Focus area: **Knowledgeable Conversational AI**

Title: **From Personalized Education to Scientific Discovery with AI: Rapid Deployment of AI Domain Experts**

PI: **Volodymyr Kindratenko (kindrtnk@illinois.edu)**

Executive summary

Technical description

We aim to research, design, and implement a conversational AI that can rapidly become an "expert" on a user-defined topic. Our team at NCSA has been developing an AI chatbot platform to create "teaching assistants" trained on course materials [1]. Our prototype is capable of rapidly ingesting instructor-supplied course materials (textbook, instructor notes, lecture slides, lecture videos) and using this information to ground GPT-4 model in order to respond to student questions related to the course. In the process of implementing this framework, we identified opportunities for such chatbots to become very effective in answering questions withing a narrowly defined area as well as challenges in adapting large language models (LLMs) for such tasks. With this proposal, we are looking to address some of these challenges and develop and demonstrate a framework capable of quickly becoming an "expert" on a user-defined topic. The proposed system will operate as follows:

1. A user specifies the source or subject, e.g., biology textbook for the class they are taking, a series of historical documents for the research paper they are writing, or a literature piece they are reading for the book club, to name a few.
2. LLM(s) is "trained" on these sources, relevant additional materials, and their derivatives. To include deeper levels of primary materials for each topic, we will include in the training corpora all references in the citation tree of each source. Additionally, separate LLMs will use the primary material to generate task-specific versions of the data, e.g., Q&A pairs, "explain like I'm 5" answers, summaries and more.
3. The user interacts with the newly tuned LLM via a chatbot interface, either text-based or via voice.

Key innovations will include:

1. **Scalable Oversight:** We aim to refine generated responses using a *factual consistency* model. This model will evaluate whether the answer is backed by a corpus of verified information sources like textbooks and scientific databases [2]. Non-factual responses are self-corrected using prompt engineering to force models to stick to the facts. Appendix A details prompts for (1) question generation, (2) answer generation, (3) answer scoring with and without ground truth answers. Appendix B shows the Evaluator application that will enable our work on large scale corpuses.
2. **Novel *factuality loss* in RL with AI Feedback.** We will introduce a novel training penalty, beyond cross entropy, termed *factuality loss*, a method of *retrieval-augmented RL with AI feedback*. Given a document, textbook or conference paper, we preprocess it by generating Q&A pairs (as in contribution 1 above). During training, retrieved documents are included in the prompt and the generated answers are evaluated by our factual consistency model that directly produces a final reward value. As we've done before, we will use TRLX for distributed RLHF/RLAIF training.
This additional factuality loss builds on the fundamental successes of Reinforcement Learning from

Human Feedback [8] (RLHF) and Anthropic's Constitutional AI [9], in a deep research effort to teach the model that hallucination is never acceptable.
3. **Supervise the reasoning *process* in addition to outcomes.** In addition to (1) we propose decomposing the question into PageRank style considerations such as "Is the author an expert in this field? Do they have prior work on this topic? ", etc. We assert that by forcing the model to think through the *process* of research, in an extended 'chain of thought' loop, it will be capable of both world class question answering and even the creation of new knowledge. The models will have access to tools such as relevant scientific databases, WolframAlpha and more [3-6], enabling to verify hypothesis by direct examination of existing data and even proposing new experiments to fill gaps in the existing literature.

The performance of any AI model leans heavily on the quality of the training data. Although we will use an existing general-purpose LLM to generate responses, integrating and grounding this model with high-quality background information is crucial to elicit top-notch responses. We will investigate both how to fine-tune such a model and how to prompt it with user-supplied context. One approach is to use Q&A pairs extracted from the initial background text to generate content for model training as well as for forming/enlarging the necessary knowledge base for model prompting. The ground truth Q&A pairs can be extracted by running SOTA models, such as GPT-4. We will build a database of questions to ask about a given user-supplied context necessary to probe various contextual details. For example, there will be questions related to understanding the spatial and temporal aspects of the contextual material (e.g., when and where the event is taking place), who are the main characters appearing in the context and what is their relation to the event, place, each other, etc. We will study what types of such Q&A pairs work best for what kind of context and user questions, how many of them are necessary, when is the best time to generate them (e.g., if these Q&A pairs are used for model fine-tuning, they need to be generated when a new "expert" is build, otherwise this can be done as a prompt engineering step at run-time), what additional computational cost is associated with both generating and using such Q&A pairs, etc.

Very frequently the initial context is not sufficient to become an "expert" and therefore finding and "reading" other relevant materials becomes necessary. Assuming the initial context comes with some sort of references to related/prior work, it may be relatively easy to extract these references, download them from the web and include as part of the initial training set. There are existing APIs and databases with citations that we can make us of. It is not clear however how deep our model should go in recursively expanding the citation tree. On the other hand, given a context without a citation list, it may be rather difficult to find relevant background information. Therefore, a significant part of the effort in this project will be dedicated to automating ways of building relevant supplemental knowledge base both in the presence of a reference list (e.g., as given in a scientific paper) and in its absence (e.g., when asking the chatbot to be come an expert on 20th Century Industrial Revolution). As an example, for the later scenario, we will need to map the key phrase "20th Century Industrial Revolution" into an initial list of reliable sources that can be further recursively expanded.

We will investigate several LLMs for building conversational AI "expert". On one side, smaller models can be fine-tuned and can execute rather fast on relatively low-end hardware, thus making our solution inexpensive to operate. However, the quality of the generated answers produced with such models may not be as good as with much larger LLMs. On the other hand, large LLMs are difficult to fine-tune and are more expensive to operate, however they can be efficiently used with appropriate prompt engineering, thus eliminating the need for fine-tuning in the first place. We will explore both approaches, ranging from models such as minimally sized LLaMA with just a few billion model parameters to GPT-3/4 scale models. We will investigate both fine-tuning and prompt engineering approaches. We will also investigate how model plugins can be used to integrate with 3rd party tools, such as databases, WolframAlpha, and others to interface with data and specialized tools.

Expected deliverables/outcomes

We expect to achieve a world class performance on factual document Q&A. Our main deliverable will be a prototype implementation that will operate as described in the Technical Description section. Along with this implementation, we will write a conference paper and will open-source the project. A graduate student will be mentored to work on the model development implementation as well as undergraduate students will be involved with the project through Independent Study classwork (e.g., ECE 397).

Milestones

Milestone 1: Implementing and Refining Scalable Oversight via Factual Consistency

Already completed during the application period, my last year of experimental engineering helped me find optimal prompts for Factual Consistency using documents, reported in Appendix A.

Milestone 2: Successful training and evaluation of our novel Factuality Loss in RL with AI Feedback

Successful training of a large open source LLM, such as Llama or Falcon-40B, using TRLX for RL with AI feedback from the Factual Consistency model. One technical challenge is using LLM inference during the training process, but we've done this before using Ray and PyTorch for distributed training and inference simultaneously on the Delta HPC cluster at UIUC.

The system will utilize a method of retrieval-augmented RL with AI feedback to ensure that the model avoids hallucination. This milestone concludes when the system can effectively implement and utilize the factuality loss feature, building on the successes of RLHF and Constitutional AI.

Milestone 3: Supervising the Reasoning Process

The final milestone focuses on the stretch goal of supervising the reasoning process in addition to outcomes. The system will decompose the question into PageRank style considerations to force the model to think through the research process. The models will have access to tools such as relevant scientific databases, WolframAlpha, and more, enabling them to verify hypotheses by examining existing data and even proposing new experiments. This milestone is concluded when the system is capable of world-class question answering and the creation of new knowledge through supervised reasoning.

Plans for open sourcing

We will open-source our implementation from the beginning.

Appendix A; Prompt Engineering

These prompts were developed in the creation of uiuc.chat and the RLHF training on top of Alpaca we conducted using hand-written completions from Electrical Engineering students at UIUC. These prompts cover QA generation and automated grading, even based on documents without ground-truth labels.

Answer Generation over Documents

```
"Please answer the following question. Use the context below, called 'official course materials,'
only if it's helpful and don't use parts that are very irrelevant. It's good to quote the official
course materials directly, something like 'from ABS source it says XYZ' Feel free to say you don't
know. \nHere's a few passages of high-quality official course materials:"

---
Document filename: ‘Day, Kastan - MS Thesis CP Comments 532023’, pagenumber: 29

Summary: Self-supervised transformers can easily utilize k-nearest neighbor search or cosine
similarity to find corresponding data samples in other modalities. These methods are most powerful
when combined into a multi-modal transformer.
---
Document filename: ‘Deep Learning with Python, Second Edition - Francois Chollet.pdf’, pagenumber:
573

Summary: Self-supervised transformers have self-attention, which focuses on relationships between
sequence elements regardless of their order in a sequence. The Transformer model is order-agnostic
but injects order information through positional encoding.
---

{additional retrieved contexts} +

Now please respond to my query: ' +

{search_query such as ‘What are some interesting attributes of self-supervised transformers?’}
```

Answer Generation over Documents with LaTeX Equations and Chain of Thought

We prepare the prompt based on type of question: factoid, causal, summary, elongate or listing.
We offer special support for equations in LaTeX, pass equation=True. To add chain-of-thought to any prompt pass cot=True.

```
def prepare_prompt(self, question: str, context: str, equation: bool=False, cot: bool=False)->str:

    factoid = ["What", "Where"]

    elongate = ["Detail", "Explain", "Discuss", "Expand", "Clarify", "Outline"]

    causal = ["Why", "How"]

    listing = ["List", "Break down"]

    summarize = ["Summarize", "Summarise", "Sum up"]

    first_word = question.split()[0]


    if first_word in factoid:

      prompt = """Generate an objective and logical answer to this question, based on the context. The answer should be short, to-the-point while being substantial as per a freshmen-level language.

            Do not include any irrelevant information. Give examples."""

    elif first_word in causal:

      prompt = """Generate a reasoning-based, precise answer to this question, based on the context.

            The answer should have a freshmen-level tone and be concise, logic-oriented.

            Give examples."""

    elif first_word in listing:

      prompt = """Generate a list-type answer to this question, based on the context.

            The answer should have a freshmen-level tone and be concise.

            It should contain reasons and examples. """

    elif first_word in elongate:

      prompt = """Generate a detailed, explanatory answer to this question, based on the context. The answer should have a freshmen-level language.  Give examples and talk about real-world applications of the concept.

            The answer should be long and discuss the concept."""

    elif first_word in summarize:

      prompt = """Summarize this context and answer the question. The answer should have a freshmen-level tone and be concise.

            Build an in-depth summary using examples."""

    else:

      prompt = """Generate a concise, short and to-the-point answer to this question, based on the context. The answer should have a freshmen level easy to understand language and tone. """
```

```
    if equation:
      prompt = prompt + "Add all necessary equations to explain this."
    if cot:
      prompt = "Generate a short answer to this question, based on the context. Use freshmen-level language.\n" + prompt + "\nLet's think step by step."
    return prompt + "\nContext" + context.replace("\n", " ") + "\nQuestion:" + question.replace("\n", " ") + "\nAnswer:"
```

Question Generation

Factual Document Chat prompt engineering. Prepares prompt based on type of question - factoid, causal, summary, elongate or listing.

We offer special support for equations in Latex, pass equation=True. To add chain-of-thought to any prompt pass cot=True.

You are a smart assistant designed to help high school teachers come up with reading comprehension questions.

Given a piece of text, you must come up with a question and answer pair that can be used to test a student's reading comprehension abilities.

When coming up with this question/answer pair, you must respond in the following format:

```
{{
    "question": "$YOUR_QUESTION_HERE",
    "answer": "$THE_ANSWER_HERE"
}}
```

Everything between the ``` must be valid json.

Question Generation (successful in Electrical Engineering domain)

[

> “Generate one multiple-choice type question about this context. The question should consist of reasoning and procedural steps including all the possibilities of the context.\nThe question should be precise and factual, and should address the real world curiosity of the topic.”,
>
> “Generate one objective, to-the-point and firm question about this context. The question should break down the context and specifically pin-point key terminologies and concepts.\nThe question should be precise, factual and truthful. The question should identify the important terms and accurately summarize them.”,
>
> “Generate one thoughtful, creative, steps-based procedural question about this context that starts with Why/How/Where/Who/When.\nThe question should be detailed, free-flowing, conversational and analytical. The question should spur curiosity and inspire discussions and debates.”

]

Factual Consistency Grading (Using Ground Truth)

Inputs: {question}, {generated answer}, {true answer}

You are a teacher grading a quiz.

You are given a question, the student's answer, and the true answer, and are asked to score the student answer as either Correct or Incorrect.

Example Format:

QUESTION: question here

STUDENT ANSWER: student's answer here

TRUE ANSWER: true answer here

GRADE: Correct or Incorrect here

Grade the student answers based ONLY on their factual accuracy. Ignore differences in punctuation and phrasing between the student answer and true answer. It is OK if the student answer contains more information than the true answer, as long as it does not contain any conflicting statements. If the student answers that there is no specific information provided in the context, then the answer is Incorrect. Begin!

QUESTION: {question}

STUDENT ANSWER: {generated answer}

TRUE ANSWER: {true answer}

GRADE:

Your response should be as follows:

GRADE: (Correct or Incorrect)

(line break)

EXPLANATION: (Without mentioning the student/teacher framing of this prompt, explain why the STUDENT ANSWER is Correct or Incorrect. Use one or two sentences maximum. Be as concise as possible.)

GPT-4 Optimized Grading Prompt (Using Ground Truth)

```
Inputs: {question}, {generated answer}, {true answer}
```

```
You are assessing a submitted student answer to a question relative to the true answer based
on the provided criteria:

    ***

    QUESTION: {question}

    ***

    STUDENT ANSWER: {generated answer}

    ***

    TRUE ANSWER: {true answer}

    ***

    Criteria:

      relevance:  Is the submission referring to a real quote from the text?"

      conciseness:  Is the answer concise and to the point?"

      correct: Is the answer correct?"

    ***

    Does the submission meet the criterion? First, write out in a step by step manner your
reasoning about the criterion to be sure that your conclusion is correct. Avoid simply
stating the correct answers at the outset. Then print "Correct" or "Incorrect" (without
quotes or punctuation) on its own line corresponding to the correct answer.

    Reasoning:
```

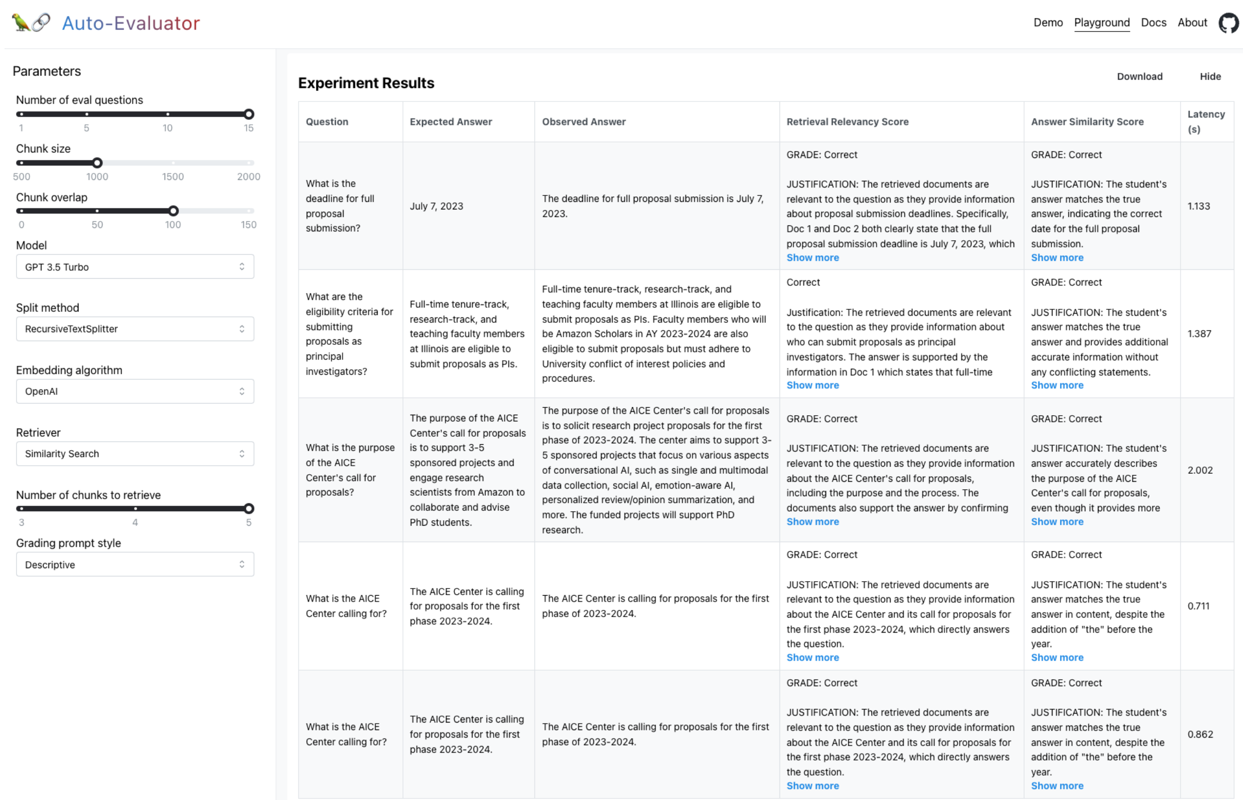
Factual Consistency Grading (Without Ground Truth)

Inputs: {question}, {generated answer}, {true answer}

```
Given the question:

    {query}

    Here are some documents retrieved in response to the question:

    {result}

    And here is the answer to the question:

    {answer}

    Criteria:

      relevance: Are the retrieved documents relevant to the question and do they support
the answer?"


    Your response should be as follows:


    GRADE: (Correct or Incorrect, depending if the retrieved documents meet the
criterion)

    (line break)

    EXPLANATION: (Write out in a step by step manner your reasoning about the criterion
to be sure that your conclusion is correct. Use one or two sentences maximum. Keep the
answer as concise as possible.)
```

Appendix B. Building on Auto-Evaluator

An example of generating a factual QA dataset [7]. This software is not necessary for our work, but it displays the key concept of QA evaluation.



Question	Expected Answer	Observed Answer	Retrieval Relevancy Score	Answer Similarity Score	Latency (s)
What is the deadline for full proposal submission?	July 7, 2023	The deadline for full proposal submission is July 7, 2023.	GRADE: Correct JUSTIFICATION: The retrieved documents are relevant to the question as they provide information about proposal submission deadlines. Specifically, Doc 1 and Doc 2 both clearly state that the full proposal submission deadline is July 7, 2023, which Show more	GRADE: Correct JUSTIFICATION: The student's answer matches the true answer, indicating the correct date for the full proposal submission. Show more	1.133
What are the eligibility criteria for submitting proposals as principal investigators?	Full-time tenure-track, research-track, and teaching faculty members at Illinois are eligible to submit proposals as PIs.	Full-time tenure-track, research-track, and teaching faculty members at Illinois are eligible to submit proposals as PIs. Faculty members who will be Amazon Scholars in AY 2023-2024 are also eligible to submit proposals but must adhere to University conflict of interest policies and procedures.	Correct Justification: The retrieved documents are relevant to the question as they provide information about who can submit proposals as principal investigators. The answer is supported by the information in Doc 1 which states that full-time Show more	GRADE: Correct JUSTIFICATION: The student's answer matches the true answer and provides additional accurate information without any conflicting statements. Show more	1.387
What is the purpose of the AICE Center's call for proposals?	The purpose of the AICE Center's call for proposals is to support 3-5 sponsored projects and engage research scientists from Amazon to collaborate and advise PhD students.	The purpose of the AICE Center's call for proposals is to solicit research project proposals for the first phase of 2023-2024. The center aims to support 3-5 sponsored projects that focus on various aspects of conversational AI, such as single and multimodal data collection, social AI, emotion-aware AI, personalized review/opinion summarization, and more. The funded projects will support PhD research.	GRADE: Correct JUSTIFICATION: The retrieved documents are relevant to the question as they provide information about the AICE Center's call for proposals, including the purpose and the process. The documents also support the answer by confirming Show more	GRADE: Correct JUSTIFICATION: The student's answer accurately describes the purpose of the AICE Center's call for proposals, even though it provides more Show more	2.002
What is the AICE Center calling for?	The AICE Center is calling for proposals for the first phase 2023-2024.	The AICE Center is calling for proposals for the first phase of 2023-2024.	GRADE: Correct JUSTIFICATION: The retrieved documents are relevant to the question as they provide information about the AICE Center and its call for proposals for the first phase 2023-2024, which directly answers the question. Show more	GRADE: Correct JUSTIFICATION: The student's answer matches the true answer in content, despite the addition of "the" before the year. Show more	0.711
What is the AICE Center calling for?	The AICE Center is calling for proposals for the first phase 2023-2024.	The AICE Center is calling for proposals for the first phase of 2023-2024.	GRADE: Correct JUSTIFICATION: The retrieved documents are relevant to the question as they provide information about the AICE Center and its call for proposals for the first phase 2023-2024, which directly answers the question. Show more	GRADE: Correct JUSTIFICATION: The student's answer matches the true answer in content, despite the addition of "the" before the year. Show more	0.862

University background IP

References

1. UIUC Course AI, https://www.uiuc.chat.
2. Liu et al. BRIO: Bringing Order to Abstractive Summarization, https://arxiv.org/abs/2203.16804
3. Asai et al. Task-aware Retrieval with Instructions, https://arxiv.org/abs/2211.09260
4. Schick et al. Toolformer: Language Models Can Teach Themselves to Use Tools, https://arxiv.org/abs/2302.04761
5. Nakano and Hilton et al. WebGPT: Browser-assisted question-answering with human feedback, https://arxiv.org/abs/2112.09332
6. Talmor et al. MultiModalQA: Complex Question Answering over Text, Tables and Images, https://arxiv.org/abs/2104.06039
7. Automated Document Question-Answering, https://github.com/langchain-ai/auto-evaluator
8. Ouyang and Wu et al. Training language models to follow instructions with human feedback, https://arxiv.org/abs/2203.02155
9. Anthropic Constitutional AI, https://www.anthropic.com/index/claudes-constitution

Budget

Item	**Amount**	**Basis of Estimate**
Research Programmer Support (Kastan Day)	$11,225	1 month salary, including fringe benefits
GRA Salary Support	$32,943	9-months GRA at 50% effort, including fringe benefits
Other Direct Costs:	$24,390	$2,000 Travel; $2,000 Publications; $1,000 Software;
		$19,198 Tuition Remission;
		$192 Institutional Support Fee (Computer Services)
Total	**$68,558**	