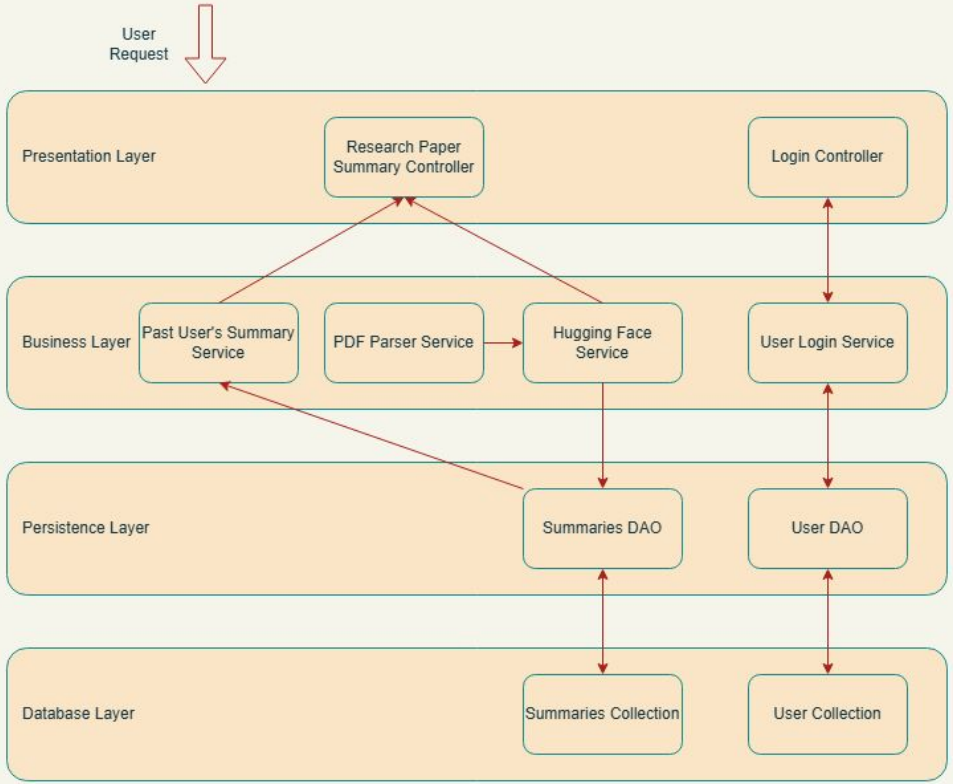
TLDResearch

Mid-Way Report

2/23/2025

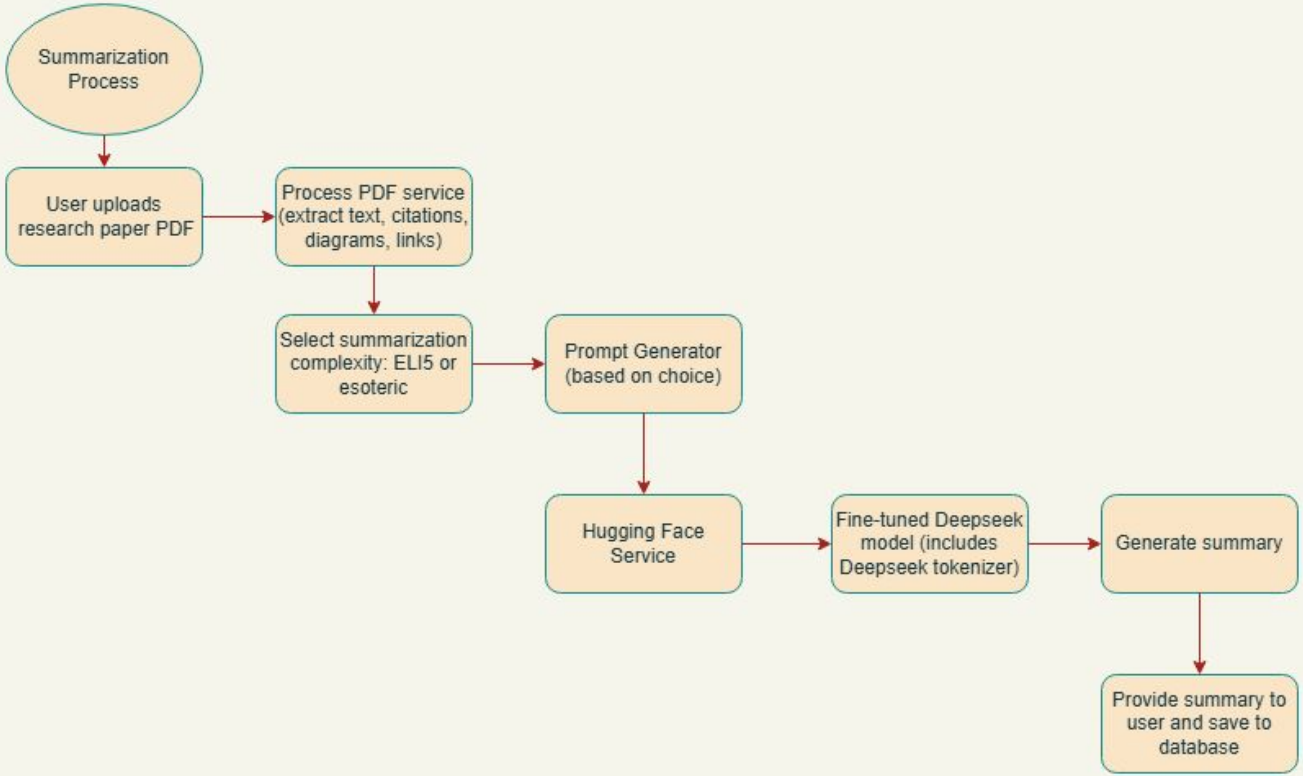
Architecture Pattern and Design Illustration



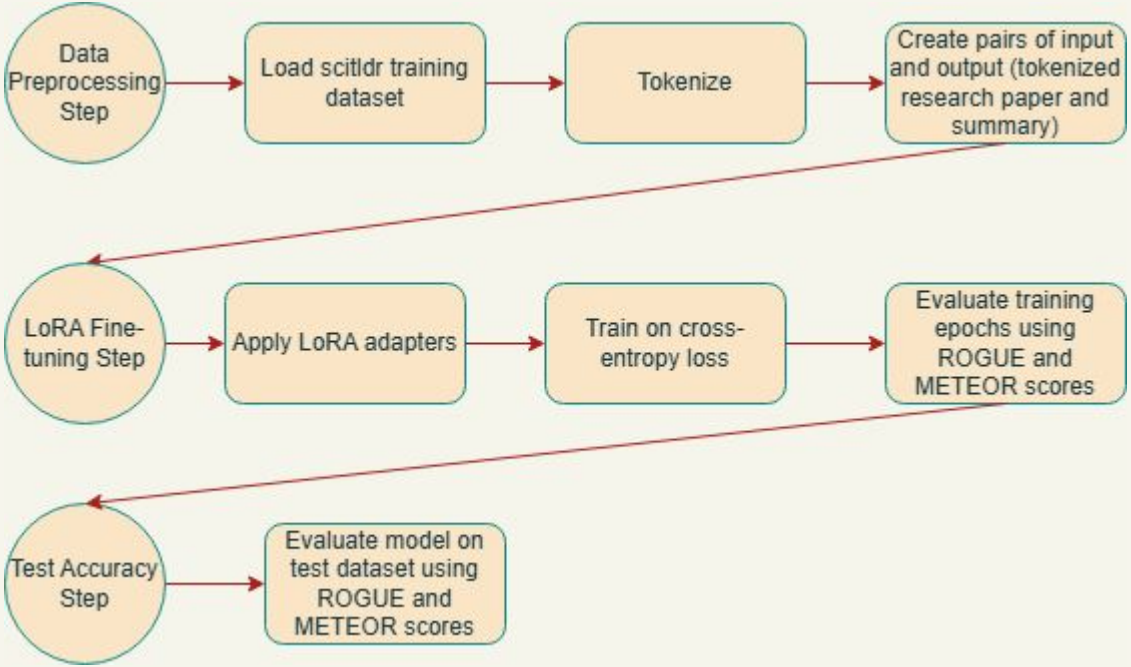
MVP Approach

- TLDResearch will consist of a box to drop a paper which the user wants to summarize. TLDResearch will then provide a summarization of the paper. The user can also look at previous generations.

Detailed Workflow / Pipeline (business interaction)



Detailed Workflow / Pipeline (Model fine-tuning (LoRA))



Tech Stack Decisions

- **Backend**: Django - provides a framework which simplifies API calls to handle research paper data fed to the Database
- **Frontend**: React - Powerful frontend to simplify implementation while providing seamless user experience
- **Database**: MongoDB - Flexible document based-structure which accommodates diverse data formats
- **Large Language Model**: Deepseek fine tuned using loRA - Testing effectiveness of new LLM , which is made lightweight enough to use via loRA.
- **Deployment**: Google Cloud - reliable and easily accessible environment for backend and LLM
- **Middlewares**: Hugging face transformers will be middleware between our backend and the LLM (we will train/fine-tune our model using our $50 credits and then put the checkpoint on hugging face)

Exemplar Data Points

The following slides contain 3 examples of input and expected output pairs that we will use to fine-tune our Deepseek model. Due to the large length of a research paper, we have provided excerpts to the full paper (as well as a link to the full text) along with the expected summary.

Exemplar Data Point #1

Input (paper): Due to the success of deep learning to solving a variety of challenging machine learning tasks, there is a rising interest in understanding loss functions for training neural networks from a theoretical aspect. Particularly, the properties of critical points and the landscape around them are of importance to determine the convergence performance of optimization algorithms. In this paper, we provide a necessary and sufficient characterization of the analytical forms for the critical points (as well as global minimizers) of the square loss functions for linear neural networks.

Output (summary): We provide necessary and sufficient analytical forms for the critical points of the square loss functions for various neural networks, and exploit the analytical forms to characterize the landscape properties for the loss functions of these neural networks.

Link: https://arxiv.org/abs/1710.11205

Exemplar Data Point #2

Input (paper): The backpropagation (BP) algorithm is often thought to be biologically implausible in the brain. One of the main reasons is that BP requires symmetric weight matrices in the feedforward and feedback pathways. To address this \u201cweight transport problem\u201d (Grossberg, 1987), two biologically-plausible algorithms, proposed by Liao et al. (2016) and Lillicrap et al. (2016), relax BP\u2019s weight symmetry requirements and demonstrate comparable learning capabilities to that of BP on small datasets. However, a recent study by Bartunov et al. (2018) finds that although feedback alignment (FA) and some variants of target-propagation (TP) perform well on MNIST and CIFAR, they perform significantly worse than BP on ImageNet.

Output (summary): Biologically plausible learning algorithms, particularly sign-symmetry, work well on ImageNet.

Link: https://arxiv.org/abs/1811.03567

Exemplar Data Point #3

Input (paper): We introduce the 2-simplicial Transformer, an extension of the Transformer which includes a form of higher-dimensional attention generalising the dot-product attention, and uses this attention to update entity representations with tensor products of value vectors. We show that this architecture is a useful inductive bias for logical reasoning in the context of deep reinforcement learning.\n Deep learning contains many differentiable algorithms for computing with learned representations. These representations form vector spaces, sometimes equipped with additional structure. A recent example is the Transformer (Vaswani et al., 2017) in which there is a vector space V of value vectors and an inner product space H of query and key vectors.

Output (summary): We introduce the 2-simplicial Transformer and show that this architecture is a useful inductive bias for logical reasoning in the context of deep reinforcement learning.

Link: https://arxiv.org/abs/1909.00668

Naive Baseline Performance

The following three slides contain the Naive Baseline Performance of the three exemplar data points, when using ChatGPT-4o to create summaries. To evaluate the performance of the summaries, METEOR was used to provide baseline data for our team.

Naive Performance #1

Reference Solution: We provide necessary and sufficient analytical forms for the critical points of the square loss functions for various neural networks, and exploit the analytical forms to characterize the landscape properties for the loss functions of these neural networks

Model Provided Summary: In this paper, we provide a necessary and sufficient characterization of the critical points of square loss functions for linear and shallow ReLU networks, demonstrating that while linear networks have no spurious local minima, ReLU networks can. Our work extends prior research by removing assumptions on data matrices and network dimensions, offering new proofs, and deepening the theoretical understanding of neural network optimization landscapes.

METEOR SCORE: 0.380

Naive Performance #2

Reference Solution: Biologically plausible learning algorithms, particularly sign-symmetry, work well on ImageNet

Model Provided Summary: In this work, we investigate the biological plausibility of learning algorithms by evaluating sign-symmetry (SS) and feedback alignment (FA) on large-scale datasets like ImageNet and MS COCO, using standard architectures such as ResNet-18, AlexNet, and RetinaNet. Our results show that sign-symmetry, which relaxes backpropagation's weight symmetry requirement by preserving only the sign of the feedforward weights in feedback pathways, achieves performance close to backpropagation, establishing it as a viable alternative for both artificial neural network training and biologically plausible learning in the brain.

METEOR SCORE: 0.29544

Naive Performance #3

Reference Solution: We introduce the 2-simplicial Transformer and show that this architecture is a useful inductive bias for logical reasoning in the context of deep reinforcement learning.

Model Provided Summary: We introduce the 2-simplicial Transformer, an extension of the Transformer architecture that incorporates higher-dimensional attention to capture logical reasoning in reinforcement learning tasks, particularly for environments like BoxWorld. Our experiments demonstrate that the 2-simplicial Transformer outperforms standard models by efficiently learning higher-order relationships, leading to faster convergence and improved puzzle-solving capabilities.

METEOR SCORE: 0.426312

Competitive Product Analysis

Name	TLDResearch Built in AI, Summarizes/chatbot
Features	Free, for now. Summarizes PDF Saves previous generations Regeneration button
User Reviews	None

Competitive Product Analysis

Name	Scholarcy
Features	● Summarizes Text and Video ● Text can be in PDF, ebook, academic journal, plaintext, or google drive format ● Saves previous generations ● Export generations to a variety of formats ● Change complexity level ● Pulls references cited in paper and compiles them ● Extract data from visualizations
User Reviews	● Does not provide chat feature ● Free version limits to 3 summaries / day ● No saving, organizing, or customizing summaries with free version ● Paid version $4.99 ● Best alternative is elephas, which has no free tier ● Sometimes hallucinates. ● Segments can be uninformative and arbitrarily cut off.

Competitive Product Analysis

Name	Papers.cool Summarize pre existing research papers in Chinese
Features	Summarizes papers only in Chinese Saves previous generations Only papers that already exists, cannot upload PDF Cannot chat with bot or change response only one summary Uses KIMI
User Reviews	No user reviews Example : https://papers.cool/arxiv/2407.05231 Source: https://github.com/bojone/papers.cool

Competitive Product Analysis

Name	Adobe Acrobat, Built in AI, Summarizes/chatbot
Features	Can summarize PDF's in different languages Priced at $5 a month for individuals and $2 for students Can chat with bot to specify summarization Can save chats for the time you use it but unsure if you can save all your conversations from different pdfs Prompts follow up questions for users to ask
User Reviews	https://www.reddit.com/r/sysadmin/comments/1buv7vh/update_as_long_as_you_dont_push_the_shiny_ai/ "Adobe has recently faced scrutiny over concerns that its AI features within Acrobat and Reader may be secretly scanning documents. However, according to Windows Latest, citing a response from the company, that's not the case." "Like other Adobe AI features, Adobe Acrobat AI Assistant was developed and deployed in alignment with Adobe's AI principles of accountability, responsibility and transparency. In addition, the features are governed by data security, privacy and AI ethics protocols and no customer content is used to train LLMs without customers' consent. "To me, they're not being transparent in this at all, given that they hide how to disable the generative AI features org-wide. Even now, their help article still says: Note: If you're an admin and want to revoke access to generative AI features for your team or org, contact Adobe Customer Care."

Conclusion

TLDR

We summarize papers so you can digest research more efficiently.