Using your own content in LLM's –

Retrieval Augmented Generation (RAG)

Gen-AI can help with education in many ways

Saves time

- AI lesson planning reduces the time teachers spend preparing lessons.
- Can automate marking using AI – image recognition or voice technologies to assess students as they read aloud.

Improved quality

- Adapt high quality resources to your context (languages or images).
- Can help teachers clarify concepts they may have forgotten.
- Adapt lessons/examples to children's previous answers.

Improved scalability

- Access on phones (e.g. chatbot to support teachers) enables wide adoption.
- Rapid integration of new national policies, curricula and best practices by automatically updating knowledge bases of LLMs.

But many of these require introducing your own information into the models.

The training data for major LLM's isn't specific to any one country and often doesn't include your education resources

If your content isn't in the training data it can only ever be approximated in the outputs.

So how you integrate local information – like your own teacher guides?

There are different techniques to use integrate your content and shape the outputs

1. Prompt Engineering
2. RAG systems
3. Fine-tuning
4. Rebuilding foundational models.

In this series, we will talk through different ways using core education materials.

Today we focus on Retrieval Augmented Generation (RAG)

1. Prompt Engineering
2. **RAG systems**
3. Fine-tuning
4. Rebuilding foundational models.

RAG can get LLM's in education from good to great

Improved Accuracy & Grounded in Curricula:

Teacher can access quality assured, standardised information (e.g. curricula, supporting materials).

Streamlined Access to Curricular Resources:

RAG can help educators quickly find relevant lesson plans, worksheets, and other materials that directly support the curriculum they teach. These materials can be updated quickly without expensive updates.

Flexibility or adaptability

The information can be curated differently depending on who is using (or what for).

So how the system uses the information on structured pedagogy can be adapted differently for school support officers or teachers.

What is Retrieval Augmented Generation (RAG)?

- RAG is an extension of Large Language Models (LLMs)
  - LLMs are things like GPT4, Gemini, Claude, LLaMA, Mistral, etc.

- So, to understand what RAG is, it first helps to recap what an LLM is (and what it isn't)

What an LLM is

- An LLM is a text completion machine – it predicts the next best words.
  - Given a partial string of text, it will produce the best string of text that should follow it.
  - "Best" here means 2 things:
    1. *The most likely*
    2. *The most valuable*

What do we mean by most likely?

- The “most likely” completion is learned by experiencing examples of real text.
  - During “pre-training”, the LLM is exposed to hundreds of millions of examples of real text, which it attempts to complete (like fill in the _GAP_)
  - When it is wrong, it is penalised, and its internal state is updated to reflect that
  - This continues until it is really good at predicting real text

What do we mean by valuable?

- The “most valuable” completion is learned by testing it with real people (human feedback).
  - Pre-training teaches the LLM how to produce real text
  - But there is no explicit concept in this of the **value** of any text completion
    - Inaccurate, or irrelevant, or rude text will just as likely appear in the training data as the opposite
  - In the second phase of training, the LLM is exposed to the reactions of real people to what it is producing.
  - Similar to pre-training, when the reactions are negative, it is penalised, and its internal state is updated to reflect that
  - This continues until it is good enough at producing completions that people value.
  - Technically this is called Reinforcement learning from human feedback

What an LLM isn't

- It is not a lifelong learner
  - Once training is finished, the internal state of the LLM never changes again
  - When using something like ChatGPT, it may **appear** to learn, in the sense that it can recall earlier instructions or reference earlier parts of the conversation
  - That is because each output it produces is the result of completing the text of the full conversation up to that point (ChatGPT is not an LLM, it is a system that operates around an LLM)
- It is not infinite
  - The "window" of text that an LLM can use for completion is finite
  - It can be very long (up to a hundred pages or more these days), but there is a limit
- It is not a search engine
  - An LLM can't search for new information on the internet (though it can be linked to one, see above point about ChatGPT)

So what is RAG?

- RAG systems are specialised tools designed for LLMs to facilitate the process of gathering and using information.
  - We save our resources in a database (vector DB)
  - We **R**etrieve what's relevant
  - We use this information to **A**ugment the **G**eneration of the answer

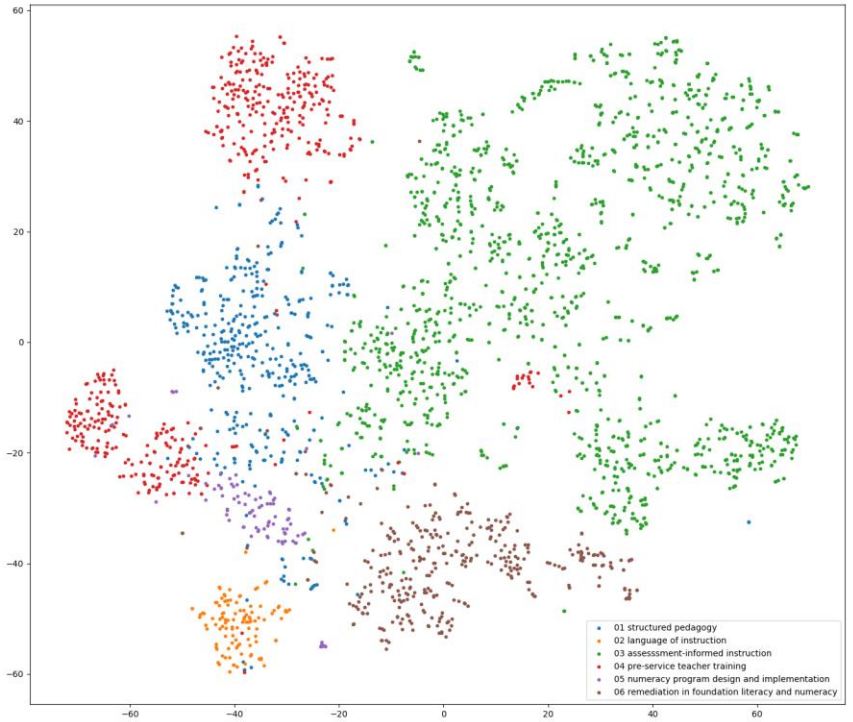
- To start, we need to represent the content numerically as **embeddings**

Embedding our text

- Embeddings are a way of turning words or sections of text into a numerical code (list of numbers).
- These numerical codes capture the semantics of what the text describes: the relationships between the codes reflect the relationships between the words they represent.
- Embeddings come from the statistical structure of language: words that hang out together, or in groups are likely to have similar meanings, or to describe things in the world that often occur together.
- Embeddings provide a computer-friendly representation of language that preserves much of the meaning.
- This allows computer algorithms to work with language: for example, finding which chunk of a document has the closest embedding to a query prompt gives the chunk with the most relevant information.

What does this mean in reality? Using the Science of Teaching Resources

- Each point corresponds to a chunk of text from a document from the https://scienceofteaching.site/ materials.
- The different colours represent different topic areas.
- The position of each point corresponds to its numerical code (reduced to 2 numbers here for visualisation)
- Chunks that are closer together are more similar in meaning.



There are many considerations for a RAG model

Split into the two key components: Retrieval and Generation

Retrieval: Finding relevant information

- It's like going to a huge library and finding the most relevant books to answer the question

Generation: How do we use the retrieved information for response

- Like an expert scholar summarises the information in the books you have picked out

A step by step example of how RAG works in a chatbot

- **The user asks a question** – "Hi, can you help me with a lesson plan to explain gravity to year 5"
  - *The system needs to generate a query to match against the content in the embeddings*

- **To do this** it uses the LLM to generate text that conveys the meaning of what we want
  - this can be done through being designed to chat and get the information.
  - or by inferring what is needed and picking between its tools.

- **The query** is then converted to embedding vector – and the model finds the nearest points in the database to this vector (nearest "n" where n can change).

- **These points – the content** – are then sent back for Generation (or just for reference).
- *So here it can return any lesson plan structures (if they are standardised and saved) and any useful text on gravity.*

- **This text is then appended to the Prompt**
  - *And the LLM processes the answer.*

Next up – a move towards compound AI systems

- RAG models are just one example of how by combining components, AI-Edtech tools can get increasingly powerful.
- By paying careful attention to the design of the system, we can improve the quality and accuracy of AI.
- Many education tasks – lesson planning; test development; marking – maybe better suited to multistep designs.
- This is especially important in education where control and trust is needed.

Stayed tuned for more AI guidance, knowledge and tools at https://ai-for-education.org