Impactful Resilient Infrastructure Science and Engineering (IRISE)

-Project Scope of Work-

(FY 2023-24 (IRISE Year 6) Annual Work Program)

SUMMARY PAGE

Project Title: Adaptation of a Large Language Model for Facilitating Pavement-Related Information Retrieval and Knowledge Discovery

Person Submitting Proposal: Dr. Lev Khazanovich

Proposed Funding Period: January 1, 2024 - December 31, 2025

Project Duration: 24 months

Project Cost: $204,543.

Project Title: Adaptation of a Large Language Model for Facilitating Pavement-Related Information Retrieval and Knowledge Discovery

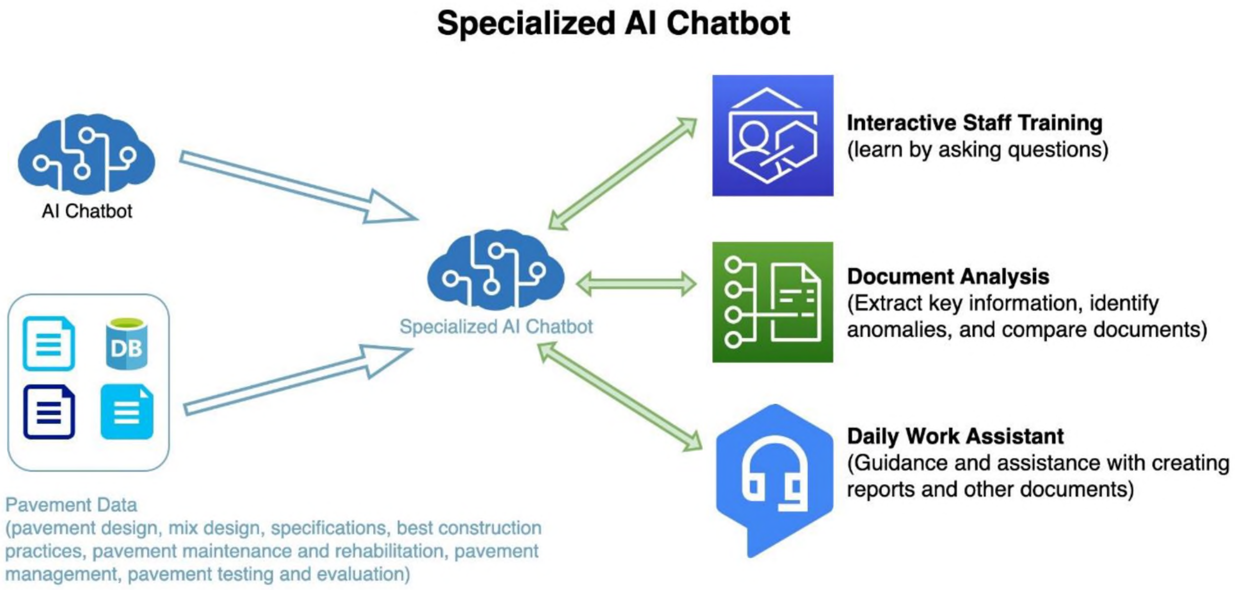
Problem Statement:

The recent introduction of Large Language Models (LLMs), such as OpenAI's GPT-3, Google's T5, and Facebook's RoBERTa, has generated significant excitement and interest. These artificial intelligence models are trained on vast amounts of text data to learn the patterns and relationships in natural language. They can generate new text that is likely to be grammatically correct and semantically meaningful. Language models offer many opportunities for pavement engineering improvement. For example, they could quickly answer common engineering questions, perform common calculations, or analyze large volumes of pavement-related data to identify patterns and trends that humans may not immediately recognize. This can help pavement engineers make more informed decisions about pavement design, construction, and maintenance.

Although ChatGPT has generated the most media attention in the last few months, there are several other systems similar to ChatGPT, including OpenAI GPT-2, Google's BERT, Microsoft's Turing-NLG, and Facebook's RoBERTa. Each of these systems has its strengths and weaknesses, but all of them are general-purpose systems trained on non-specialized datasets. As a result, they lack domain-specific knowledge and may answer technical questions incorrectly. However, LLMs can be finetuned to improve their robustness for specific knowledge domains. In the process known as domain adaptation, an LLM pre-trained on a large corpus of general text is additionally trained on a smaller, domain-specific corpus of text. Thus, the model learns domain-specific language patterns, terminology, and concepts, and it performs better on tasks related to the domain.

Project Objectives:

Our goal is to create an Artificial Intelligence (AI) model that can process and generate natural language to answer common pavement engineering and pavement construction questions quickly and accurately. The model will be able to answer questions about pavement design, mix design, specifications, best construction practices, pavement maintenance and rehabilitation, pavement management, as well as pavement testing and evaluation. It will have a human-like ability to provide reliable answers and will be suitable for interactive training. The tool will have the capability to assess a user's knowledge by posing questions and offering feedback on their answers, making it an effective tool for skill development and evaluation. Additionally, the model will be able to assist pavement engineers in retrieving and summarizing relevant information from a large corpus of pavement-related literature. This can facilitate knowledge discovery and help pavement engineers stay up-to-date with the latest developments in their field.


Specialized AI Chatbot
AI Chatbot
Specialized AI Chatbot
DB
Pavement Data
(pavement design, mix design, specifications, best construction practices, pavement maintenance and rehabilitation, pavement management, pavement testing and evaluation)
Interactive Staff Training
(learn by asking questions)
Document Analysis
(Extract key information, identify anomalies, and compare documents)
Daily Work Assistant
(Guidance and assistance with creating reports and other documents)

Project Scope:
The project aims to develop a specialized AI language model for pavement engineering by identifying an appropriate platform and exposing it to a large corpus of relevant text data, including PennDOT manuals, training materials, research reports, as well as reports from the FHWA, industry sources, and the Transportation Research Board. The model will be optimized to predict the next word in a sequence based on the context of previous words and will be finetuned through repeated iterations to learn patterns in the data. The model will be further finetuned on specific tasks, such as answering questions and generating educational materials. The final product will be a system that can generate learning materials, simulate real-world scenarios for problem-solving and decision-making practice, and assist users in accessing and understanding relevant information.

Proposed Work:
The objectives of this project will be realized through the completion of the following tasks:

Task A – Review available LLMs

Available open-source LLMs will be reviewed. Information on a variety of LLMs, including OpenAI GPT-2 and GPT-3, Google BERT (Bidirectional Encoder Representations from Transformers), Google T5 (Text-to-Text Transfer Transformer), Facebook RoBERTa, Facebook LLaMA (Large Language Model Meta AI), Hugging Face DistilBERT, Stanford Alpaca, and other products, will be collected and evaluated using the following criteria:

- Generalization: The ability to generalize to new, unseen data through cross-validation.
- Robustness to Noise and Errors: The ability to handle noisy or erroneous input data, which is common in real-world applications. This can be assessed by introducing noise or errors into the input data and evaluating the model's performance.

- Computational Efficiency: Training time, response time, and memory usage. This is particularly important because the application will require real-time processing.
- Model Size: A smaller model size may be less universal but would require a less powerful computer.
- Interpretability: The ability to understand how the model is making predictions.

Based on the results of this evaluation, the most suitable LLM for the pavement-related knowledge domain will be selected. The best efforts will be made to identify the most suitable model having a potential to be utilized on a laptop computer. This will include a specific software that implements machine learning algorithms, including the code and pre-trained data, also known as "weights."

Task B – Select domain knowledge sources and prepare them for finetuning

Relevant pavement-related data will be collected and cleaned. The data should be representative of the pavement knowledge domain and should cover a wide range of pavement engineering. The data will be collected from various sources, such as academic publications, FHWA, state DOTs, and industry research reports, PennDOT manuals, training materials, etc. Once the data is collected, it will be cleaned to remove any irrelevancies or redundancies, such as duplicates, incomplete records, and irrelevant text, to ensure that the data is consistent and of high quality.

In the next step, each document will be converted into a file in plain text format. A Python script using open-source libraries, such as PyPDF2, will be developed to automate this conversion. The extracted textual data will be further filtered to remove the sections of the documents unlikely to contain textual information. Finally, for each text file, a synthetic question-answer (QA) dataset will be generated using state-of-the-art LLMs.

Graduate and undergraduate student volunteers will be recruited to perform and check randomly selected questions and answers. This will help ensure that the dataset is accurate and representative of the domain. Whenever possible, multiple people will be asked to validate the same examples. This will help identify any inconsistencies or errors in generation and improve the overall quality of the dataset.

Task C – LLM finetuning

Using the dataset prepared in Task B, the LLM selected in Task A will be The model's parameters, such as weights and biases, will be optimized, improving its accuracy and efficiency. Various finetuning approaches will be considered, such as prompting, few-shot learning with exemplars, the chain of thoughts, prefix-tuning, low-rank rank adaption (LoRA), and adapters (e.g., LLaMA-Adapter). The finetuned model will learn to extract and represent the most relevant information for a given task while filtering out noise and irrelevant information. Finetuning will improve the robustness of the model to variations in the input data, such as differences in writing style, syntax, and terminology. This will enhance the model's generalization capabilities and enable it to handle a wider range of input data.

The following steps comprise the anticipated finetuning process:

1. Our pavement-related dataset, prepared in Task B, will be split into training and test sets with a Python script along with the free, open-source scikit-learn machine learning tool for Python. 80% of the data will be used for training and 20% for testing. These dataset splits will be used as a starting point.
2. The training dataset will be expanded with adversarial questions and answers. The dataset will comprise positive (correct question, answer, context) and negative examples, such as random context paired with a question, or a pair of negative examples (a question based on the same document section and another most similar to the context).
3. The Python code will create training and testing datasets, and the dataset creation process will be similar for both the finetuning and discriminator models, which check whether the generated output is relevant. The process will be applied separately for training and testing sets to avoid examples from the training set appearing in the test set.
4. The dataset will be formatted for finetuning. Depending on the LLM selected in Task A, additional formatting of the datasets for finetuning might be required using the freely available tools that LLM creators provide or recommend. This step might require developing Python code for automating this step.
5. Finally, the finetuning will be performed using scripts from LLM creators and a machinelearning library such as Pytorch. The finetuned LLM state, the final checkpoint, will be saved to a file that will be used for running the model.

Task D: Finetuned LLM Testing

After the LLM is finetuned, it must be validated and tested to ensure that it provides meaningful and accurate responses. To test the finetuned model, a script will be developed using a free ML library, such as PyTorch. The script will load data from the checkpoint file produced in Task C and initiate inference by running data points into the LLM to generate output. The testing process will include various benchmarks, such as the testing dataset generated in Task C, common sense (BoolQ, PIQA, OpenBookQA, etc.), scientific (ScienceQA), and domain-specific pavement engineering benchmarks. Additionally, feedback from pavement engineering experts will be gathered to assess the quality of the LLM-generated responses. The performance of the model will be evaluated based on the frequency of syntax, logical and mathematical errors, and inconsistencies.

The pavement engineering experts will grade the finetuned LLM’s responses according to their relevance, using a three level scale (High, Medium, or Low). If the responses are rated as low-quality or non-relevant, the domain-specific LLM reinforcement learning update will be performed to improve the model's performance.

Task E: Draft final report

Results and observations from all previous Tasks will be compiled in a final report summarizing the process for developing the finetuned LLM model as well as the results of its evaluation. In addition, the User Guide and the recommendations for the best use of the tool will be provided.

Task F: Final report

A Final Report taking into consideration comments that were received on the Draft Final Report will be prepared and electronically dsubmitted to the technical panel for the final approval.

Deliverables:

- *Deliverable #1* – Task A: A memo report summarizing the literature review, due 5 months from the Notice to Proceed date.
- *Deliverable #2* – Task B: A memo report detailing the results of the training and testing data sets preparation, due 16 months from the Notice to Proceed date.
- *Deliverable #3* – Task C: A memo report detailing the LLM model finetuning, due 19 months from the Notice to Proceed date.
- *Deliverable #4* – Task D: A memo report documenting the results of the model testing, due 21 months from the Notice to Proceed date.
- *Deliverable #5* – Task E: A draft final report, due 22 months from the Notice to Proceed date.
- *Deliverable #6* – Task F: Final report, due 24 months from the Notice to Proceed date.

In addition to the deliverables listed above, it is also anticipated that the findings of this research will be published and presented at key technical conferences and in journal publications with a prior approval from the Technical Panel.

Key Personnel:

Principal Investigator: Dr. Lev Khazanovich is to provide the technical expertise, project management, and oversight on all project activities.

Co-Principal Investigator: Dr. Vandenbossche is to provide the technical expertise on tasks B and D of this project and assist the Principal Investigator in project management and oversight on all project activities.

Other Personnel:

Graduate Assistant 1 (Igor Sukharev) is a graduate student who will assist in tasks A through F of this project. Dr. Sukharev earned a PhD in Electrical Engineering from Voronezh State Technical University, Russia. He has 20 years of experience as a Software Engineer and Architect for various IT companies, including 15 years at IBM. Currently, he is pursuing a PhD in Civil Engineering at Pitt under the supervision of Prof. Khazanovich.

Schedule:

	2024				2025			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Task A. Review available LLMs	🟩	🟩						
Task B. Select and convert domain knowledge sources for finetuning		🟩	🟩	🟩	🟩	🟩		
Task C. Finetune LLM				🟩	🟩	🟩	🟩	
Task D. Test and valided finetuned LLM					🟩	🟩	🟩	
Task E. Prepare draft final report							🟩	🟩
Task F. Prepare final report								🟩

Proposed Person-Hours by Task:

Project Role	Name	Task A	Task B	Task C	Task D	Task E	Task F	Total
PI	Lev Khazanovich	16	48	40	48	30	20	202
Co-PI	Julie Vandenbossche		24		24			48
Grad Student 1	TBN	320	650	350	260	108	40	1728
Hourly Student 1	TBN	40	500	120	120	40		820

Budget: The total project cost is $204,543.

Acknowledged By:

Lev Khazanovich

Dr. Lev Khazanovich
Principal Investigator