

# Smart PDF Chatbots

Enhancing Customer Support with AI that  
Understands Your Documents

Bassel

Isaak

Murilo

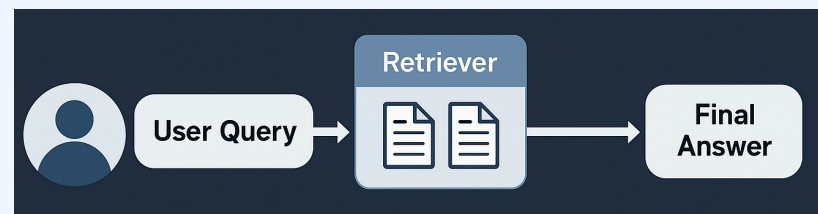
[04.04.2025]

# The Challenge

- Customers often struggle to get quick, accurate answers from large and complex documents like PDFs.
- How can we make document access smarter, faster, and easier?

# Our Solution




- We've built a smart AI chatbot that can:
  - - Understand the content of long documents
  - - Instantly answer user questions
  - - Provide consistent and accurate information
  - - Powered by cutting-edge AI language models and retrieval technology.
  - - You can choose Gemini or OpenAI as AI Model



# How It Works (Simplified)

- 1. Document Upload – Start with your PDF content
- 2. Content Processing – AI breaks it down into manageable parts
- 3. Knowledge Storage – Creates a searchable knowledge base
- 4. Conversational Access – Users chat with AI for instant answers

# Business Benefits

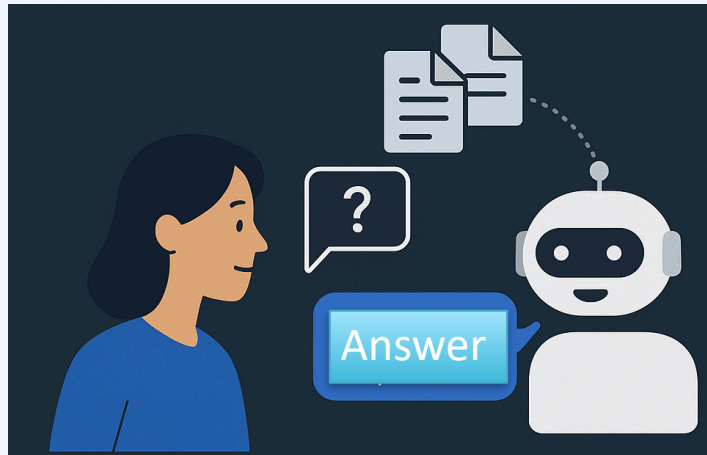
-  Accurate answers from your content
-  Reduced support load
-  Enhanced customer experience

# Example Use Cases

- - Product manuals
- - Policy documents
- - Internal knowledge bases
- - Customer support FAQs

# Examples

question = "What are the key benefits of using BERT in NLP tasks?"



## BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova  
Google AI Language  
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

### Abstract

We introduce a new language representation model called **BERT**, which stands for **Bidirectional Encoder Representations from Transformers**. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

Based on the provided text, the key benefits of using BERT in NLP tasks are:

- **Fine-tuning approach:** BERT can be fine-tuned directly on downstream tasks, allowing task-specific models to benefit from the larger, more expressive pre-trained representations, even with small downstream task data.
- **Bidirectionality:** BERT's pre-training tasks and bidirectionality contribute to improvements.
- **Unified approach:** BERT uses the self-attention mechanism to unify the encoding and decoding stages, effectively including bidirectional cross-attention between sentences.
- **Flexibility:** BERT can be adapted to various tasks by plugging in task-specific inputs and outputs and fine-tuning all parameters end-to-end.
- **Minimal new parameters:** Task-specific models are formed by incorporating BERT with one additional output layer, so a minimal number of parameters need to be learned from scratch.

# Smart Summaries from PDF Documents

Turn complex documents into clear, concise insights with AI

Bassel

Isaak

Murilo

[04.04.2025]



# The Need for Summaries

- Long PDFs are time-consuming to read and often underutilized.
- > What if your documents could summarize themselves?




# Our Solution

- We built a solution that uses AI to summarize PDF content in seconds.
- It provides:
  - - Short, concise overviews
  - - Bullet-point summaries
  - - Detailed breakdowns
- Based on your exact needs.
- You can choose Gemini or OpenAI as AI Model





# How It Works (Simplified)

- 1. PDF Upload & Chunking – The document is broken down for easier processing.
- 2. AI Summarization Techniques – Choose from:
  - - First 5 pages summary
  - - Short overall summary
  - - Detailed bullet-point summary
- 3. Instant Insight – Tailored summaries are delivered in real-time.

# Types of Summaries

-  Brief Overview (First 5 Pages)
- Just want the beginning? Get a quick start.
-  Short Full Summary
- Need the whole picture at a glance?
-  Detailed Bullet Summary
- Get rich, structured summaries with key points for decision-making.

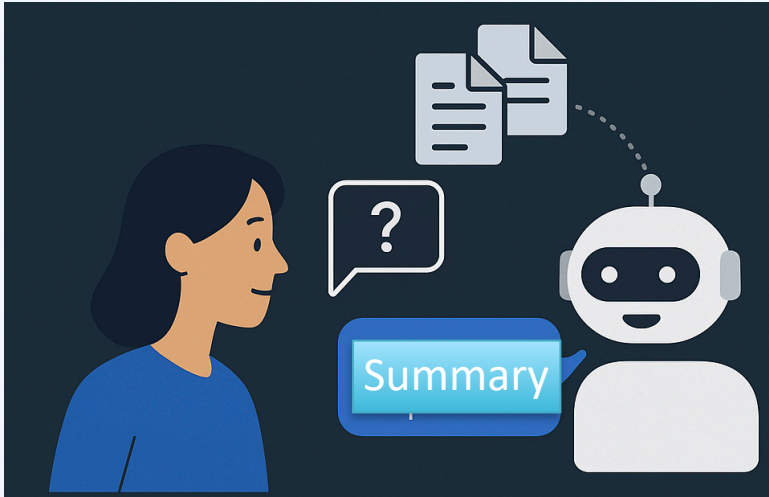
# Customer Benefits

-  Saves time and manual reading
-  Boosts understanding and decision-making
-  Works with any PDF document
-  Flexible levels of detail

# Use Cases

- - Policy and legal document summaries
- - Product specification digests
- - Internal reports and research papers
- - Meeting or training material recap

# Example



## BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova  
Google AI Language  
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

### Abstract

We introduce a new language representation model called **BERT**, which stands for **Bidirectional Encoder Representations from Transformers**. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

### Brief Summary:

BERT is a powerful, pre-trained language model based on a bidirectional Transformer architecture. It achieves state-of-the-art results on numerous NLP tasks by being fine-tuned with minimal modifications. BERT uses a Masked Language Model (MLM) and Next Sentence Prediction (NSP) during pre-training to learn contextual word representations. It outperforms previous models like OpenAI GPT and ELMo, and is robust to hyperparameter choices. Research explores BERT's performance on benchmarks like GLUE and SQuAD, analyzes the impact of pre-training tasks and model size, and compares fine-tuning and feature-based approaches.

# Project Notebooks:

- [https://github.com/bassel-zdn/DS\\_CapstoneProjects/blob/main/Project\\_1/Gemini\\_Project1.ipynb](https://github.com/bassel-zdn/DS_CapstoneProjects/blob/main/Project_1/Gemini_Project1.ipynb)
- [https://github.com/bassel-zdn/DS\\_CapstoneProjects/blob/main/Project\\_1/OpenAI\\_Project1.ipynb](https://github.com/bassel-zdn/DS_CapstoneProjects/blob/main/Project_1/OpenAI_Project1.ipynb)