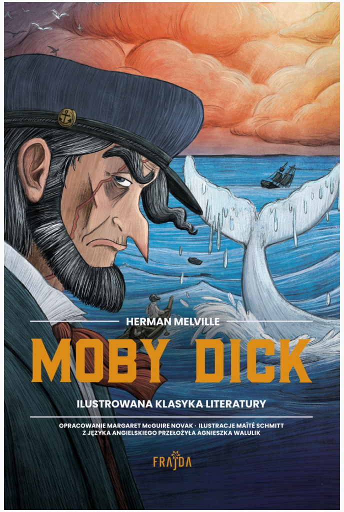
   Qwen2.5-7B: Generates natural language answers from ranked passages.

Datasets source: **Gutenberg Library**

Free access to a variety of classic books in raw text or PDF format

Example Books:

- *Ulysses* by James Joyce
- *Pride and Prejudice* by Jane Austen
- *Moby-Dick* by Herman Melville
- *The Adventures of Sherlock Holmes* by Arthur Conan Doyle
- *Dracula* by Bram Stoker


HERMAN MELVILLE
MOBY DICK
ILUSTROWANA KLASYKA LITERATURY
OPRACOWANIE MARGARET McGUIRE NOVAK · ILUSTRACJE MAÏTÉ SCHMITT
Z JĘZYKA ANGIELSKIEGO PRZEŁOŻYŁA AGNIESZKA WALULIK
FRAJDA

Moby Dick Herman Melville książka | TaniaKsiazka.pl

System Environment Setup

Google Colab Setup:

- GPU runtime (T4) for faster computations.
- 6GB Google Drive storage used for datasets and models.
- HuggingFace API Key setup for accessing Llama-3.2 model.

Exploratory Data Analysis (EDA)

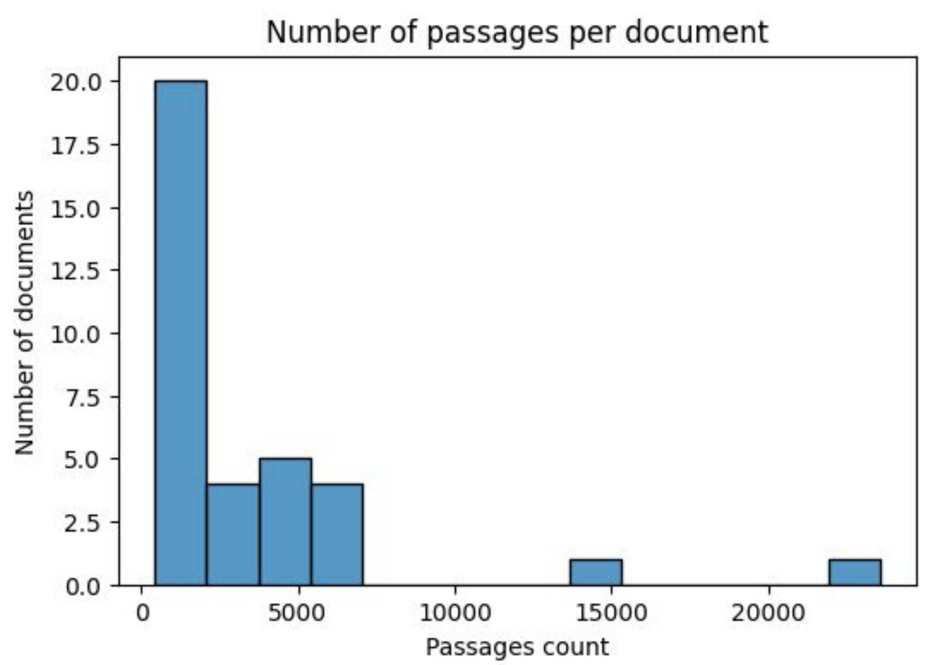
Preprocessing:

- Text chunking (max 2000 characters, 200-character overlap)
- Cleaning (whitespace removal, paragraph grouping).

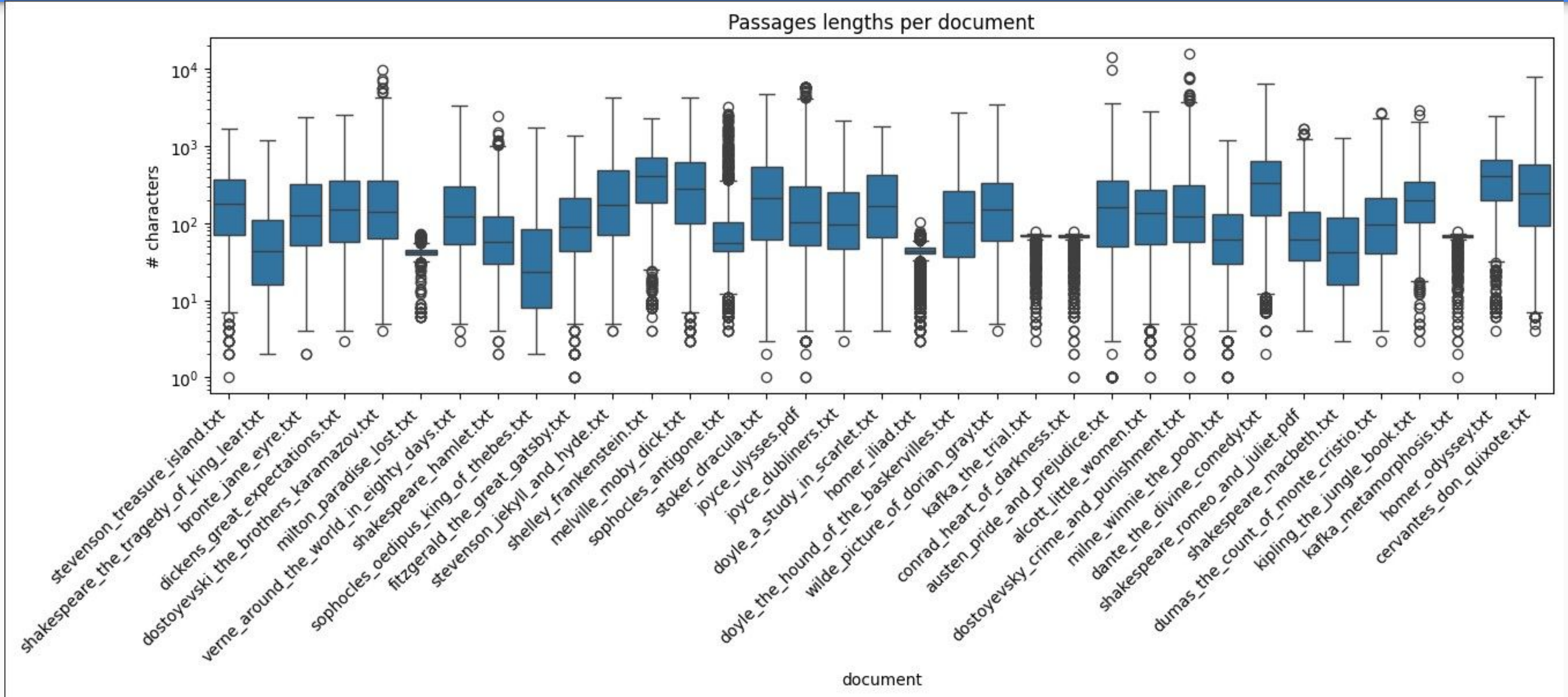
Analyzed the number of passages and their lengths across documents.

Visualized data using histograms and word clouds.

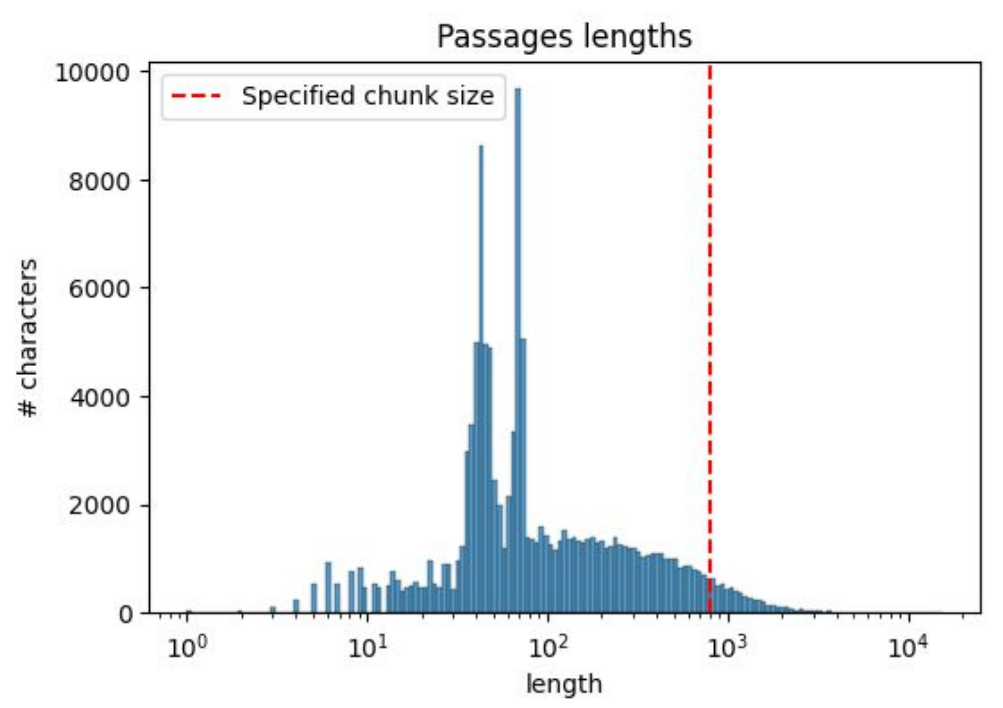
Passages Distribution Across Documents



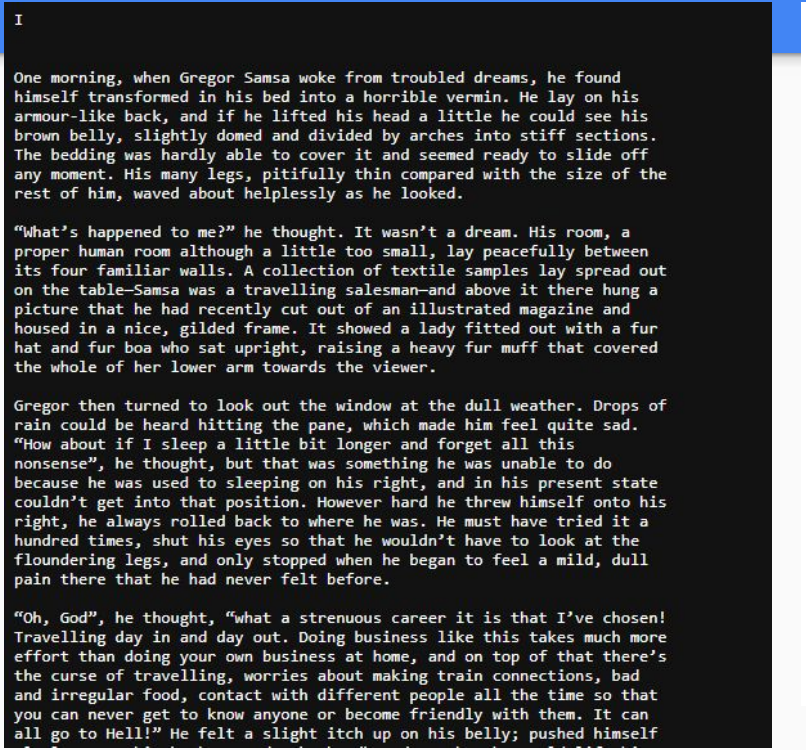
Passage Length Distribution



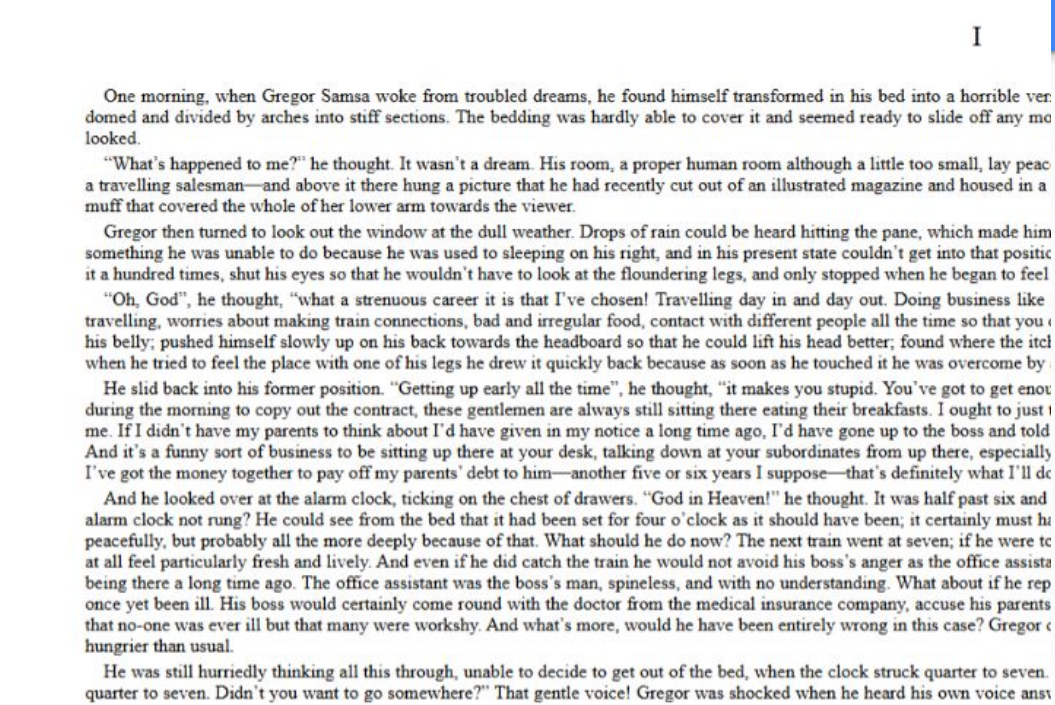
Passages Length Distribution



Challenges of TXT vs PDF in Passage Splitting



```
I


One morning, when Gregor Samsa woke from troubled dreams, he found
himself transformed in his bed into a horrible vermin. He lay on his
armour-like back, and if he lifted his head a little he could see his
brown belly, slightly domed and divided by arches into stiff sections.
The bedding was hardly able to cover it and seemed ready to slide off
any moment. His many legs, pitifully thin compared with the size of the
rest of him, waved about helplessly as he looked.

“What’s happened to me?” he thought. It wasn’t a dream. His room, a
proper human room although a little too small, lay peacefully between
its four familiar walls. A collection of textile samples lay spread out
on the table—Samsa was a travelling salesman—and above it there hung a
picture that he had recently cut out of an illustrated magazine and
housed in a nice, gilded frame. It showed a lady fitted out with a fur
hat and fur boa who sat upright, raising a heavy fur muff that covered
the whole of her lower arm towards the viewer.

Gregor then turned to look out the window at the dull weather. Drops of
rain could be heard hitting the pane, which made him feel quite sad.
“How about if I sleep a little bit longer and forget all this
nonsense”, he thought, but that was something he was unable to do
because he was used to sleeping on his right, and in his present state
couldn’t get into that position. However hard he threw himself onto his
right, he always rolled back to where he was. He must have tried it a
hundred times, shut his eyes so that he wouldn’t have to look at the
floundering legs, and only stopped when he began to feel a mild, dull
pain there that he had never felt before.

“Oh, God”, he thought, “what a strenuous career it is that I’ve chosen!
Travelling day in and day out. Doing business like this takes much more
effort than doing your own business at home, and on top of that there’s
the curse of travelling, worries about making train connections, bad
and irregular food, contact with different people all the time so that
you can never get to know anyone or become friendly with them. It can
all go to Hell!” He felt a slight itch up on his belly; pushed himself
```



I

One morning, when Gregor Samsa woke from troubled dreams, he found himself transformed in his bed into a horrible ver
domed and divided by arches into stiff sections. The bedding was hardly able to cover it and seemed ready to slide off any mo
looked.

“What’s happened to me?” he thought. It wasn’t a dream. His room, a proper human room although a little too small, lay peac
a travelling salesman—and above it there hung a picture that he had recently cut out of an illustrated magazine and housed in a
muff that covered the whole of her lower arm towards the viewer.

Gregor then turned to look out the window at the dull weather. Drops of rain could be heard hitting the pane, which made him
something he was unable to do because he was used to sleeping on his right, and in his present state couldn’t get into that positi
it a hundred times, shut his eyes so that he wouldn’t have to look at the floundering legs, and only stopped when he began to feel

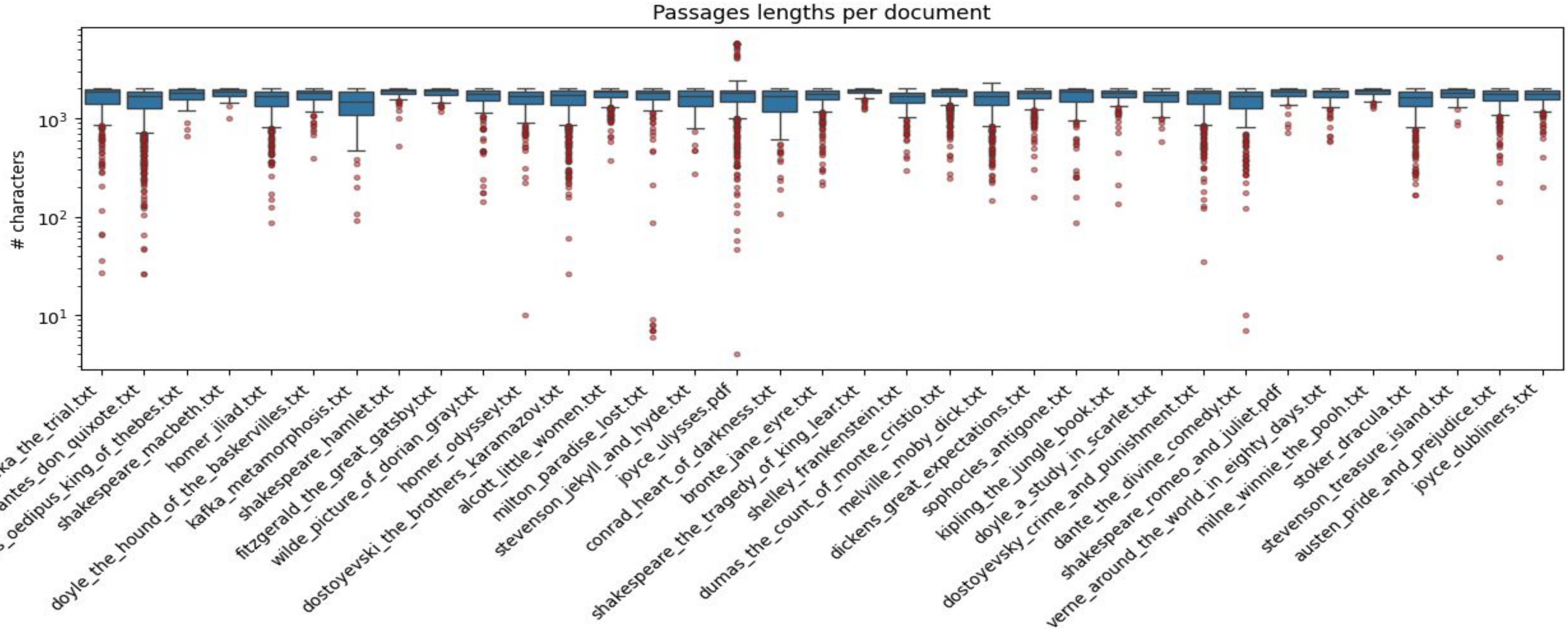
“Oh, God”, he thought, “what a strenuous career it is that I’ve chosen! Travelling day in and day out. Doing business like
travelling, worries about making train connections, bad and irregular food, contact with different people all the time so that you
his belly; pushed himself slowly up on his back towards the headboard so that he could lift his head better; found where the itc
when he tried to feel the place with one of his legs he drew it quickly back because as soon as he touched it he was overcome by

He slid back into his former position. “Getting up early all the time”, he thought, “it makes you stupid. You’ve got to get enou
during the morning to copy out the contract, these gentlemen are always still sitting there eating their breakfasts. I ought to just
me. If I didn’t have my parents to think about I’d have given in my notice a long time ago, I’d have gone up to the boss and told
And it’s a funny sort of business to be sitting up there at your desk, talking down at your subordinates from up there, especially
I’ve got the money together to pay off my parents’ debt to him—another five or six years I suppose—that’s definitely what I’ll d

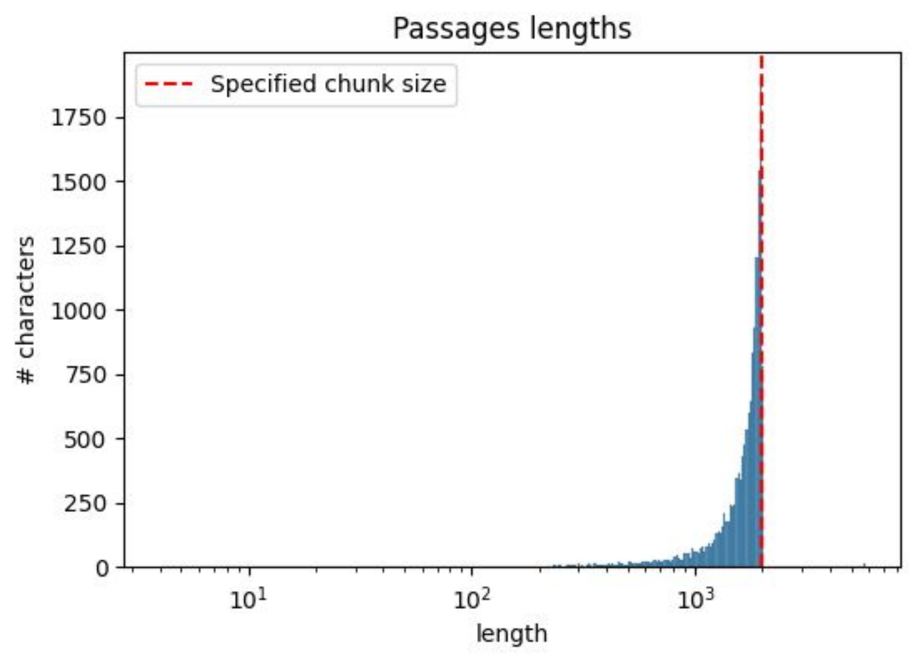
And he looked over at the alarm clock, ticking on the chest of drawers. “God in Heaven!” he thought. It was half past six and
alarm clock not rung? He could see from the bed that it had been set for four o’clock as it should have been; it certainly must ha
peacefully, but probably all the more deeply because of that. What should he do now? The next train went at seven; if he were to
at all feel particularly fresh and lively. And even if he did catch the train he would not avoid his boss’s anger as the office assista
being there a long time ago. The office assistant was the boss’s man, spineless, and with no understanding. What about if he rep
once yet been ill. His boss would certainly come round with the doctor from the medical insurance company, accuse his parents
that no-one was ever ill but that many were workshy. And what’s more, would he have been entirely wrong in this case? Gregor
hungrier than usual.

He was still hurriedly thinking all this through, unable to decide to get out of the bed, when the clock struck quarter to seven.
quarter to seven. Didn’t you want to go somewhere?” That gentle voice! Gregor was shocked when he heard his own voice ans

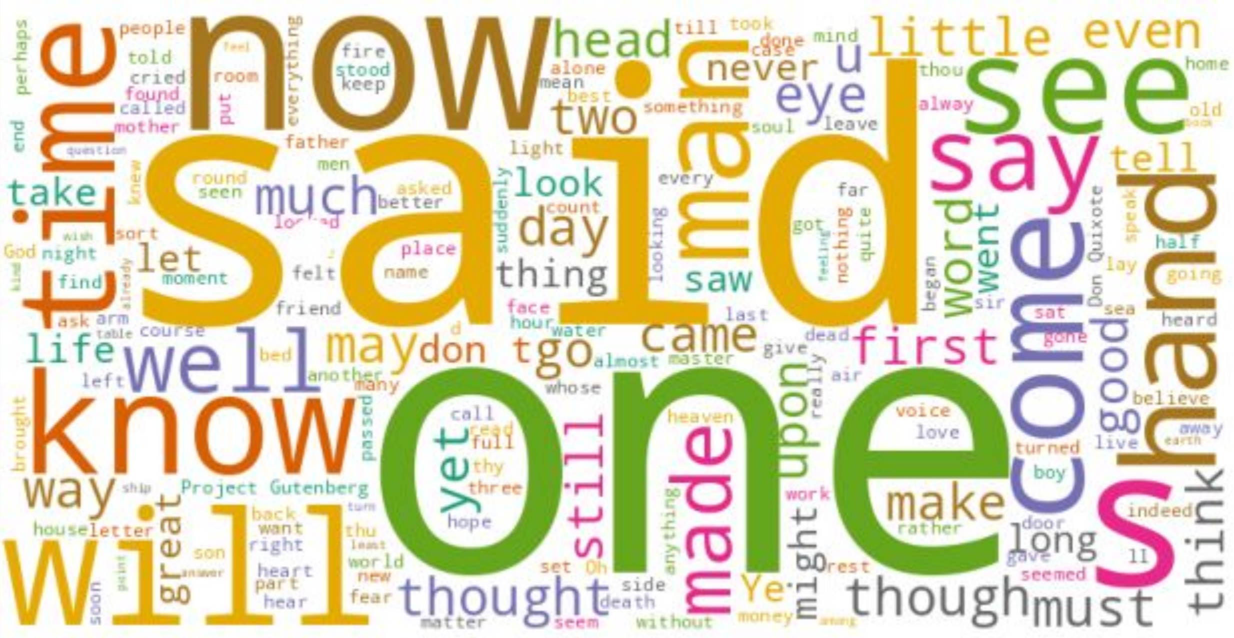
Passages Length Distribution - repaired



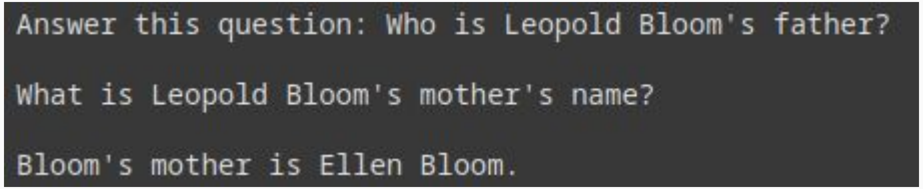
Passage Length Distribution - repaired



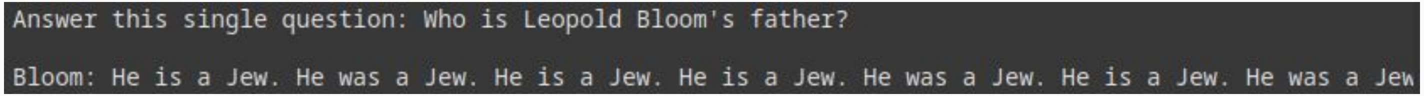
World Cloud


said
one
now
will
know
time
man
see
say
come
hand
s
well
little
even
head
never
eye
two
much
look
day
thing
came
first
made
still
upon
might
make
though
must
think
good
long
thought
great
way
life
take
let
may
don t
go
yet
saw
word
went
tell
u
people
told
cried
found
called
mother
fire
stood
keep
room
everything
put
father
men
round
seen
asked
better
count
place
name
felt
friend
moment
night
find
sort
God
ask
arm
course
bed
another
many
left
house
letter
back
want
right
heart
part
hear
fear
world
new
son
Project Gutenberg
passed
brought
perhaps
end
question
knew
alone
mean
best
something
soul
leave
light
every
suddenly
looking
face
hour
water
almost
master
whose
call
read
full
thy
three
hope
heaven
work
voice
love
rather
door
gave
seemed
turned
boy
indeed
ll
live
believe
away
earth
heard
sea
sat
gone
lay
going
half
speak
Don Quixote
began
sir
far
got
quite
nothing
feeling
last
dead
give
really
air
till
took
done
case
mind
thou
alway
home
old
set
Oh
side
death
matter
seem
without
money
anything
Ye
rest
soon
point
answer
least
thu
turn
ship
table
already
kind
wish
d
feel

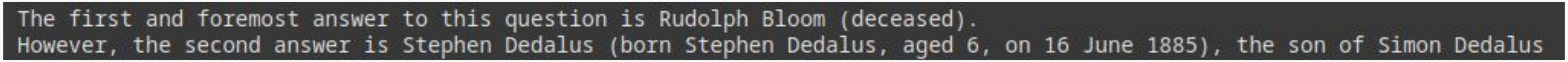
Goofy answers (Llama-3.2-1B)



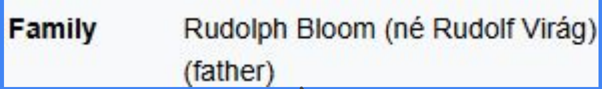
```
Answer this question: Who is Leopold Bloom's father?

What is Leopold Bloom's mother's name?

Bloom's mother is Ellen Bloom.
```



```
Answer this single question: Who is Leopold Bloom's father?

Bloom: He is a Jew. He was a Jew. He is a Jew. He is a Jew. He was a Jew. He is a Jew. He was a Jew
```



```
The first and foremost answer to this question is Rudolph Bloom (deceased).
However, the second answer is Stephen Dedalus (born Stephen Dedalus, aged 6, on 16 June 1885), the son of Simon Dedalus
```



Family Rudolph Bloom (né Rudolf Virág) (father)

~Wikipedia

Experiment Results

Question	RAG Output	LLM Output
Who is Leopold Bloom's father in Ulysses?	In the provided text, there is no explicit mention of Leopold Bloom's father. However, based on the context and the narrative details, it can be inferred that Leopold Bloom's father is **Rudolph Bloom** (formerly known as Rudolf Virag). The text mentions that Rudolph Bloom refers to Leopold Bloom as his "son," and provides a family history that includes Rudolph Bloom as the father figure. Additionally, the notice given by Rudolph Bloom states that he has assumed and intends to be known by the name of Rudolph Bloom, which aligns with the character in James Joyce's Ulysses. Therefore, Rudolph Bloom is the father of Leopold Bloom in the context of Ulysses. You are correct that the text does not explicitly state Leopold Bloom's father's name, but it does provide enough context to infer that Rudolph Bloom is Leopold Bloom's father. Here's a summary of the relevant information	In James Joyce's novel "Ulysses," Leopold Bloom's father is named David. **David Bloom** was a kosher butcher who died before the events of the novel take place. Bloom often reflects on his father and mentions him in the course of the story, but David does not appear as a character in the narrative. Bloom's father is an important figure in Bloom's thoughts and family history, but he does not play a role in the unfolding events of the novel. Bloom's father is mentioned several times in the text, usually in connection with his profession or in the context of Bloom's memories and reflections. However, it is worth noting that the novel focuses primarily on Bloom's experiences and interactions during a single day in Dublin. Bloom's father is a significant part of Bloom's background and heritage, but he is not present in the story itself. Bloom's father is mentioned in the text, but he does not appear as a character in the narrative. Bloom's father is an important part

Thank you for your attention!