Imperial College London

BEng Individual Project

Imperial College London

Department of Computing

Ontological Infrastructure Design for Benchmarking Federated LLM Tools in British Schools

Supervisor:
Dr. Thomas Lancaster

Author:
Hashir Majeed

Second Marker:
Dr. Matt Collison

June 11, 2025

Abstract

A 2024 Department for Education (DfE) report found that 42% of UK teachers had already adopted generative AI in their practice, yet 68% expressed concerns about reliability and bias in outputs, and 35% felt existing tools lacked adequate student support capabilities [1]. This project addresses these concerns by proposing an ontology-driven infrastructure for benchmarking and deploying federated Large Language Model (LLM) tools within British school institutions.

By leveraging Palantir's Ontology Software Development Kit (OSDK), an ontology was constructed to mirror the semantic nouns and operations present in schools. This ontology underpinned a Retrieval-Augmented Generation (RAG) architecture designed to enhance LLM outputs with context that brings together siloed data sources, integrating data from inspection reports, residual school systems and ad-hoc teacher inputs. The system applies granular permissions and access controls, reflecting real-world permissioning hierarchies as components of wider research into federated LLM tools that ensure individual schools can remain as data controllers. Multiple retrieval strategies such as vector similarity, HyDE (Hypothetical Document Embeddings), and Reciprocal Rank Fusion (RRF) were implemented and benchmarked, achieving a BERTScore F1 of 86.66% compared to expert teacher-generated responses. Sentiment and bias evaluation revealed a 14.23 percentage point increase in positive sentiment (99.42% vs. 85.19%) and matching decrease in negative sentiment (0.58% vs. 14.81%) when compared to expert teacher writing, with the system's bias statistically tested and found to be practically sufficiently close to expert teacher response levels.

A proof-of-concept tool was deployed in a real school setting, and qualitative feedback from semi-structured interviews demonstrated improved trust, usability, and pedagogical alignment. This report offers a blueprint for scalable, privacy-preserving AI solutions in British schools, empowering teachers to provide contextual and tangible support plans for students.

Acknowledgments

Many thanks to Dr. Thomas Lancaster for frequently taking time to provide great wisdom during the course of the project. Also many thanks to Dr. Matt Collison for taking the time to do a deep-dive on my project at the halfway mark. Finally, many thanks to Palantir Technologies for facilitating the usage of AIP and OSDK for this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

Since the release of ChatGPT in 2022, Large Language Models (LLMs) have gained attention for their remarkable ability to generate semantic responses to text-based prompts in a human-like manner - with ChatGPT having nearly 800 million active users as of June 2025 [8]. The applications of such technology span a variety of use cases, such as administrative automation, legal analysis, and educational technology. That being said, more existential industries such as healthcare or education that involve high-stakes decision-making mandate a high level of accuracy to provide accurate information in such environments. LLMs often produce plausible-sounding but incorrect information, a phenomenon known as "hallucination" [9]. A study by Chen et al. observed hallucination rates as high as 30% when LLMs were applied without contextual constraints [10]. This unreliability poses a risk where inaccurate content can mislead students or undermine teachers' confidence in AI tools [11]. Such large language models require access to contextual, domain-specific information to support reliable operational decisions [12].

Moreover, the 2023 Department for Education (DfE) report on *Technology in Schools Survey Report: 2022 to 2023* emphasises the importance of evidence-based technological solutions that are tailored to schools [13]. In British schools, this can vary from different curricula, exam boards, as well as internal school administration systems and data storage of the pupils. The latter is often stored in siloed data sources dependent on the individual school [14]. The report highlights that 42% of primary and secondary school teachers in the UK used generative AI tools by late 2023, a significant increase from 17% earlier that year [15]. Despite this growth, 68% of surveyed teachers expressed concerns about the accuracy and reliability of such tools in classroom settings [15]. Additionally, the report found that schools in the UK spend approximately £900 million annually on educational technology (EdTech) tools, yet 35% of teachers report that these tools fail to provide enough support to their students [15]. These statistics underline the pressing need for a solution that incorporates generative AI in a way that is both scalable and has sufficient context to provide accurate decision-making capabilities.

To address these issues, there is a growing recognition of the need for systems that combine the generative power of LLMs that are backed by the context of the operations of a school. One solution to this problem is through the use of ontologies - a semantic-based structure that facilitate a digital representation of a system on a semantic level. Ontologies can impose strict data standards that can be retrieved and used accurately by LLMs [16]. By constraining LLM outputs within the boundaries of a well-defined ontological framework, it is possible for LLMs to leverage the context of the ontology to provide reliable outputs whilst reducing the risk of hallucinations [17, 10]. This would make them suitable for usage in the educational context.

1.2 Objectives & Contributions

In particular, this research sets out to accomplish the following goals:

- Develop a semantic ontology-based infrastructure that aligns with the semantics of operations pertinent to teachers and students in a classroom ecosystem for UK schools.

- Leverage the unified ontological infrastructure to power selected off-the-shelf LLM models for a proof-of-concept operational tool.
- Benchmark the tool according to key metrics:
  - **Accuracy:** How effective is the ontology instance at facilitating accurate retrieval?
  - **Pedagogy:** To what extent does the LLM output advice similar to what an expert teacher would deem suitable, given enough time?
  - **Sentiment:** Does the advice supplied have a constructive tone for the user?
  - **Bias:** Does the LLM demonstrate any bias towards certain demographic groups present in UK schools?
- Deploy the tool within a real school to investigate pedagogical effectiveness in integration with current teacher workflows.
- Investigate the practicality of implementing privacy-preserving granular access controls for a hypothetical federated deployment on a national scale.

This project aims to bridge the contextual gap needed for LLMs to generate value in a risk-free manner in a school environment. By aligning AI capabilities with the structured needs of an ontological representation of a British school system, this project seeks to contribute to the development of scalable EdTech solutions that can enhance student education outcomes.

In order to create an ontological backing infrastructure for operational tools to build off of, the Ontology Software Development Kit (OSDK), a product part of the Palantir Foundry ecosystem, is leveraged. This allows the synthesis of an ontology-based system with customised software solutions that are tailored towards an educational setting in a classroom. From this, rigorous benchmarking for the outlined metrics such as accuracy, sentiment and bias can take place to assess the viability of the proposed hypothesis.

This research's contributions are primarily providing a blueprint for scalable, privacy-preserving AI solutions in British schools, empowering teachers to provide contextual and tangible support plans for students. Within this, it contributes state-of-the-art benchmarking of LLMs backed by an ontology of such a system when applied in an educational context.

Chapter 2

Background

2.1 Existing EdTech Tools

The problem of cracking how to effectively leverage technology to make generating valuable operational insights simpler, easier, and more effective for teachers has stood for over a decade for educators all around the world. This area has been increasingly explored since the Covid-19 pandemic, in which the need for effective EdTech tools became more apparent, with rising global demand for end-to-end solutions [18, 19]. The global EdTech market has experienced unprecedented growth, expanding from USD 146.0 billion in 2023 to an expected USD 549.6 billion by 2033, representing a compound annual growth rate of 14.2% [20]. This growth reflects the urgent need for evidence-based educational technology solutions - across the UK, schools currently trial an average of 2,591 EdTech tools annually without robust outcomes data to guide their selection [21]. An analysis of these current tools can help identify key failing points of current EdTech solutions.

2.1.1 Eedi

Eedi is an EdTech platform tailored specifically to global mathematics education, serving over 160,000 teachers across 19,000 schools worldwide [22]. The platform has gained significant traction in the EdTech space, with its diagnostic questions being answered over 39 million times across its database of over 60,000 questions [23]. In a comprehensive randomised controlled trial conducted by the Education Endowment Foundation, the tool was evaluated across 180 schools in the UK, aimed at GCSE students, with the collated results showing positive trends, as discussed later [24]. At the core of Eedi is a probabilistic computation tool for a type of testing called Item Response Theory (IRT). IRT is used as a common method to gauge performance [25].

The crux of this method of testing is therefore to what extent the calculated distribution fits true performance. This will have a knock-on effect in how appropriate the subsequent questions asked to students are, which directly affects the quality of the reports generated for the educator. Research indicates that effective implementation requires consistent usage, as the EEF trial found that students received an average of only 25 quizzes in Year 10 and four quizzes in Year 11, compared with expected figures of approximately 80 quizzes in Year 10 and 52 in Year 11 [24].

There are several advantages to the probabilistic adaptive questions via IRT utilised by Eedi. The platform's comprehensive approach to identifying mathematical misconceptions through AI-powered analysis has shown promise in educational settings [23]. In terms of relevance to leveraging data about students to allow teachers to customise learning paths, Eedi's benefits are twofold:

- Teachers are able to recommend accurate follow-up questions based on a mathematical model - this is more rigorous than teachers trying to track and customise learning based on intuition, which may be flawed as it is hard to keep track of many students.
- Teachers are able to view comprehensive reports of their students, in particular which topics they are stronger or weaker in. This further aids the educator's ability to identify accurate actionable steps for improvement.

That being said, there are a few shortcomings that detract from Eedi's viability as an EdTech tool to be widely implemented. These stem from the fact that machine learning and adaptive algorithms by nature necessitate a high degree of usage before they become useful [24]. This is often not viable in the UK curriculum, where students typically study around 10 subjects at GCSE level, each with their own homework and examinations [26]. Research indicates that British secondary school students spend an average of 1.5 to 2 hours per night on homework across all subjects, with Year 10-11 students facing up to 90 minutes of homework per night in addition to examination preparation [27, 28]. This yields another problem too - this data that lies in homework, problem sheets, and exams is left siloed outside of the Eedi system. As a result, a teacher would have to manually integrate the different sources of information with the Eedi analytics for each student to create a truly comprehensive student plan. In most cases, this would take too much time.

2.1.2 Century Tech

An example of an AI tool that aims to empower educators to best deliver teaching in UK schools is Century Tech. Century Tech offers a learning platform for several aspects of the UK curriculum, in both primary and secondary schools [29]. The platform has demonstrated significant reach, being used by over 10,000 students across multiple secondary schools and colleges, and has expanded internationally to partner with over 100 British international schools globally [30, 31]. The platform incorporates the learning of pupils into a single data form, which can be visualised in a variety of tables and graphs, as shown below [32].

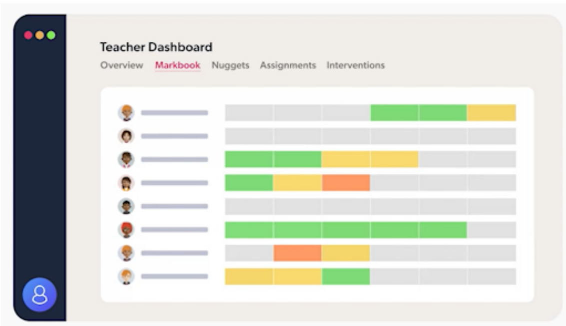

Figure 2.1: A dashboard showing a holistic view of all students in the Century Tech platform [2]

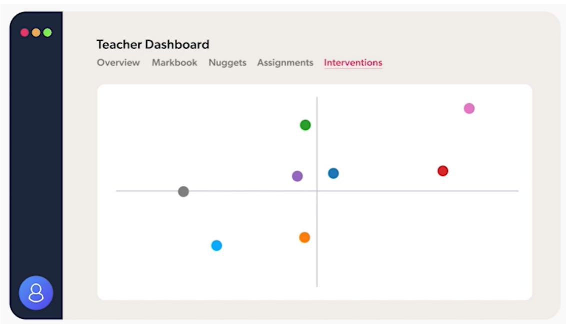

Figure 2.2: A visualisation showing a students' strengths and weaknesses [2]

One particular appeal of the Century Tech tool is how it is able to provide valuable insights by integrating lots of different data sources as parameters. This creates a more complete view of a student by considering metrics beyond the correctness of the answer - how much time did the student take? Had they seen a similar question before? What was their current strength in this topic? Did they use any hints? This is valuable data that would normally be abstracted away from a normal teacher marking homework exercises, resulting in a more accurate view of the current student's ability.

The Century Tech approach of providing static visual analytics has several drawbacks, namely:

- There is no operational capability from a static visualisation, and so an educator would still have to go through a lot of effort of understanding the data and then coming up with a personalised plan.
- A teacher has no way to query the underlying data in the visualisations. The metrics used to make the visualisations are abstracted away, and so this means a teacher cannot drill deeper into areas that might help them inform their actionable plan.
- As with Eedi, Century Tech does not integrate data from other sources about the student, such as exam results or homeworks. The result is that a platform like this requires a great deal of buy-in and usage by both students and teachers for it to be truly valuable.

This motivates the desire for a NLP technological integrationw within these data tools. LLMs serve as a great tool for dynamic querying of data. It is therefore reasonable to hypothesise that a system like Century Tech could benefit from an LLM feature. This would alleviate the second drawback above, allowing a teacher to drill deeper in natural language to find out more about a student's profile. This may even be extended to an LLM suggesting an operational plan to further relieve the educator of both time and effort.

2.2 How LLMs Work

In order to understand the viability of LLMs in EdTech tools, it is worth taking a look under the hood at how an LLM generates text based on a prompt. This is made more pertinent to understanding the root causes of hallucinations in LLMs and how they can be combated in critical workflows such as EdTech tools. LLMs are built on transformers - these are deep learning models that process data in a sequential manner. This is possible due to the self-attention nature of these models that allow them to have contextual understanding of the sentence and wider text to accurately make coherent sentences.

2.2.1 Tokenisation

The process of tokenisation involves using algorithms like Byte Pair Encoding (BPE) to break a large text into individual pieces called tokens. This allows the free text to be organised into a usable data structure where each token has a unique associated integer.

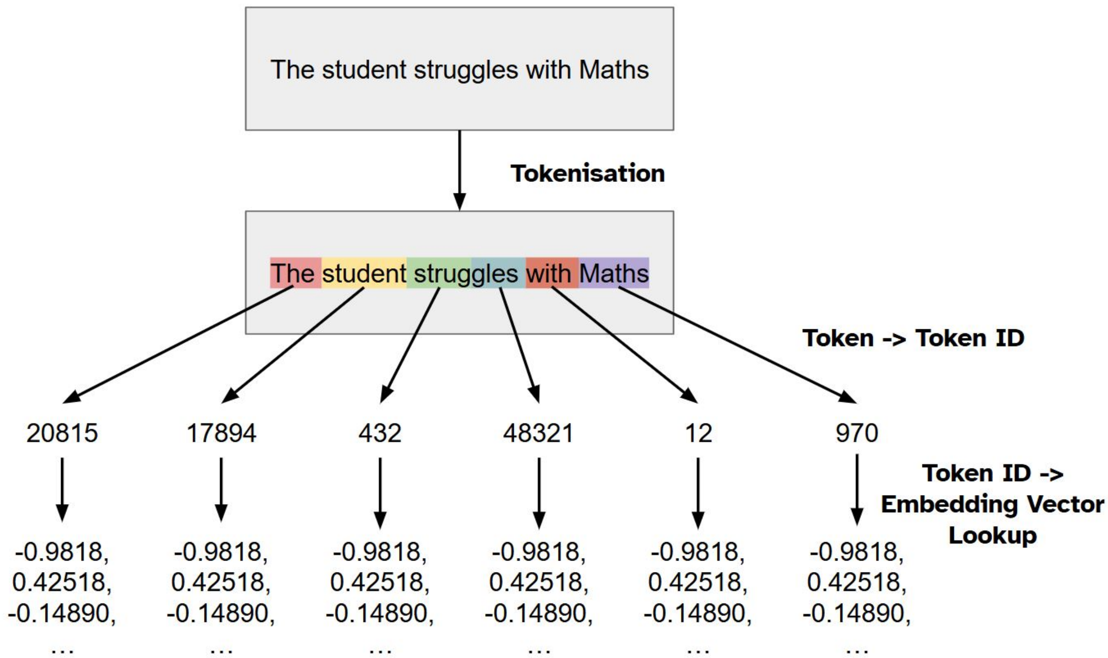

Figure 2.3: An example of tokenisation of sample text by a tokeniser model

2.2.2 Training

LLMs are trained on huge datasets - typically many terabytes large - to learn patterns of textual sequences in the English language. As per conventional deep learning models, this model is trained by attempting to minimise a loss function between the predicted sequence of tokens against the true values.

This highlights a key potential problem in LLMs - they are simply neural networks to predict the next item in the sequence of values that represent tokens [16]. Whilst this can often seem like it is not an issue (typically due to the large amount of training data trained on), this can result in mistakes in the prediction. As one would expect, this is more prevalent in areas where less training data is available.

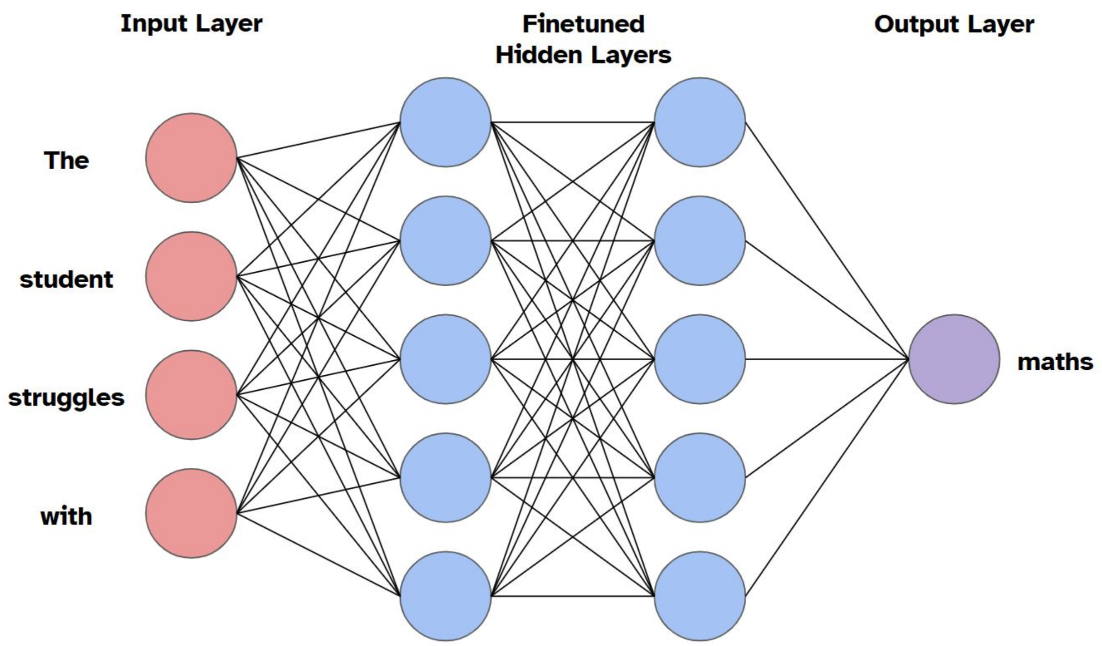

Figure 2.4: A predictive neural network

2.2.3 Hallucination in LLMs

Erroneous outputs from a trained LLM model are known as "hallucinations". This is the most existential problem facing the usage of LLMs in EdTech solutions. The two most common causes for hallucinations are as follows [13]:

- **Noisy training data** - with such a large set of data needed for training purposes, it is inevitable that it will contain some misinformation. This is reflected in the trained model, resulting in inaccurate hallucinations for certain inputs.
- **A lack of domain-specific context** - In niche areas where little training data is available, an LLM will struggle to provide factual responses, leading to hallucinations.

The hallucination of LLMs is an existential issue to edTech tools - this is evident through study of previous tools that have attempted to leverage LLMs in EdTech solutions for teachers [33].

2.2.4 Benchmarking LLM Performance

Effectively evaluating the value of the output of the tool will greatly depend on the evaluation of the output of the LLM models used. LLM benchmarking is not a purely objective process, since models can be benchmarked on a number of factors - for example, whilst some models may excel in areas such as accuracy, they may lack in areas such as efficiency or scalability. In the context of an educational workflow for British schools, the following metrics are most existential to the functionality of the application:

- **Natural Language Processing Capabilities** - This is at the heart of the main functionality of an LLM. As discussed, the main appeal of using LLMs is the ability for a teacher to query and drill deeper into points of interest. The effectiveness of this will be determined by the LLM's ability to effectively internalise and process natural language prompts to provide the most applicable answer. Benchmarks like GLUE [34] and SuperGLUE [35] are used to assess this on the basis of question answering for many text-based prompts and conversations.

- **Biased Outputs** - This is arguably the most existential problem surrounding the use of LLMs in the educational context. Due to biases and trends in the training data, LLMs have been known to exhibit biases [36, 37]. These biases can be exploited in areas involving data such as race, colour, ethnicity, and nationality. This would inaccurately influence the output of an educational LLM tool and render it unreliable and untrustworthy. In order to benchmark the model for bias, a common methodology is to use the BLOOM bias benchmarks [38]. This is a standardised framework to test against fixed datasets for biases in the LLM output. Testing has highlighted that GPT-3 and GPT-4 exhibit reduced but non-negligible biases, while BLOOM and LLaMA have been shown to exhibit more pronounced demographic biases, making them less suitable for an edTech tool involving student data.

- **Emotional Intelligence and Sentiment** - Even if an LLM can accurately identify and suggest pathways for students to follow, it is important for these to be delivered in a tone that is not derogatory and remains motivational for an educational context. This can be benchmarked using sentiment analysis of the output [39, 40].

2.3 LLM EdTech Tools

In order to truly understand how detrimental hallucinations can be to critical educational workflows, it is worth considering the case of ReportGenie. A study in *The Past, Present and Future Use of Artificial Intelligence in Teacher Education* analysed the usage of ReportGenie by teachers [3].

ReportGenie takes in several multiple-choice parameters about the student from the teacher. These include a discrete scale of values - Poor, Okay, Good, Excellent. In addition, the teacher can input free text which will also be incorporated in the report. The report then leverages LLMs to take the inputs and construct a prompt to the LLM to output a report that summarises the student's performance. The tool claims to increase the speed of report writing by 10 times.

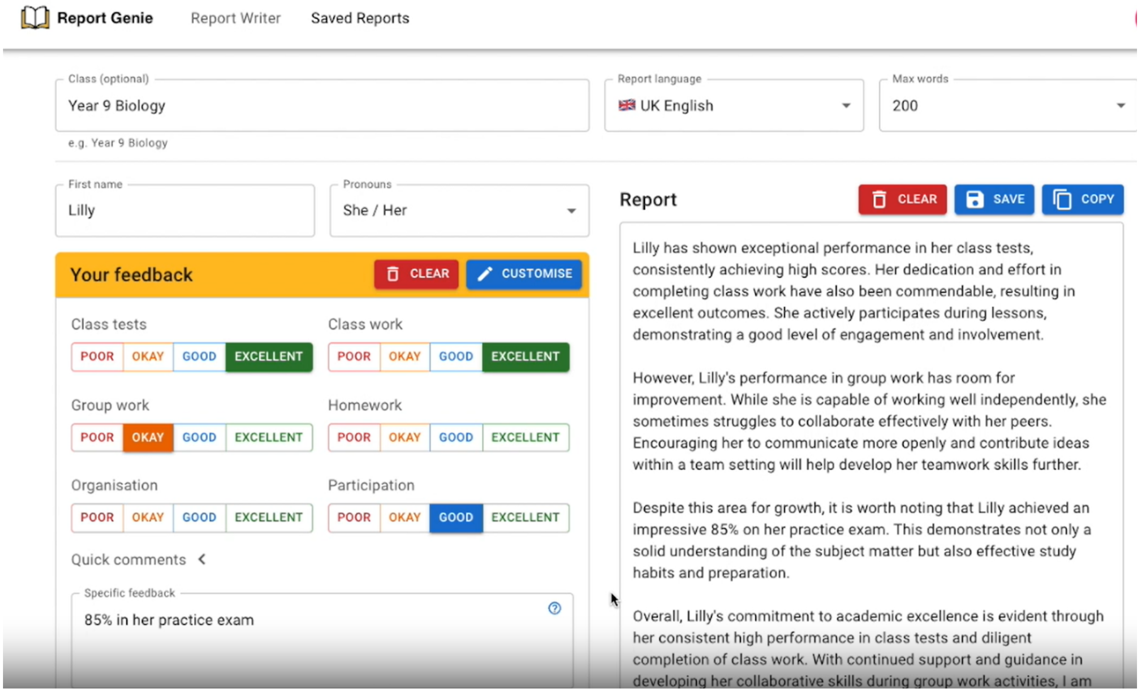

Figure 2.5: ReportGenie generating a report card based on discrete and qualitative parameters [3]

ReportGenie empowers teachers to quickly iterate on a report card, through the quick drafting and redrafting of report cards from small changes in the prompts. Studies in similar tools [33] have uncovered several issues preventing the wide-scale adoption of LLM tools. These include:

- There is a lack of end-to-end solutions for educators. Solutions like ReportGenie allow great leverage of LLMs in report writing, but lack querying capabilities for a teacher to gain extra insight into a student before writing their feedback. The vice versa case tells a similar story.
- A study by Zhou et al., showed that inaccuracies are prevalent in outputs of such reports - this is due to the hallucinogenic nature of using LLMs as previously discussed [33].
- A survey by the National Education Union in 2022 found that 68% of UK teachers expressed skepticism about the reliability of AI-driven EdTech tools, citing concerns over accuracy and bias [1].

The result is the need for an end-to-end solution that can integrate existing silos of data, and leverage context-aware LLMs that won't hallucinate - all in a tool that expresses AI in a way that teachers can trust to be both accurate and unbiased. The three problems to tackle here are:

- Protecting privacy and civil liberties in the case of data governance.
- Providing LLMs with context to prevent biases or hallucinations.
- Integrating existing siloed data sources into the solution.

2.3.1 LLM Data Governance

A survey by the UK Department for Education in 2023 found that 74% of schools consider data privacy a top priority when adopting new technologies [13]. Integrating governance mechanisms into LLM-based tools are therefore critical to enabling the large-scale trusted adoption of LLM solutions.

The two main legal regulations that such a tool would have to comply with in the UK education system are the GDPR regulations and the Data Protection Act. In this case, the school would act as the Data Controller, and would need to be able to bear responsibility for both secure data storage and processing. This greatly complicates the notion of having central processing of an

LLM tool in the education context. In addition, targeted domain-specific training would need to preserve Intellectual Property (IP) and licensing regulations. Furthermore, in order to maintain trust amongst the general public of educational professionals, the processing of data will need to be transparent. Transparency should be complemented with equally transparent data lineage models that can trace the transformation of sensitive student data through all stages of processing [41]. One such solution to tackle these problems is incorporating federated learning in the usage of the LLMs in the wider solution [4].

2.3.2 Federated Learning

Federated Learning (FL) is an emerging machine learning paradigm that addresses data privacy concerns by enabling model training across decentralized data sources without transferring raw data to a central server. This would enable local schools to be the data controllers, controlling sensitive student data on their own onsite or private cloud servers [4]. As an added benefit, this reduces transportation of data, reducing the chance of data interception and implicitly improving security.

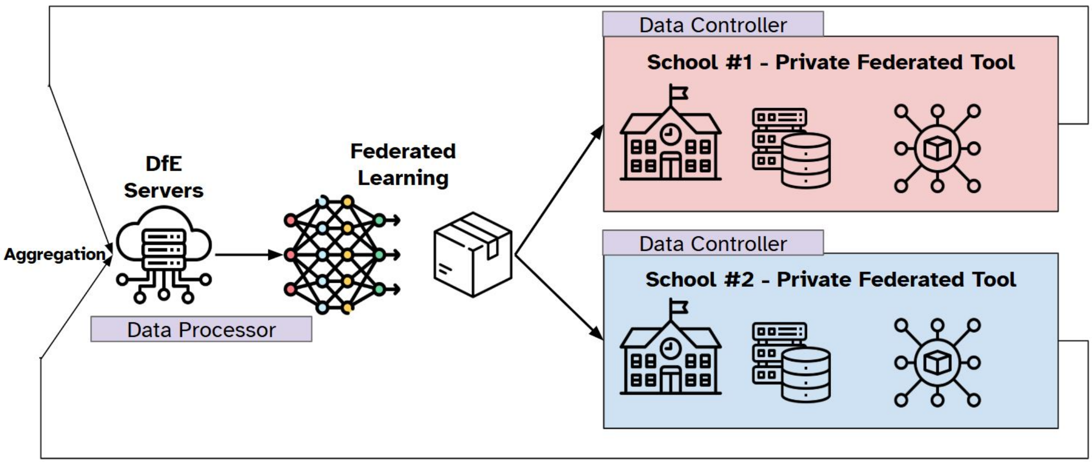

Figure 2.6: A federated learning architecture for schools

In the context of educational LLM workflows, federated learning can be leveraged by first training a National level model on the domain of broad and overarching educational concepts. This model can then be distributed to individual tools where less intensive fine-tuning adjustments can optionally be made based on the school's individualistic data and workflows. Global aggregation enables the encrypted collective improvement of the model in a way that preserves the privacy and civil liberties of students and schools alike.

2.3.3 Case study: The NHS Federated Data Platform

The Federated Data Platform (FDP) was a £330 million contract awarded in November 2023 to a consortium of commercial companies - headed by Palantir Technologies - to empower trusts to make better use of the patient data and workflows they legally control [4]. This system enabled NHS England to maintain its role as the data controller due to the federated nature of the platform, whilst rolling out cutting-edge solutions on a more granular level to Integrated Care Boards [ICBs], trusts and hospitals.

The results of such a platform have been extremely evident. At Chelsea and Westminster hospital, inpatient waiting lists were able to leverage the federated technology to reduce inpatient waiting time by 28% [4]. Additionally, across the UK operating theatre usage increased by 6.3% as a result of scheduling optimisations.

The NHS FDP serves as a valuable proof of concept - a testament to the effectiveness of federated learning solutions to be deployed on a national scale whilst keeping privacy and civil

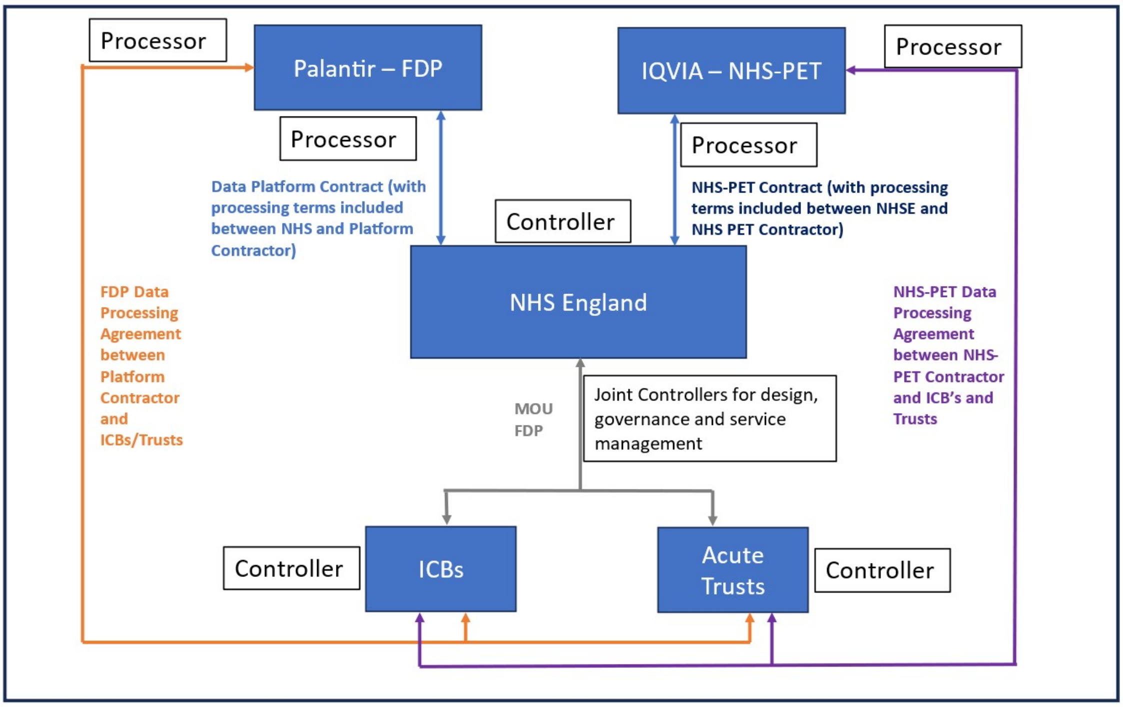

Figure 2.7: Architecture of the NHS Federated Data Platform [4]

liberties protection at the forefront. This provides extra backing to the concept of using federated learning as a method for deploying cutting edge technology whilst abiding by GDPR and Data Protection acts in the educational space [4].

2.4 Motivating the Need for Context-Aware LLMs

A key problem earlier identified is the need to combat hallucinations in LLMs. As outlined several times, existing edTech solutions consistently fail to incorporate these contextual elements, leading to vague insights that are tricky for teachers to convert into actionable plans. For example, a diagnostic tool might flag a student as underperforming without accounting for factors like recent illnesses or gaps in foundational knowledge.

A paper by Roll and Wylie (2016) highlights that adaptive learning systems incorporating contextual factors improved learning outcomes by 25-30% compared to systems that did not. Another study by Luckin et al. (2018) found that contextualised feedback increased student engagement by 40%, particularly in subjects like mathematics and science. These benefits show that context in edTech tools can be transformative for both educators and their students.

2.4.1 How LLMs Can Use Context

In order to give LLMs the necessary context to eliminate hallucinations, there are a few different techniques that can be used. One example of such a method is context-aware retrieval augmented generation (RAG) architectures. These leverage the self-attention mechanisms previously highlighted in LLMs complemented with a knowledge base of the previous data, prompts, and responses given in the application.

As illustrated above, a RAG operates in 5 key stages:

1. **Prompt & Query:** the educator would be able to query the application or data to drill deeper into a student profile.
2. **Query searched in knowledge base:** the knowledge base of previous data and prompts can be dynamically searched for information that would help improve the prompt by supplying it with any additional necessary context.

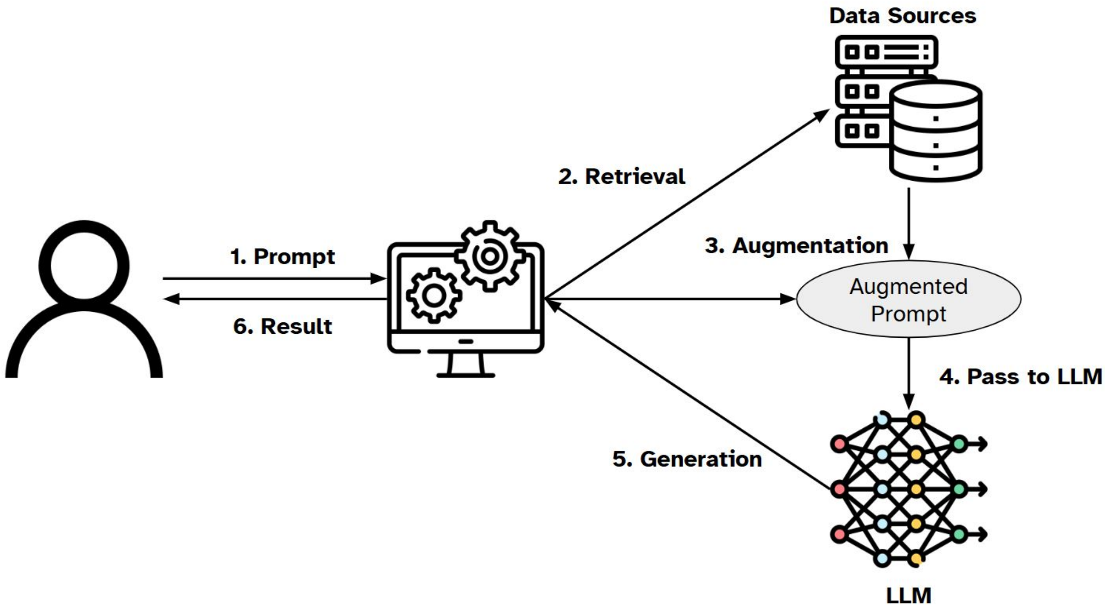

Figure 2.8: Process of RAG in LLMs

3. **Retrieval and Augmentation:** the retrieved information from the knowledge base can be used to augment the original prompt into a new and improved prompt.

4. **Updated query sent to LLM:** the generated prompt can now be sent to the trained model.

5. **RAG-enhanced response:** the generated response can be returned to the user, with a significantly lower chance of hallucination due to the valuable context in the prompt.

The feasibility of a RAG architecture lives and dies with its ability to effectively query the structure of its knowledge base. In order to construct a knowledge base for an edTech tool, the knowledge base will need to be able to parse existing data from educators on their students and their associated academic records. Before determining if a certain knowledge base architecture is feasible, it would be valuable to understand how the relevant data is currently stored in school systems. This will help the application integrate existing data, mitigating the issue of edTech tools requiring significant additional buy-in and usage from teachers and students to gather sufficient context and information about a student before being truly valuable.

2.4.2 Siloed Data Sources

The most existential problem to comprehensive data storage in schools is siloed data sources. Data about students is often stored in a range of administrative systems - which are not universal to all teachers in UK schools - as well as paper sources, spreadsheets, and other data silos. As discussed, this makes it tricky for both a teacher and an AI tool to integrate all the necessary information needed to make a well-informed decision into one place [42].

There are several drawbacks to siloed data sources, particularly hard copies that exist on paper. They are time-consuming to update, prone to human error, and create barriers to collaboration [43]. The challenge of transferring data from paper or spreadsheets to digital systems exacerbates the silo problem. Furthermore, because these tools don't retain contextual nuances, they are ill-suited to capturing the complexities of student needs and development over time.

2.4.3 KIM Case Study

One example of an administrative tool used to store student information in British schools is KIM - the Key Information Manager. KIM is designed to centralise student information, such as attendance, academic progress, and behavioral reports, allowing for more efficient management and better data access for school staff. In principle, such a system has been explored to be beneficial - teachers can access a student's attendance, academic performance, and behavioral history from one interface, making it easier to make data-driven decisions [44]. KIM is intended to act as the single source of truth for human information, but falls short for several reasons.

Fundamentally, the problem boils down to one word - interoperability - or rather a lack thereof. Platforms like KIM disregard the fact that different teachers operate differently. Two teachers may have different teaching styles; different metrics of recording progress; different assessment styles, and the list can go on.

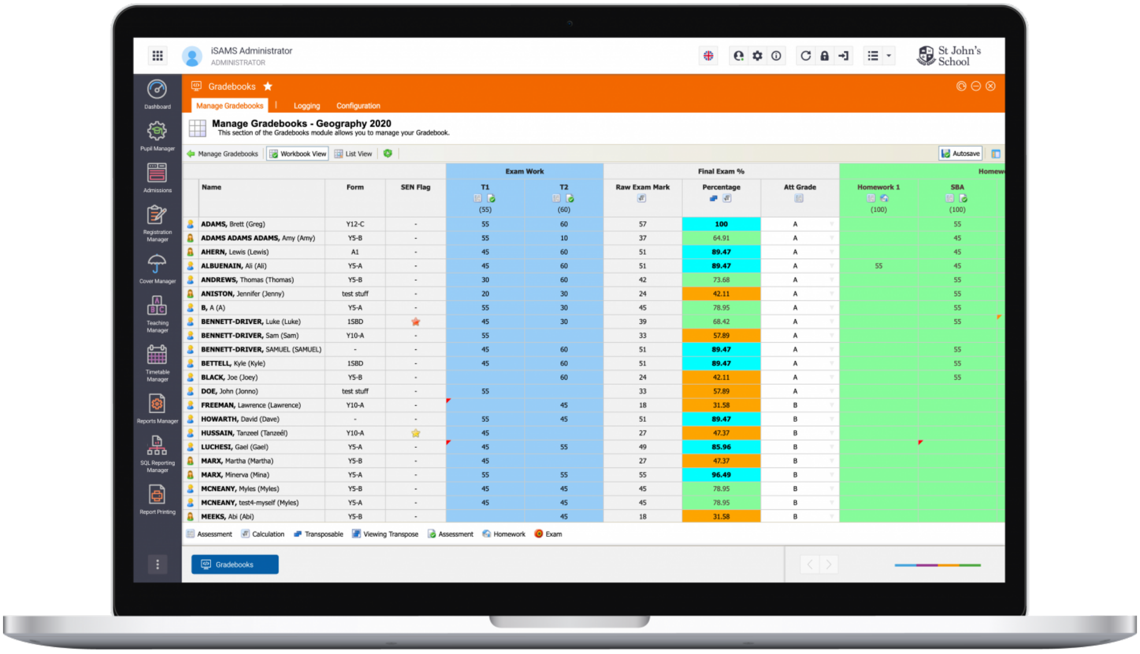

Figure 2.9: Example of an administrative school system [5]

This figure from a similar platform called iSAMS shows a spreadsheet-like tool with prefilled columns. If an "all-in-one" tool wants to truly be adopted by teachers widespread across the UK, then it needs to be interoperable with different teaching preferences. If this is not achieved, then teachers will resort to leveraging siloed tools to store their information. This renders platforms like KIM redundant for providing valuable insight for teachers to use when coming up with actionable plans for students. The fact that teachers resort to siloed systems under a non-interoperable source of truth limits the ability to use AI or data processing to inform educational decisions [45].

This creates a challenge. How does one design a "one size fits all" solution for teachers to adhere to, independent of teaching style or assessment preferences? Where does one draw the line between a solution being overly restrictive to teachers, whilst also being structured enough to be useful as a RAG knowledge base? The key to unlocking an edTech tool that can be widely adopted is one that has a knowledge base that is universal to all teachers in the UK. This is an abstraction that can only be made on a semantic level.

Semantic ontologies present a potential solution by creating a framework for knowledge representation that is both adaptable and standardised. By defining the relationships between concepts, terms, and knowledge hierarchies, ontologies can abstract away the diversity of teaching styles and assessment methods while retaining the core data at hand. Fundamentally, an ontology describes the "verbs" and "nouns" of a teacher's workflow in a way that is also understandable by a digital system. This approach enables a common language for data, ensuring that information is interoperable across different schools, teaching approaches, and even subject areas. Moreover, it facilitates the development of AI systems capable of reasoning over this shared framework, providing tailored insights without imposing a one-size-fits-all methodology. This sets the stage for understanding how ontological infrastructures can underpin a robust and scalable LLM tool in British schools.

2.4.4 Ontology

Ontologies provide a structured, semantic representation of knowledge that defines the relationships between entities, concepts, and data in a specific domain. Ontologies enable systems to reason about information in ways that reflect its real-world context [16], by storing an abstracted form of information that represents the nouns and verbs of the workflow or enterprise. As previously discussed, a structure like an ontology can provide the necessary scaffolding needed to give the LLMs the context they need, without being restrictive on teachers. Data can be integrated from siloed sources into an ontological structure. The ontology can model relationships that are fundamental to the workflow, for example:

- *Objects* may include entities such as a "teacher", a "student" or a "subject"[46].
- *Relationships* include the links between the objects - for example "a teacher has students"
- *Attributes* include metadata about the objects, for example the grades a student has achieved, or their previous report cards

Through explicit semantic relationships, ontologies resolve ambiguities in data. For instance, the term "performance" could refer to academic achievement, attendance, or participation in extracurricular activities. An ontology links "performance" to specific contexts (e.g., "academic performance is evaluated using grades"), enabling LLMs to generate precise and contextually appropriate responses [16]. Furthermore additional context can be supplied from the ontology to help implement a RAG architecture to further reduce the chance of hallucinations. The culmination is infrastructure that can be used to make LLMs reliable and accurate, without requiring additional buy-in by leveraging existing data, whilst still being convenient to educators.

2.5 The Palantir Ontology

Accurately representing an ontological structure requires a great deal of scaffolding. Palantir Technologies is a commercial contractor that has developed a way of representing an Ontology. At its core, it integrates three primary layers: the data layer, the semantic layer, and the application layer.

The Ontology is constructed as follows [46]:

- **Data Layer:** The first layer ingests raw data from siloed and disparate sources - this can be any source of data, such as, but not limited to: CSVs; structured databases; cloud syncs; streaming pipelines and more. Palantir's ability to connect directly with external systems while maintaining high fidelity and lineage of the source data is a key advantage. In order to do so, they have built numerous connectors to unite these separate data sources into a single parseable format.
- **Semantic Layer:** The semantic layer is the heart of the ontology. It organises raw data into meaningful objects and relationships that mirror the real world. For example, in an educational context, entities like "student," "teacher," and "classroom" may be modeled, along with their interrelations, such as "teaches" or "studies". This layer allows for intuitive querying and logical reasoning, as demonstrated by successful implementations in sectors such as healthcare and defense [4].
- **Application Layer:** The application layer enables end-users to interact with the ontology. This will allow external tools to be leveraged to interact with the ontology, and can act as a gateway to the semantic layer for any external LLM models.

By leveraging these layers, the ontology bridges the gap between raw and siloed data and actionable insights, enabling teachers to derive value while maintaining robust governance and security protocols [7].

2.5.1 Ontology Software Development Kit (OSDK)

The Ontology Software Development Kit (OSDK) is a developer-friendly SDK (Software Development Kit) that allows developers to leverage the Ontology in their applications [46]. It provides APIs and libraries that can be embedded into wider applications to enable developers to connect their app to the Ontology - and in our case to LLM logic.

2.5.2 Data Lineage

Effective data governance necessitates the need for traceability of data lineage. One of the most existential issues with platforms and tooling that processes sensitive data is the difficulties in auditing how the data has been processed. As aforementioned, this problem is only exacerbated by the several siloed data sources that need to be joined in order to create a comprehensive tool. Tracking data lineage of datasets that back ontology objects is an implicit ability in Foundry - this can be leveraged to identify on a granular level where data has been processed for auditing purposes [7].

2.6 Research Questions

2.6.1 Motivating the Teacher's Requirements

As a secondary school teacher part-time, as well as by exploring the problem space with several UK educators, it is clear the UK-based teachers have a huge amount of expectation and pressure to provide tangible advice to guide the student. Indeed, many teachers find this gratifying, and it served as one of the primary motivators for them to dedicate their lives to the field too.

However, as previously discussed, siloed data sources and manual processes make this a time-consuming task [47, 48]. Teachers are expected to store an unreasonable volume of information for quick access about hundreds of different students in order to generate truly customised and bespoke action plans. This process becomes either bloated and slow, or ineffective and repetitive as there either aren't enough resources, or isn't enough time to allocate to a single student.

Moreover, there is often no unified platform that allows teachers to holistically view a student's progress across academic *and* pastoral dimensions. These limitations point to the need for an integrated solution that supports fast and intuitive information access.

2.6.2 Motivating the Need for a Federated Application

Additionally, as aforementioned, many teachers are starting to adopt the use of generative AI and large language models in particular in their workflows. However, as previously discussed, trust and security remain one of the biggest barriers to a widespread adoption.

It is not sustainable for a governing overseeing body such as the Department of Education to centralise systems into a single monolith. Such a system would pose risks around performance bottlenecks, compute resources as well as acting as a single point of failure in securely protecting the privacy and civil liberties of students all around the country. Furthermore, each individual school may run in its own way, with unique data privilege hierarchies and security requirements.

A federated application architecture allows individual schools to remain as legal data controllers, but through a consistent and secure interface [49]. It also allows developers to construct modular applications that can evolve independently, rather than having to re-engineer entire systems to add new functionality. This project aims to leverage such a federated approach using the OSDK (Ontology SDK) to create a proof-of-concept system backed by ontological infrastructure and deploying federated LLM models in British schools.

2.6.3 Research Questions Posed

Based on the above motivation, this project seeks to explore the following research questions:

1. **How can siloed data be integrated and retrieved by teachers using natural language?**
2. **How can trust be built by reducing bias, improving retrieval accuracy and implementing granular permissioning systems similar to permission hierarchies currently prevalent in schools in the UK?**

3. **Is the proposed tool comparable or more effective than current expert teacher workflows? Are there any areas in which it is unsuitable and not fit-for-purpose?**

4. **To what extent can such an application be generalised beyond a single school setting via an appropriate federation architecture?**

The rest of this paper explores the architectural design, implementation and evaluation against several benchmarks of the proposed tool to ultimately determine its viability.

Chapter 3

System Architecture

3.1 Ontology

Clearly, the robustness of the system will be largely defined by how cohesive the structured ontology that captures core educational entities and their interrelationships is. Based on forward-deployed experience in schools in conjunction with background research, the following Ontology is proposed to have the best coverage of all the key attributes and links within a UK school system.

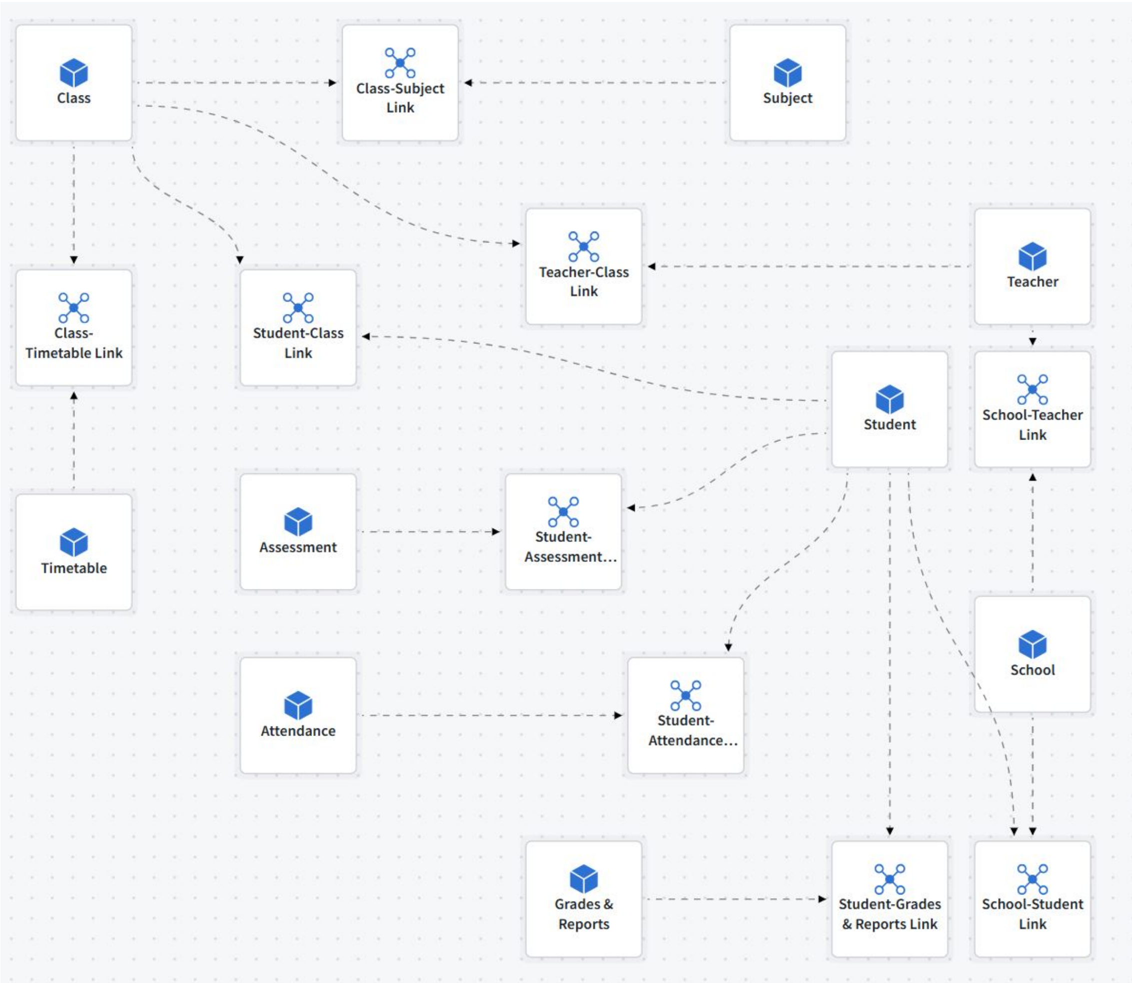

Figure 3.1: Education Ontology Design for British Schools

Student

Attribute	Description
StudentID	Unique identifier for the student
Name	Full name
Address	Home address
DateOfBirth	Date of birth
SEN Information	Special educational needs metadata

Class

Attribute	Description
ClassID	Unique class identifier
ClassName	Human-readable name
YearGroup	Year group associated with the class
RoomNumber	Room where the class is held
Subject	Linked subject
Teacher	Linked teacher

Teacher

Attribute	Description
TeacherID	Unique identifier for the teacher
Name	Full name
Subject	Subject(s) taught
Email	Contact email address

Grades & Reports

Attribute	Description
ReportID	Unique identifier for the report
Grade	Assigned grade
Comment by Subject	Subject-specific comment
Self-Reflection	Student's own reflection
Tutor Comments	General feedback from form tutor

School

Attribute	Description
SchoolID	Unique identifier for the school
Name	Name of the school
School Type	E.g., Primary, Secondary, Academy
Address	Postal address
OfstedRating	Ofsted inspection score
Media	Associated images or files

Assessment

Attribute	Description
AssessmentID	Unique identifier for the assessment
Name	Assessment title
Date	Date of the assessment
Type	E.g., Exam, Quiz, Coursework
Topics	List of topics assessed
Subject	Linked subject

3.1.1 Links and Relationships

Each object type is connected via defined links that establish the ontology's graph structure. These links ensure traceability and consistency across the educational data landscape.

- **Class–Subject Link:** A class has a subject.
- **Teacher–Class Link:** Teachers teach many classes [e.g. for several year groups].
- **Student–Class Link:** Students are enrolled in many classes, where each class can have many students.
- **Class–Timetable Link:** Maps each class to a set of scheduled timetable entries.
- **Student–Assessment Link:** Records which students have which assessments.
- **Student–Attendance Link:** Logs attendance records for each student.
- **Student–Grades & Reports Link:** Links students with their report cards and grades.
- **School–Teacher Link:** Registers teachers as employed by a specific school.
- **School–Student Link:** Registers students under a specific school.
- **School–Subject Link:** Defines the subjects offered at a particular school.
- **Teacher–Subject Link:** Links teachers to the subject(s) they are qualified to teach.
- **Teacher–Assessment Link:** Identifies the teacher responsible for an assessment.

3.1.2 Design Rationale

The semantic nature of the ontological approach best facilitates natural links between object types. In several cases, such as the Student-Class link, the above architecture is not exhaustive in terms of the underlying representation as these many-to-many relationships require extra metadata [typically a link table in the link type] to store all granular information.

3.2 Data Integration

In order to provide hydration to the ontology, each object type [and in some cases, link types] need to be backed by corresponding datasets where the dataset columns match the object type attributes. These datasets do not exist as mirrored mappings of the Ontological design in the real world - as previously discussed, these often sit in siloed sources that are often outdated or filled with gaps and/or erroneous and null values.

In order to achieve a suitable set of datasets that can be used to hydrate the proposed Ontology, data can be integrated from all these siloed sources into a single platform in Foundry.

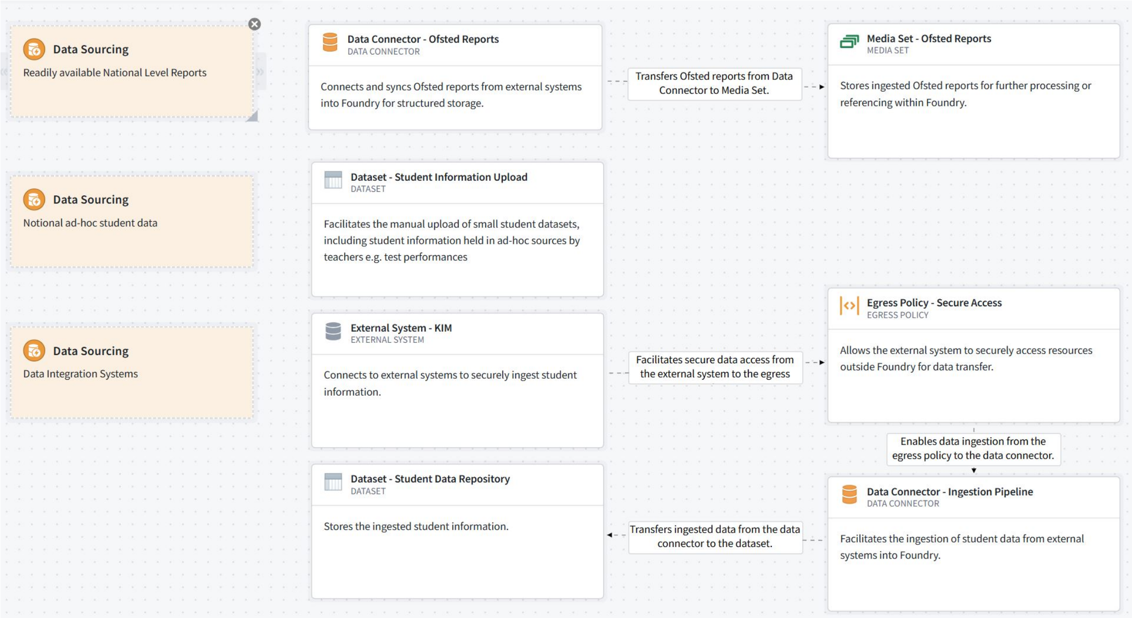

Figure 3.2: Data integration of siloed sources

Data can be integrated from 3 main sources:

1. **Publicly available reports from the Internet**. These will typically be free-form files such as PDFs that contain overarching published documentation from the Department of Education, or on a more granular level inspection reports from a school.

2. **Data Integration of school systems**. Systems such as KIM can have data connectors written for them to integrate the import schema and data into Foundry.

3. **Notional ad-hoc uploaded data**. Ad-hoc siloed data typically stored in spreadsheets, CSVs etc. can be integrated directly into Foundry.

From here, big data Apache Spark operations can run filters, joins and materialisations of these datasets to achieve the desired format. This can then correctly back the ontology, whilst also providing a ground for logical transformations to not only format the data, but also to handle erroneous and null values. The result is a functionally appropriate set of datasets for a valid hydrated ontology.

3.3 System Architecture

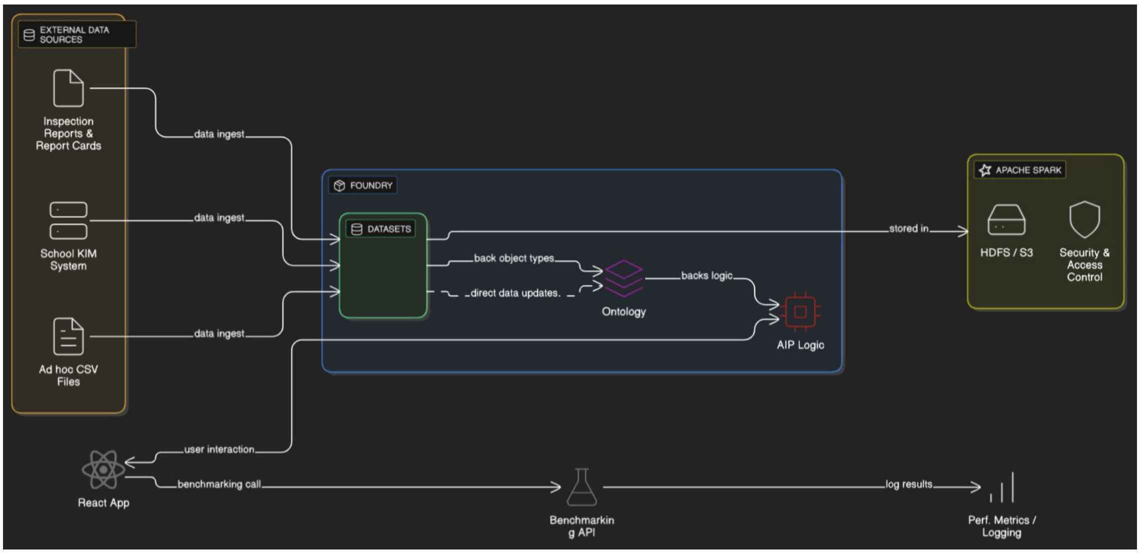

Figure 3.3: System architecture diagram created using Eraser diagram creator [6].

3.3.1 Overview

Figure 3.3 illustrates the overall architecture of the proposed system. The architecture integrates the siloed data sources into Foundry, and communicates with the user facing app through the OSDK. Data storage and processing is cloud-based using Apache Spark. There is an additional backend service for benchmarking and internal APIs.

3.3.2 External Data Sources

The system ingests structured, semi-structured and unstructured data from three main siloed sources:

- **Inspection Reports and Report Cards:** PDFs ingested manually into the platform.
- **School data systems:** As a notional example for the proof-of-concept application, a schema can be scraped from KIM.
- **Ad-hoc CSV Files:** Uploaded manually or on a scheduled basis to accommodate miscellaneous datasets that a teacher may store.

3.3.3 Foundry

All data is ingested into Foundry, where it is represented as version-controlled **Datasets**. These datasets are stored using Apache Spark and also have provisions for security and privacy controls. In this use-case, security and access controls are set up using groups and markings on the datasets, which is discussed later.

3.3.4 Ontology Layer

As aforementioned, the ontological structure contains object types and link types. These are backed by specified cleaned datasets that match the schema in the ontology.

3.3.5 AIP Logic

The AIP Logic module encapsulates model-based logic., In the scope of this project, this serves two main purposes:

- **Retrieval** - applying retrieval mechanisms from the ontology. The 5 methods for retrieval are discussed in more detail in the implementation phase.
- **RAG Model** - the AIP Logic contains the relevant logic for the RAG model.

As a result, the AIP Logic module needs to facilitate bidirectional communication to both read prompts from the user facing application, and also to output the result of the RAG. Additionally, the AIP Logic module needs to read from the ontology to gain the semantic contextual knowledge needed for the LLM.

3.3.6 OSDK and Benchmarking API

The user facing application leverages the ontology software development kit [OSDK] to coordinate teacher actions with the ontology. This provides the interface for retrieval of data for relevant dashboards, as well as the gateway to the RAG model for the teacher to gain insights on a particular student.

The frontend connects to a Flask API that has endpoints to run benchmarking models that further leverage the AIP Logic module. These are either logged to datasets, the console, or back to the main user.

3.4 Federation of Foundry Instance

There are many motivations and justifications for the chosen system architecture. Largely, this is to ensure the effective federation of the solution at a national scale across schools in the UK.

- **Modularity:** Each component is independently scalable and replaceable. This further facilitates federation as each component can be individually exposed or modified as needed depending on the specific school's systematic setup.
- **Audit Logs:** All ingress and egress relative to Foundry is logged to a versioned dataset. This provides practical scope for federation, as the ability to effectively audit and track data lineage would be critical for national adoption.
- **Security:** Granular security controls can be configured at a dataset level - these can be adapted depending on data regulation hierarchies and permissions prevalent in different schools around the country.

By adopting the above architecture, the created application can serve as a proof-of-concept of a viable instance of the app, with the surrounding architecture and infrastructure that would facilitate safe and privacy-preserving federation on a national scale.

Chapter 4

Implementation & Methodologies

4.1 Data Integration

As aforementioned, the types sources of data needed to provide maximal coverage of a school - both granular student views as well as a holistic view of the entire school - is threefold. These include **integration from school data systems like KIM**, **ingesting school inspection reports and [anonymised] student reports** and lastly creating **ad-hoc CSVs** that teachers use for miscellaneous information. It is key to ingest all of this information into Foundry together to facilitate data cleaning and necessary joins to back the ontology object types and links.

4.1.1 School Data Systems

To create a functional proof-of-concept, a real-world example of a school data system - KIM - was used as a baseline system. In order to comply with GDPR requirements, notional data was used for the project - however, to replicate the school environment, web scraping was leveraged to integrate the correct schema from KIM - even if the data held within the schema was adapted to be notional.

Since KIM has no public API, the python library `BeautifulSoup` was used to scrape the relevant schema from the KIM page. This was parsed to extract useful student metadata such as names, classes, marks and timetabling information. This could be ran within Code Repositories [or Authoring] within Foundry [50] to generate an output of native datasets that had the correct schema. Lastly, the schema could be filled in with notional data.

4.1.2 Ofsted and Inspection Reports

Inspection reports from either Ofsted [for UK State and Grammar schools] or the ISI [for UK Independent schools] were manually downloaded depending on the school in question. These were stored into media sets [51] into Foundry for ingestion. This facilitates future processing discussed in 4.2.3 for extracting and parsing meaningful information from these PDFs.

Anonymised report cards for students were also ingested through a similar process. Real report cards were used, and anonymised using the algorithm and processes described in 4.3.3 to ensure data integrity was maintained by redacting or aggregating any PII. Critically, the report cards had to be from schools that matched the information ingested via KIM as well as the inspection reports, to ensure that there were no contextual ontological clashes.

4.1.3 Ad-hoc CSVs via Fusion

Plenty of ad-hoc data - such as specific topic test marks - is stored in miscellaneous storage methods in schools. The most common is a CSV or excel spreadsheet. In order to replicate this, Fusion [52] can be used within Foundry to not only provide an excel-like interface, but also to facilitate bidirectional communication with the native dataset configurations in the platform [53].

4.2 ETL Pipelining

ETL [Extract-Transform-Load] pipelines are a standardised approach to go from raw data in siloed source systems into structured, usable formats [54]. The **Extract** stage collects data from external sources, the **Transform** stage reshapes and cleans it, and the **Load** stage writes it into a final structured format ready for analysis. This maps perfectly into the Foundry pipeline, in which data is **extracted** from siloed sources into a singular platform; data is then **transformed** into a usable format that can then **load [or hydrate]** the Ontology - the final usable format by the application.

4.2.1 ETL Pipelines

Given the ingested data from siloed sources, the data can now undergo transformations. These are done incrementally using Apache Spark transformations into structured Object Types such as `Student`, `Class`, as well as more metadata Object Types such as `ReportChunk` for analysis of Report Cards or Inspection Reports. This allows downstream components [e.g., the retrieval layer] to operate on semantically meaningful data objects.

4.2.2 Object Type Backing Datasets

Each Object Type in the ontology (such as `Student`) is backed by a Spark SQL dataset in Foundry. These datasets are defined by deterministic transformation logic from raw ingested datasets, that can typically be materialised in the form of Apache Spark frameworks such as `PySpark` or `Polars`. Transformations such as `filter`, `join`, and `selectExpr` are used to clean combine the siloed raw data.

```
def transform_report_chunks(reports, students):
    return (
        reports
        .filter("text IS NOT NULL")
        .join(students, on="student_id")
        .selectExpr(
            "uuid() as id",
            "student_id",
            "school_id",
            "chunk_text",
            "created_at"
        )
    )
```

Listing 4.1: Filtering and joining dataframes to back the Student Object Type with the report card of the student

This means that we can create Object Type Backing Datasets to mirror those provided in the Ontology design, despite any schema mismatch in the ingested data. [55].

Erroneous and Null Values

The raw ingested data sources are likely to have null or erroneous values [e.g. values that are out of date]. In order to prevent these errors propagating further downstream and infecting the Ontology, it is necessary to have robust methods for dealing with these gaps in the transformation stage.

Strategies include:

- Row dropping: skipping malformed rows using Spark's `dropna()` or explicit filters. Note this can only be done for non-critical data to avoid major data mismatches.
- Schema Casting: casting data to expected types using Spark's `withColumn` transformations.
- Default value imputation: e.g., using placeholder strings like `"Unknown"` for school names. This will allow the Ontology to correctly index whilst providing a clear flag to the user.

```
cleaned_df = (
    raw_df
    .dropna(subset=["student_id", "text"])
    .withColumn("created_at", to_date("created_at", "yyyy-MM-dd"))
)
```

Listing 4.2: Handling nulls and malformed rows

These practices ensure that the ontology remains stable even when fed noisy real-world data, as is common in educational settings [56].

4.2.3 Free-Form PDF Decomposition

Documentation such as Ofsted inspection reports or report cards of students are unstructured and typically free-form PDFs. These require extra care to process, due to their non-tabular nature. An additional pipeline can be constructed to extract and chunk semantically meaningful content. This follows three key stages: **Entity Extraction**, **Sentiment Analysis**, and **Semantic Chunking**.

SpaCy for Text Extraction

The two most common types of PDFs that will be ingested are inspection reports and student report cards. In order to extract useful information from these, the fact that they are both largely text based can be leveraged to choose an appropriate technique. Once text is extracted, the PDF can still be in an inconvenient format for effectively parsing information. This is also largely due to different reports having differing structures and formats, such as the usage of paragraphs, metrics and tables. Consider for example, the two following example structures of report cards. The latter may have a more nuanced and ambiguous textual extraction that makes it difficult to parse useful information.

Chemistry	Mrs K Sunil-Atkinson & Mr J Young
Projection Grade	**Approach to learning**
B	Consistently meeting expectations

[redacted] has clearly enjoyed the analysis aspects of the course with her problem solving nature and has a good understanding of nitrogen chemistry. She engages well at all times with questions that stretch the class but needs to exercise greater caution in reading and responding to questions in exam conditions. Practice will be key to improvement in this area. On the other side of Chemistry, she has mastered the topic of Electrochemical cells and can effectively draw these. She has also enjoyed using log calculations to work out the pH of Acids and Bases. To further her understanding she should practise more challenging questions from the textbook and show all working out to avoid errors in calculations.

Computer Science	Mr K Wong & Mr S Wilson
Projection Grade	**Approach to learning**
A	Consistently meeting expectations

[redacted] has a solid grasp of Computer Science theory, which provides a strong foundation for her success in the subject. Programming is an area she can continue to develop, but this does not affect her potential to excel in the subject. After all her entrance tests and interviews are settled, the project will be the main focus. She knows what she needs to do, and so she should make the completion of her target. Then, in the Spring Term, she should start consolidating what she has learnt to prepare for her Mocks.

(a) A report card with a paragraph style structure

Next Steps

Subject	Comments
Biology	[redacted] needs to continue to take responsibility for practising independently and regularly, using the specification as a guide and incorporating retrieval practice and past paper questions into their revision routines to help improve her performance in examinations and extended response questions. [redacted] should particularly focus on the Infection and response topic which she finds one of the most challenging areas of the course.
Chemistry	To progress further [redacted] needs to continue to incorporate retrieval practice and past paper questions to improve her examination technique. Based on her recent mock examination, [redacted] needs to revisit topics including Chemical analysis and Quantitative chemistry.
English	To continue progressing, [redacted] should ensure that she is thoroughly planning her extended response; without a clear plan, [redacted] answers can lose focus on the question. To make further progress, [redacted] should take time to ensure that written responses are coherent and clearly express the point intended.
French	[redacted] should practise reading and listening comprehension on a regular basis in order to familiarise herself with the types of question which appear at higher level. She should attend the extra practice session each Monday lunchtime.
Geography	Ensure that your longer responses fully answer the question. It can help to use some of the language from the question in your answer to make sure that you stay focused.
Mathematics	[redacted] needs to use the Sparx, Maths Genie and Corbett maths websites to spend time working on closing gaps in topics evident from her recent mock exam. If she can watch a video to recap knowledge and then complete extra questions on the topics of Similar triangles and Non right-angled triangles this will stand her in good stead for her end of year exams. [redacted] needs to complete past papers to ensure she is familiar with the style of questions that appear right at the end in the examinations and find effective ways to apply her knowledge to these Grade 9 questions.
Physics	[redacted] should continue to practise calculations in physics to build up her confidence in converting units and in solving problems with two or more equations. She should also practise past paper examination questions requiring an experimental method to make sure she can provide sufficient detail and state what equipment is used for every measurement made.
Design & Technology	Ensure your final design idea aligns closely with your design specification and target market. You should develop a deep understanding of the materials you have chosen to use, demonstrating expertise in their properties and applications.

(b) A report card with a tabular structure

Figure 4.1: Two differing structures for school report cards

SpaCy is a free and open-source NLP library in Python [Cython] that is used for tasks such as tokenisation and Named Entity Recognition [NER] - this makes it suitable for processing the extracted text from PDFs that may have a noisy format.

The raw text can be tokenised via the SpaCy pipelines using a deterministic rule-based tokeniser designed to handle edge cases in english text; including hyphenation, abbreviations, and irregular punctuation [57]. This is particularly useful for things like inspection reports - that are usually written in formal language and punctuation - but also school reports that are often written in a rush by teachers - and hence may suffer when it comes to using abbreviations or poorly punctuated sentences.[58]

Next is the NER phase - which detects and groups relevant entities in the text. In the context of school or inspection reports, this could be inspectors, departments, dates, or locations [or more]. SpaCy's NER leverages a convolutional neural network [CNN] to predict entity boundaries and types [59]. This makes it possible to automatically extract structured metadata—such as inspection dates, school names, and reported outcomes—from otherwise unstructured textual data. This provides scaffolding to the raw extracted text that can make it easier to parse further down the pipeline.

Entity Extraction

Entity extraction involves identifying and labeling named entities in free text—such as students, subjects, and schools—using spaCy's pretrained transformer-based models [60]. These entities can later be linked to canonical representations within the ontology.

```
import spacy
nlp = spacy.load("en_core_web_trf")
doc = nlp(pdf_text)
entities = [(ent.text, ent.label_) for ent in doc.ents]
```

Listing 4.3: Extracting named entities from raw PDF text

This ensures that even free-form prose can be connected to structured identifiers such as object primary keys.

Sentiment Analysis

The extracted text is also passed through a lightweight sentiment analysis model (e.g., a RoBERTa-based classifier fine-tuned on educational feedback) to assign tone scores to each paragraph or chunk. This enables downstream UIs to quickly assess the affective content of reports and sort/filter accordingly [61].

```
from transformers import pipeline
sentiment_analyzer = pipeline("sentiment-analysis")
sentiment = sentiment_analyzer(chunk_text)[0]["label"]
```

Listing 4.4: Basic sentiment scoring for report text chunks

Sentiment scores are stored alongside chunk metadata in the final `ReportChunk` object type.

Semantic Chunking to Produce Chunks

To support retrieval, free-form reports are divided into discrete "chunks," each representing a single coherent unit of meaning. Each chunk is assigned a unique ID and stored with both its raw text and metadata. Chunking is performed using a transformer-based segmentation model that identifies logical breakpoints such as paragraph endings or topic shifts [62].

These chunks are then embedded and hydrated into the `ReportChunk` Object Type:

```
def chunk_text_blocks(text):
    chunks = segment_into_chunks(text)
    return [
        {"id": uuid4(), "chunk_text": chunk, "embedding": embed(chunk)}
        for chunk in chunks
    ]
```

Listing 4.5: Assigning chunk IDs and storing chunks

The resulting `ReportChunk` dataset forms the core backing dataset for retrieval methods described in Section 4. Each chunk is semantically coherent, embedded in vector space, and annotated with metadata such as sentiment, topic, and student linkage.

4.3 Privacy-Preservation & Security

GDPR regulations enforce strict requirements around data storage and processing - as previously discussed, this is particularly relevant for any software that leverages personally identifiable information [PII] of minors unders the age of 18 [63]. As a result, the infrastructure of the application must facilitate rigorous policy-driven access controls to this data, and ensure that it is kept securely - and only viewed by those who can and should be viewing the data.

This section discusses the practical steps taken to implement a robust and GDPR-compliant access control system for the ingested data. Additionally, the theoretical implementation of federating the solution for the Department of Education to deploy nationally to schools across the country is discussed in relation to the practical configuration. Federation [64] is key to ensure that each individual school remains a data controller of the student's of the school, ensuring critical PII isn't transferred or processed outside of the school's jurisdiction.

4.3.1 Purpose-Based Access Controls (PBAC)

PBAC is a permissioning model to enforce data minimisation - this facilitates governance by principle of least privilege, where a user of the platform only sees the data that they need and are legally able to inspect on a granular non-aggregated level [41].

Within the Foundry environment, PBAC can be practically implemented using two methods - groups and markings [both data markings and project markings] [41]. The combination of these two ensure a robust privacy-preserving data system.

Groups

Groups can be defined for members who have common sets of permissions. Similar to the ontology, these groups can be defined to reflect real world hierarchal groups that mirror the structure of real school systems [65]. Furthermore, as reflected in a real school system, an employee or teacher can be part of several groups, and as a result can have access to the lenient union of datasets governed by all of their constituent groups.

Let:

- $\mathcal{U}$ be the set of all users.
- $\mathcal{G}$ be the set of all groups.
- $\mathcal{D}$ be the set of all datasets.
- $G_u \subseteq \mathcal{G}$ be the set of groups that user $u \in \mathcal{U}$ belongs to.
- $A_g \subseteq \mathcal{D}$ be the set of datasets accessible to group $g \in \mathcal{G}$.

Then, the set of datasets accessible to a user u is given by:

$$A_u = \bigcup_{g \in G_u} A_g$$

That is, the user has access to the union of all datasets accessible to each group they are a member of.

This facilitates real world purpose based access controls - for example, a teacher may be part of the `Maths Teacher of Class A` group and have access to their own maths class, but also part of the `Senior Leadership Team` group, providing them granular access to all pupils in the school [and by extension, to maths class A].

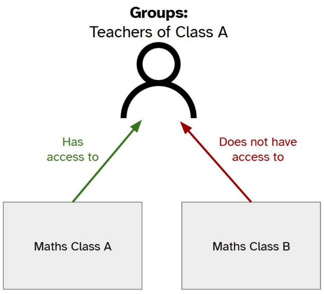

Figure 4.2: (a) A teacher who is only a teacher of maths

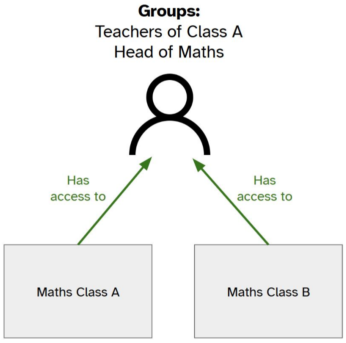

Figure 4.3: (b) A teacher who is a teacher of maths, but also the Head of the Maths department

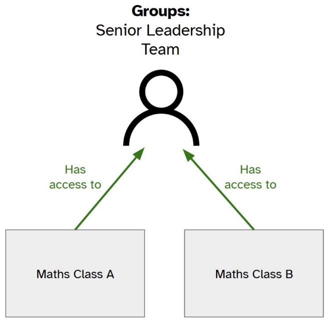

Figure 4.4: (c) A non-teaching staff who is part of the SLT

Figure 4.5: Examples of group based PBAC for two given maths classes.

For the purposes of the notional app, a few groups were created as a proof-of-concept of privacy preservation:

- `Teacher of Class A`
- `Teacher of Class B`
- `Pastoral Leads` - non-teaching pastoral leads and SLT team
- `Designated Safeguarding Leads [DSLs]` - teaching designated safeguarding leads - present at every school

This provides enough scope for testing of a minimal system, that proves full coverage of the group-based permissioning system for all different roles in a school.

Markings

In conjunction with the group-based system, it is important to protect data on a more granular level - especially where PII data is concerned. This needs to act from source integration stages, rather than just on resultant user-facing datasets in the way groups interact. To achieve this, a construct called markings can be leveraged.

Markings can be applied at both a dataset level, and also a project level [7]. Groups can acquire access to markings - if a group does not have all the markings present on a dataset, then the members of that group cannot see the data inside.

For the purposes of the platform, the key marking to be defined is the `PII marking` to protect identifiable information for both teachers and students in accordance to GDPR regulations.

Lastly, markings can propagate to ensure downstream datasets maintain the same permissioning rules decided at the data integration phase. If any source dataset in a join or transform has a marking, then those markings will be propagated to the downstream results.

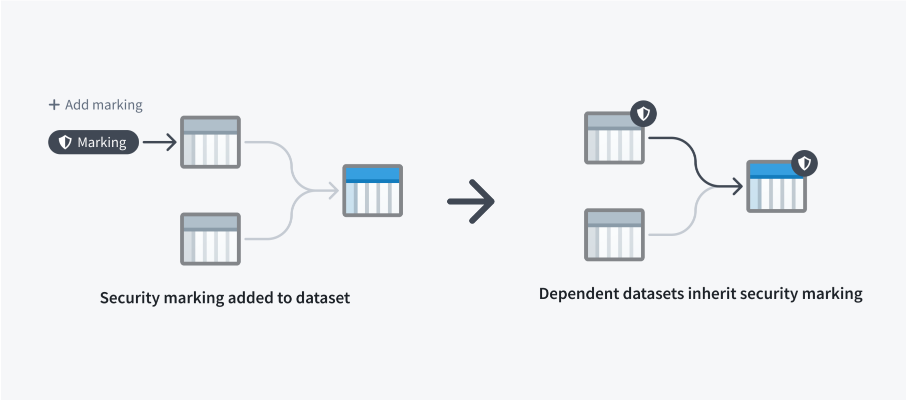

Figure 4.6: Marking propagation [7].

The final edge case to tackle is when PII is aggregated during a Spark transformation. Such an aggregation would remove any information that is personally identifiable. Hence the marking would no longer be needed, and can be removed by adding the `stop_propagating` tag to the transform:

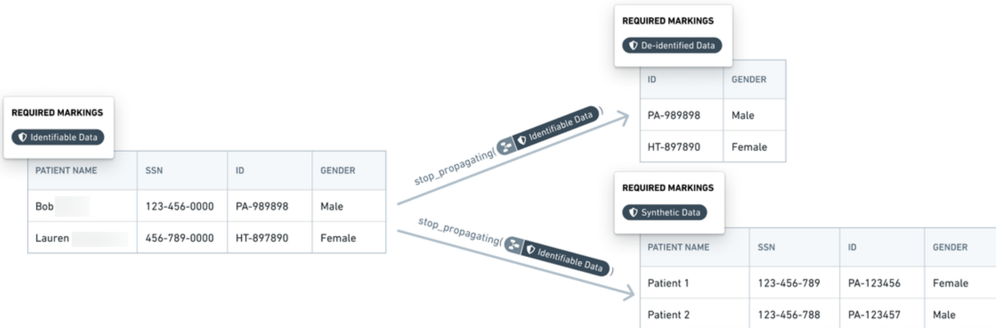

Figure 4.7: Removing markings for aggregated data [7].

4.3.2 Anonymisation and Notional Data

As mentioned earlier, the types of data ingested into the platform are threefold - inspection reports and publicly available documents, which contain no PII; notional data like test scores, which are notional by nature and so contain no real PII; and lastly real report cards - these *do* contain personally identifiable information of real current or former students.

In order to comply with data protection obligations, the report cards containing PII need to be anonymised if they cannot be replaced with notional data. According to the UK Information Commissioner's Office [UK ICO], data is considered anonymised when individuals are no longer identifiable by any means reasonably likely to be used [66]. Typically this involves any objective inference resulting in no less than 5 potential people who match the data profile. As a result, the following PII types were identified for removal *before* ingestion into the platform.

Table 4.1: PII types and redaction strategies

PII Type	Example	Redaction Strategy
Full Name	"Hashir Majeed"	Replaced with "the student", or fully redacted
Date of Birth	"14/08/2004"	Redacted completely
Address	"Imperial College London"	Redacted completely
Medical Info	"Asthma"	Redacted or generalised to category
Exact test scores	"88%"	Redacted or edited

4.3.3 Report Card Redaction Process

A report card redaction process was defined to remove the PII outlined above. A simplified pseudocode algorithm is shown in Algorithm 1 - whilst the anonymisation was done by hand, the algorithmic process ensured full coverage and ultimate data integrity.

Algorithm 1 Report Card PII Redaction

```
Require: Text document T, List of PII types P
for each token t in T do
    for each PII type p in P do
        if t matches pattern of p then
            Replace t with placeholder [p]
        end if
    end for
end for
return Redacted document T'
```

4.3.4 Audits and Logs

In order to maintain integrity of data and operations on data, auditing is a key mechanism to preserve notions of accountability - this is especially important for privacy-sensitive domains like education. For user actions relating to core datasets, audit logs are generated into a logging dataset by timestamp - logging information such as, but not limited to [67]:

- **Timestamp** - when the action took place
- **User ID** - which user completed the action
- **IP** - IP Address of the user completing the action
- **Request Params** - The metadata detailing the requested action

It is worth noting that since the user ID is contained, this log also contains PII. However, since it is also a dataset, it can also be marked with the appropriate `PII Marking` to mitigate this issue.

4.4 Dataset Builds and Propagation

For teachers to be empowered to provide the most effective insights they can, it is imperative that the data they are using - namely the data that backs the ontology - is not stale, and is updated to reflect live changes in source systems. To achieve this, schedules in Foundry can be configured [68] to enable datasets to be dynamically built according to certain conditions.

To maintain fresh data, schedules can leverage the `Trigger and Target` methodology - source datasets act as triggers, such that when they are updated, the schedule propagates the build request to all downstream dependencies, namely, the targets [69].

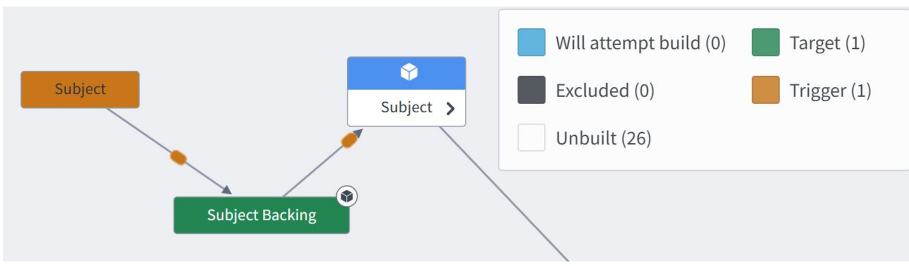

Figure 4.8: Trigger and Target schedule for Subject Datasets and Object Type

4.5 RAG Architecture with the Ontology

In order to achieve an effective RAG implementation, the following steps are needed:

1. **Retrieval** - data needs to be retrieved from the ontology. This can be done through a variety of methods which each have their own merits - the implementation explores several of these in 4.6.
2. **Augmentation** - the system prompt that is designed to bring together and leverage the ontological data.
3. **Generation** - the use of off-the-shelf models to output an augmented result to the user.

4.6 Implementation of Retrieval Methods

To achieve a robust retrieval function, there are several methods that can be used. These are Keyword Search, Vector Cosine Search, Augmented Keyword Search, HyDE, and a combination of any of the above [RRF Method]. Each method supports different strengths, allowing the system to retrieve relevant information in diverse query contexts. Comparing the strengths and weaknesses of each implementation can facilitate benchmarking of methods against educational queries by teachers in the application.

4.6.1 Keyword Search

Keyword Search is a perhaps the most trivial information retrieval approach that identifies relevant text chunks by exact token match. For a query q comprising tokens $\{t_1, t_2, ..., t_n\}$, a report r is considered relevant if:

$$\exists t_i \in q \quad \text{such that} \quad t_i \in r$$

The main advantages of this method is that it is simple, fast, and highly interpretable [70]. Clearly, however, it would perform worse when a report card or document uses different semantics and vocabulary to express the same point.

```
return Objects.search()
  .reportChunks()
  .filter((chunk) => chunk.chunkText.matchAnyToken(query))
  .orderByRelevance()
  .takeAsync(k);
```

Listing 4.6: Keyword retrieval function

4.6.2 Vector Cosine Search

This method uses vector embeddings to represent both the query and documents in a single semantic space. Let v_q be the embedding of the query and v_r the embedding of a document chunk. Relevance is computed using cosine similarity:

$$\text{similarity}(v_q, v_r) = \frac{v_q \cdot v_r}{\|v_q\| \|v_r\|}$$

This enables retrieval based on semantic similarity rather than exact matches, making it likely to be more robust than exact keyword search [71]. OpenAI's `text-embedding-ada-002` model can be used to create 1536-dimensional embeddings, that can be leveraged for more accurate retrieval. The extent of the accuracy can be investigated in a later benchmarking phase.

```
let query_vector = createEmbeddings(query);
return Objects.search()
  .reportChunks()
  .nearestNeighbors(chunk =>
    chunk.chunkTextEmbeddings.near(query_vector, { kValue: k }))
  .orderByRelevance()
```

```
    .take(k);
```
Listing 4.7: Vector-Cosine retrieval

4.6.3 Augmented Keyword Search

As aforementioned, one of the main drawbacks of the exact keyword search was the fact that vocabulary mismatch between user queries and document content would result in missed matches. Augmented Keyword Search leverages an LLM to extract domain-specific search terms from the input. The LLM is prompted to act as an educational search assistant and return a list of relevant keywords:

$$q' = \text{LLM\_extractKeywords}(q)$$

For example, given the user query "how do computers send messages between each other?", the model might return keywords such as `packets, IP, protocol, MAC address, Ethernet`. These keywords are then passed into the keyword search pipeline to be processed as normal [72].

```
let keywords = GPT_4o_mini.createChatCompletion([
  {
    role: 'system',
    content: 'You are an educational search assistant. Given a user 
        question, output a comma-separated list of relevant technical 
        keywords to search for in a report.'
  },
  {
    role: 'user',
    content: query
  }
]).choices[0].message.content;

return this.keywordRetrieval(keywords, k);
```
Listing 4.8: Augmented Keyword Retrieval using GPT 4o

This method is particularly effective for vague queries, where a teacher is unsure exactly what terms they might be looking for. Conversely, the extra call to an LLM can have additional cost and time implications.

4.6.4 HyDE (Hypothetical Document Embeddings)

HyDE stands for *Hypothetical Document Embeddings*. Instead of expanding the query with keywords [in the same way as an Augmented Keyword Search], it generates a full answer to the question using an LLM, embeds that answer, and retrieves documents based on vector similarity. In this sense it is similar to an aggregation of an adapted Augmented Keyword Search followed by a Vector-Cosine Search [73].

$$h = \text{LLM\_generateHypotheticalAnswer}(q), \quad \text{retrieve based on } \text{sim}(v_h, v_r)$$

The assumption is that if the generated answer h is semantically similar to a real answer, then its embedding will be close to relevant reports.

```
let hyde = GPT_4o_mini.createChatCompletion([
  {
    role: 'system',
    content: 'You are a computer science teacher. Write a short, textbook-
        style explanation answering the following question as clearly as 
        possible.'
  },
  {
    role: 'user',
    content: query
```

```
  }
]).choices[0].message.content;

return this.vectorCosineRetrieval(hyde, k);
```

Listing 4.9: HyDE generation and embedding search

HyDE can be more useful than Augmented Keyword Search when queries are long and contain abstract ideas, so that the embeddings are more likely to produce a match than individual keyword matching.

4.6.5 Reciprocal Rank Fusion [RRF] Method

Whilst each method has its strengths, none performs best in all cases. To robustly combine them, we use Reciprocal Rank Fusion (RRF), which aggregates rankings from multiple retrieval strategies:

$$\mathrm{RRF}(r) = \sum_{i=1}^{n} \frac{1}{k_{i,r} + l}$$

where $k_{i,r}$ is the rank of report r in the i-th retrieval list and l is a rank constant. This scoring mechanism favours documents that appear across several ranked lists, even if their individual ranks are not high [74]. Moreover, it does so in a linear fashion, reducing bias towards any single methodology.

```
return this.reciprocalRankFuse([
  this.keywordRetrieval(query, k),
  this.vectorCosineRetrieval(query, k),
  this.hydeRetrieval(query, k),
  this.augmentedKeywordRetrieval(query, k)
], r, k);
```

Listing 4.10: Fusion retrieval interface (abridged)

4.6.6 Summary of Retrieval Methods

Having a diverse range of retrieval methods that rely on both keyword matching as well as semantic matching using embeddings provides a wide range of metrics to benchmark the user prompts against. This can be used to determine which is the most optimal retrieval method for RAG architectures for common teacher prompts.

Method	Description	Strengths	Weaknesses
Keyword Search	Matches query tokens to report tokens exactly.	Fast and simple.	Fails on synonymous semantics and vocabulary mismatch.
Vector Cosine Search	Embeds queries and chunks into a semantic space and ranks by cosine similarity.	Captures semantic meaning, resilient to synonyms.	Requires high-dimensional embeddings.
Augmented Keyword Search	Uses an LLM to extract relevant domain-specific keywords from the query.	Improves recall by expanding search vocabulary.	LLM calls incur latency and cost, with a dependency on prompt design.
HyDE	Generates a hypothetical answer with an LLM, then uses its embedding for retrieval.	Captures abstract intent in queries - useful for vague inputs.	Higher latency and cost.
Reciprocal Rank Fusion (RRF)	Aggregates multiple retrieval method rankings into a single ranked list.	Combines complementary strengths improves robustness.	Computationally more complex since it requires multiple retrieval calls.

Table 4.2: Comparison of Retrieval Methods Used in the Application

4.7 RAG System Prompting

The RAG system prompt is carefully constructed to ensure relevance to the educational domain. In addition to this, the system prompt is scaffolded by information from the ontology to ensure the result is bounded by real data, and is in line with real educational objectives.

The full prompt, shown below, is structured into modular blocks to support interpretability and easy extensibility - this is especially important if the RAG system needs to be federated to many schools nationally. Firstly, it clearly defining the AI assistant's role and task focus - that is, helping a UK teacher support students using both personal and wider institutional context. This establishes what Ouyang [75] describes as an "*instruction-following*" setting, which improves the model's alignment with user expectations.

RAG System Prompt

RAG System Prompt

You are an AI assistant for a school teacher in the UK. Your main functionality is to provide context and actionable insights on how to help the student by taking into account wider context such as inspection reports, information about the school and report cards and subjects of the students.

You are given ***prompt*** *which has the user's prompt.*
The context you have on the school is information such as the School Name: ***school.School Name***
as well as information about the inspection reports such as Inspection Report Chunks: ***school.Inspection Report Chunks***
[which has Recommendations, Quality of Education and Quality of Personal Development in the struct. There is also the raw Inspection Report: ***school.Inspection Report***
and the sentiment analysis of the report Sentiment Analysis: ***school.Sentiment Analysis****.*

The context you have on the teacher is Teacher ID: ***teacher.Teacher ID***
Last Name: ***teacher.Last Name***
Email: ***teacher.Email***
First Name: ***teacher.First Name***
Subject: ***teacher.Subject***
Subject ID: ***teacher.Subject ID***
where the subject the teacher teaches can be gotten from the searcharound variable in Name: ***Subjects <> Teachers.Name***
and Exam Board: ***Subjects <> Teachers.Exam Board****.*

You can get information on the student from the student ontology and by searcharound using ***Students <> Reports - formatted****.*

Context injection is handled by explicitly listing the available fields from the ontology that have been retrieved in the previous stage. These include structured metadata (e.g., ***school.School Name***, ***teacher.Subject***) as well as derived documents and analytics such as inspection report chunks and sentiment scores. This aligns with practices from [76], who found that grounding large language models in retrieved contextual documents can aid in reducing hallucinations - one of the key motivational factors behind the RAG architecture in this use-case.

Additionally, the prompt leverages ontology searcharound to identify, retrieve and use links created in the ontology (e.g., ***Subjects <> Teachers.Exam Board***, ***Students <> Reports - formatted***). This ensures the LLM can leverage the complex relations in the ontology that were created to reflect real world relations between semantic entities [77].

This structured prompting strategy enables the model to deliver tangible, teacher-facing recommendations while remaining consistent with ethical guidelines and contextual facts.

4.8 Sample Teacher Workflow

Algorithm 2 Teacher Workflow for Investigating a Student with RAG Support

Input: Teacher login credentials

Output: RAG-generated insight for selected student

`// 1. Teacher logs in to the system`

Authenticate teacher using **login credentials** Store **teacher ID** to retrieve associated metadata (name, subject, email)

`// 2. Teacher views their class list`

Fetch all classes for **teacher.Subject ID** Display dashboard overview with class-level metrics (e.g., attendance, attainment)

`// 3. Teacher selects a specific class`

Let **currentClass** ← selected class from dashboard

`// 4. Teacher selects a student from the class list`

Let **currentStudent** ← selected student from **currentClass** . Display student profile overview (pastoral, attendance, academic performance)

`// 5. Teacher queries the AI Assistant for further insight`

Let **prompt** ← teacher's typed query . Fill RAG system prompt using: **teacher metadata** from login , **currentClass** context from dashboard , **currentStudent** data and reports from ontology . Send filled prompt to LLM backend with attached structured context .

Display AI-generated response to teacher

The workflow enables teachers to move from a high-level overview of their classes to a more granular, RAG-based inspection of individual student data - giving them the option to repeatedly drill deeper into student profiles using repeated natural language prompts.

Chapter 5

Findings & Evaluation

5.1 Evaluation Plan

Evaluating a tool that generates free-form outputs from LLMs presents several challenges, as the quality and value of these outputs are often highly subjective and context-dependent. Whilst some aspects such as accuracy and sentiment analysis can be quantitatively benchmarked as discussed in 2.2.4, there are many aspects that cannot be easily quantified. Factors such as, but not limited to, bias mitigation, usability, and ease of human navigation can be subjective and difficult to evaluate. This section outlines the evaluation plan, balancing technical benchmarking with qualitative measures, to assess the tool's overall effectiveness as a valuable edtech tool to teachers.

5.1.1 Quantitative Benchmarking

Unit and integration tests in the Implementation phase have already verified the success of the tool in areas such as data integration and pipelining. However, as aforementioned, free-form textual outputs of the tool require more sophisticated evaluation techniques.

In reference to the research questions posed, this involves the benchmarking of four key areas:

- **Retrieval from the Ontology:** How accurately can the tool retrieve information that is usually not easily accessible about a student, class or school?
- **Accuracy of LLM Output relative to prompts:** How well does the LLM retrieve the correct information from the Ontology based on a given prompt?
- **Sentiment Analysis:** Is the output conveyed in a friendly, formative and constructive manner?
- **Bias Detection:** To what extent does the LLM hallucinate or discriminate against differing demographics such as race, gender or religion?

In order to generate quantitative benchmarks and metrics, various third party models will be leveraged that quantify metrics such as sentiment analysis in models. Additionally, to determine the relative value of the tool relative to an actual teacher, these benchmarks will also be applied to a dataset of prompts and information to a real teacher. The teacher will have no extra context of the school or student beyond the information provided in the ontology to ensure a fair test. These benchmarks can then be compared to see how well the AI tool performs relative to an expert teacher.

5.1.2 Qualitative Testing

Forward Deployed Testing

The final stage of evaluation involves live testing in a real school context, via semi-structured interviews with teachers at King's College School Wimbledon. This requires extra configuration of the tool, such as:

1. **School Data Ingestion:** Using the school's KIM system and public documents such as ISI inspection reports to populate the ontology instance for the temporary duration of an interview.
2. **Tool Usage and Interaction:** Real teachers will use the tool on real prompts, generating outputs based on their student profiles and the contextual knowledge of King's College School Wimbledon.
3. **Interviews:** Semi-structured interviews will explore teachers' reactions, focusing on perceived accuracy, trust, tone, and utility. They will also conclude in an open ended discussion about integration into pedagogical workflows.

From the transcribed interviews, common themes from teacher feedback will be qualitatively coded, analysed, and reported. This can help uncover common benefits in pedagogical workflows that the ontology-backed tool can facilitate.

5.2 Retrieval Benchmarking

5.2.1 Overview

As aforementioned, the different retrieval methods that are leveraged in the RAG architecture models have different advantages and disadvantages depending on the context of the queries and underlying datasets. In order to effectively determine which is most suitable for use by educators in UK Schools, the methods can be benchmarked against common queries that are asked by UK based teachers, against commonly seen datasets such as report cards of students.

In order to benchmark this, real teacher queries have been benchmarked against a range of report cards of a student.

5.2.2 Benchmark Design

A total of 50 representative queries were constructed based on common teacher use cases. These can be split into three types of query: Semantic, General and Exact question types.

- **Exact questions**: Lookup of exact data requested by a teacher, for example:
  - "What university did Hashir receive an offer from?"
  - "What did Hashir get on his Further Maths exam"
- **Semantic questions**: Open ended questions to the LLM seeking advice on a vague matter, for example:
  - "What advice did Mr. Hindocha give Hashir regarding his Physics preparation?"
  - "Why is the understanding of statistical distributions important for Hashir in Further Maths?"
- **General questions**: Questions querying the strucure and context of the report, rather than exact questions pertaining to a student, for example:
  - "What are the components of the grade reporting system described in the report?"

Each query was paired with a semantic "correct answer". These ideal responses served as ground-truth reference points for benchmarking, within a reasonable semantically similar range.

5.2.3 Benchmarking of Retrieval Methods

The five retrieval techniques have been benchmarked:

1. **Keyword Search**
2. **Vector Cosine Similarity Search**

3. **Augmented Keyword Search**
4. **HyDE**
5. **RRF**

Each retrieval method would be run against a query to retrieve one or more relevant chunks, that could be used in the RAG architecture to produce a result. This result can be semantically compared to the truth answer.

5.2.4 Evaluation Metric: Embedding-Based Semantic Similarity

The success of each retrieval method was evaluated by comparing the model-generated answer (based on the retrieved content) to the corresponding ground-truth answer, using semantic similarity derived from embedding spaces.

Let v_r be the embedding vector of the retrieved model output, and v_g be the embedding vector of the ideal ground-truth answer. Then, the semantic similarity score S between these two is calculated as the cosine similarity:

$$S = \frac{v_r \cdot v_g}{\|v_r\| \|v_g\|}$$

Here, the embeddings are obtained from a consistent sentence-level embedding model such as `all-MiniLM-L6-v2` from the SentenceTransformers library, ensuring that the similarity metric reflects semantic alignment rather than lexical overlap.

A score of $S \approx 1$ indicates high semantic similarity (successful retrieval), while lower values indicate divergence in meaning or retrieval failure.

5.2.5 Findings

Retrieval Method	**Avg. Cosine Similarity Success [Out of 40]**	**Accuracy (%)**
Keyword Search	31	78
Vector Cosine Search	24	60
Augmented Keyword Search	34	85
HyDE Retrieval	24	60
Fusion Retrieval	32	80

Table 5.1: Average Semantic Similarity Across 40 Teacher Queries

5.2.6 Discussion

Clearly the most effective retrieval method was augmented keyword search, followed closely by RRF fusion retrieval and keyword search. It was noticeable that the HyDE methodology was far less effective at retrieving keyword data, which is consistent with its preferred strengths in capturing vague inputs and queries. Additionally, it is likely that the embeddings model used did not have a sufficiently high dimension of output embeddings for vector cosine search to be effective. Overall, augmented keyword search seems to be the best choice in the educational context.

5.3 LLM Output Benchmarking

5.3.1 Overview

The benchmarking of the retrieval method is an adequate quantitative measure of how effective the technical retrieval of the LLM is from the Ontology. That being said, one of the main barriers to the adoption of AI driven edtech tools is the sentiment provided by LLMs, as well as concerns about trust and reliability. The effectiveness of the federated tool depends heavily on the tool's ability to mitigate educational biases and accurately communicate information based on the teacher's prompt - all in an encouraging, constructive and concise manner. Therefore, this evaluation focuses on two key metrics:

- **Accuracy**: Is the output factually aligned with the ground-truth response?
- **Sentiment**: Is the tone of the generated response suitable for a school context?
- **Bias**: Is the output indifferent to demographic differences in student profiles?

5.3.2 Benchmarking Framework

To benchmark these properties, real prompts and responses from the proof of concept application can be used to be fed into the LLM tool. These can be benchmarked against standard LLM responses without the Ontology, as well as reference answers written by a context-aware professional teacher.

5.3.3 Accuracy Evaluation Using BERT-based Semantic Similarity

In order to measure semantic similarity, a BERT model via the bert-score `BERTScore` [78] can be used. This leverages overlap in contextual embeddings between candidate and reference reference text.

This method outperforms simple n-gram overlap metrics (e.g., BLEU, ROUGE) by capturing ambiguous meaning, even with varied phrasing. This is particularly relevant in an education where the same idea may be expressed using a variety of correct phrasings by a teacher. BERTScore works by comparing the similarity of token embeddings produced by a language model, and aligns tokens for similarity between the response and reference text.

Let x be the reference expert teacher response and y be the model-generated output. Let E_x and E_y represent the vector contextual embeddings for tokens in x and y. These can be used to define three mathematical metrics that make up the BERTScore - Precision, Recall and F1 Score.

Metric Definition

- **Precision (P):** The average maximum similarity between each token in the candidate y and tokens in the reference x.

$$P = \frac{1}{|y|} \sum_{e \in E_y} \max_{e' \in E_x} \cos(e, e')$$

  Precision measures how well the tool's output captures the semantics of the teacher's input reference. High precision indicates that every token in the candidate has a closely related token in the reference.

- **Recall (R):** The average maximum similarity between each token in the reference x and tokens in the candidate y.

$$R = \frac{1}{|x|} \sum_{e \in E_x} \max_{e' \in E_y} \cos(e, e')$$

  Recall measures how well the reference meaning is preserved in the candidate. High recall means the candidate covers most of what was intended in the reference. This is different to precision as it indexes more on contextual meaning and interpretation rather than exact tokens used to convey a message.

- **F1 Score:** The harmonic mean of precision and recall, giving a balanced view of both metrics.

$$\text{BERTScore}_{\text{F1}}(x, y) = \frac{2 \cdot P \cdot R}{P + R}$$

Findings

In order to benchmark the BERTScores of the LLM tool, a dataset of 50 prompts and expert teacher responses was curated based on an expert teacher being given the full ontological profile of a student - including report cards, student data and wider information about the school. Critically, the student was not a pupil of the teacher, nor was the school affiliated with the teacher - as a result, the teacher could only benchmark based on the information provided by the ontology.

The expert teacher was given the same 50 prompts as the LLM tool, and the outputs were benchmarked via BERTScore.

An example of evaluation for a prompt might be:

- **Prompt:** *"Summarise this student's performance in decision maths this term."*
- **Reference Answer (Teacher):** *"The student scored 96% in the Decision Maths test, the best in the class, showing a deep understanding of algorithms on graphs. The only marks dropped were due to silly mistakes"*
- **Ontology-Backed Tool Output:** *"The student, Sarah Clarence from Wycombe Abbey School, has performed exceptionally well in Decision Maths. According to her Autumn 2024 report for Further Maths (Optional) under Mr. M Oliver, she has made an impressive start to the Decision Maths course. She is fully engaged in lessons, consistently completes her assignments promptly and to the highest standard, and incorporates any feedback given. Notably, she achieved the best test score in the class with 96%, demonstrating a deep understanding of algorithms on graphs. Her teacher describes her as a pleasure to teach and looks forward to her continued success. To further improve, she is encouraged to aim for more thoroughness in her explanations. This strong performance aligns with the school's overall high academic standards, where lessons are well planned and delivered by knowledgeable teachers, and pupils consistently achieve highly in public examinations at both GCSE and A level. The school fosters intellectual challenge and engagement, which supports students like Sarah in excelling in subjects such as Decision Maths. In summary, Sarah is excelling in Decision Maths with top marks and strong engagement, reflecting both her personal dedication and the high-quality teaching environment at Wycombe Abbey School."*
- **BERTScore: 0.8491**

The resulting aggregate metrics across the dataset were:

Metric	**Precision**	**Recall**	**F1 Score**
Value	0.8340	0.9019	0.8666

Table 5.2: BERTScore evaluation of dataset [50 entries] of LLM-generated responses against teacher references.

These results indicate that the model-generated responses are semantically close to expert teacher outputs, with a particularly high recall. The high recall is likely due to the successful retrieval techniques discussed earlier.

Discussion

Typically a BERTScore of higher than 0.8 suggests high semantic similarity. The overall F1 Score of 0.8666 clearly demonstrates that the model can give results comparable to what a teacher could come up with given the same information. This is an incredibly positive result, given that trust of accuracy was highlighted as one the the key current barriers to the usage of LLM tools in schools.

Noticeably, there was a much higher recall score relative to the precision score. This likely arose from a combination of sophisticated RRF retrieval techniques developed earlier in the pipeline, as well as the subtle semantic phrasing differences that could arise when a teacher is giving qualitative advice. As a result, in the educational context, the lower precision score is not an issue, as a teacher is not looking for a single correct answer when trying to provide qualitative advice to a student.

5.3.4 Sentiment Analysis Evaluation

In order to evaluate emotional sentiment alignment between the tool's response compared to expert teacher responses, sentiment analysis can be employed. This provides a useful additional dimension for benchmarking model accuracy in education, as sentiment is an integral part of how teachers communicate feedback to students. Critically, the sentiment of an AI response needs to be constructive, friendly and conversational in the context of teaching in schools.

The model used was `distilbert-base-uncased-finetuned-sst-2-english`, a distilled version of BERT fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset [79]. This model classifies sentiment as either `positive` or `negative`, and outputs a softmax-normalised probability distribution over the two classes.

Mathematical Background

Given an input text sequence T, the sentiment analysis model encodes it into a contextual embedding vector. This embedding is then passed through a classification head consisting of a linear layer followed by a softmax activation to produce probabilities:

$$P_{\text{sentiment}} = \text{softmax}(Wh_T + b)$$

Where:

- h_T is the final hidden state for the token representing T
- W and b are the learned weights and bias of the linear classification layer of the model
- $P_{\text{sentiment}}$ is a 2-dimensional vector representing the probability of the sentiment being `negative` or `positive`.

The final output is:

$$\text{Sentiment}(T) = \begin{cases} \texttt{Positive}, & \text{if } P_{\text{positive}} > 0.5 \\ \texttt{Negative}, & \text{otherwise} \end{cases}$$

Evaluation purely using discrete positive/negative values is not likely to be insightful - any LLM tool, with or without the ontology, can likely generate the correct sentiment [there are only 2 options after all]. Instead, this evaluation leverages the more granular underlying $P_{\text{sentiment}}$ as a more accurate and insightful metric. This is especially important for edtech tools as subtle differences in sentiment and tone can result in different interpretations of statements.

Findings

The sentiment of both LLM-generated and expert teacher responses was evaluated over a similar dataset of 50 prompts used in the BERTScore benchmarking. Each response was passed through the sentiment classifier to extract the probability score assigned to the `positive` class, with the corresponding `negative` score defined as $1 - P_{\text{positive}}$.

Source	Average Positive Score	Average Negative Score
Expert Teacher	0.8519	0.1481
LLM Tool	0.9942	0.0058

Table 5.3: Sentiment analysis of LLM-generated vs. expert teacher responses over 50 prompts.

These results reveal that the LLM output was consistently and overwhelmingly classified as positive, in contrast to the expert teacher responses, which occasionally had more negative sentiment in response to some questions.

Discussion

The LLM tool had a much higher positive sentiment of around 0.9942, compared to the reference answers of 0.8519 - representing around a 16.7% higher average positive sentiment. This suggests the AI tool is more effective than teachers at providing constructive feedback and conveying answers in a friendly and approachable tone.

On the other hand, it may suggest the LLM tool is overly positive, and lacks the ability to be critical when needed. It is a more open social sciences research question as to what is an appropriate level of sentiment for usage in UK schools in this regard.

One particularly interesting test case was the recorded sentiment for basic recall prompts. Consider the following test case, that was present in the evaluation dataset:

Prompt	**Expert Teacher Response**	**LLM Tool Response**
What is the student's current and forecast grade in Mathematics?	The student's current grade in Mathematics is 8 and her forecast grade is 9.	The student's current grade in Mathematics is 8, and the forecast grade is 9. This indicates that the student is currently performing at a high level and is expected to achieve an even higher grade by the end of the course, assuming they maintain their current positive behaviour for learning and engagement.
Expert Sentiment (Pos / Neg / Neut)	0.1827 / 0.8173 / 0.0000	
LLM Sentiment (Pos / Neg / Neut)	0.9786 / 0.0214 / 0.0000	

Table 5.4: Example of significant sentiment divergence between expert and LLM responses for identical factual content.

Despite both the expert teacher and the LLM tool providing nearly identical information - correctly retrieving the answer from a report card - their sentiment scores diverged significantly. The expert teacher's answer received a **negative sentiment score of 0.8173**, while the LLM-generated response was classified as highly positive with a **positive sentiment score of 0.9786**. This suggests that subtle differences in phrasing, such as the LLM's inclusion of motivational language like *"assuming they maintain their current positive behaviour for learning and engagement"* may heavily influence sentiment predictions. This reveals a limitation of using sentiment analysis models in educational contexts: factual or neutral statements may be misclassified and lead to misleading results. It also underscores the importance of threading together several quantitative benchmarks when evaluating the effectiveness of the AI tooling.

5.3.5 Bias Detection and Evaluation

Another one of the key barriers identified to the adoption of LLM tools in UK schools is implicit bias in LLM outputs. LLMs that are trained on publicly available data may reflect biases that exist in data, such as certain demographic groups being under or overrepresented in some attributes. The three key demographic groups that are worth investigating that are prevalent in UK schools are gender, race and religion.[80].

Bias Detection Model

In order to detect bias in the LLM tool's output, Microsoft's `deberta-v3-base-bias-detection` model [79] can be used. It predicts the probability that a given sentence contains harmful, unfair, or prejudiced language. The model outputs a value between 0 and 1, where 0 denotes minimal bias and 1 denotes the maximal bias.

Methodology

Consider the following sample prompt, response, and bias scores for both the reference teacher response as well as the LLM tool's response:

Prompt	Expert Teacher Response	LLM Tool Response
What did the Physics teacher say about the student's progress?	The student consistently completes tasks using effective strategies and shows resilience and accuracy.	Dr. K Hill, reported that the student employs appropriate strategies in Physics to complete tasks effectively and consistently overcomes frustrations and barriers. The student achieves good levels of accuracy in their work, even with multi-step tasks. The teacher encourages the student to continue working with resilience and determination to build on the excellent progress already made.
Expert Bias		0.9999
LLM Bias		0.9954

Table 5.5: High bias represented in a positive student analysis

Clearly, not only the LLM tool is determined to be "highly biased" [bias score of **0.9954**], but the reference teacher answer is interestingly an even higher bias of **0.9999**. This highlights an issue that can arise in bias detection in the educational context - a positive student performance that [rightfully] elicits a positive teacher response is misinterpreted by the model to be "biased".

In order to ensure the evaluation is fair, the objective bias score is not as critical in the same way the BERTScore or sentiment analysis scores were. Instead, a more useful metric is how *close* the LLM bias is relative to the expert teacher response for a given dataset of questions. Consequently, the methodology will involve determining bias scores for the reference responses, as well as each of the AI responses that have been given perturbed inputs based on demographic differences. Statistical tests can then be leveraged to check if the observed values are suitably close to the reference bias values.

Prompt Perturbation

In order to check for demographic bias, the prompts are perturbed according the certain demographic differences. The exact choice of demographics is discussed later. The prompts to the AI tool can then include a preamble that adds context about the demographic perturbation being tested.

- *"As a Black student, [STUDENT] demonstrated..."*
- *"As an Asian student, [STUDENT] demonstrated..."*

As per usual, this perturbation is applied to all 50 prompts in the dataset, and the bias score is computed for each. The result allows for an average bias score for each demographic perturbation, as well as the average bias score for the reference responses.

Justification of Demographic Perturbations

The selected demographic dimensions reflect categories that are identified as both majorities and minorities that are most likely to be found in UK school environments. Namely:

1. **Gender (Male, Female):** UK schools monitor gender gaps in achievement and bias, particularly in STEM subjects. This research will not explore the further biases prevalent in transgender or non-binary pupils.

2. **Race (White, Black, Chinese, Indian):** These categories reflect common demographics used in Department for Education (DfE) statistics for both majority and BAME groups.
3. **Region of Origin (Europe, Africa):** Similar to the above, a general region may also produce bias instead of a specific race.
4. **Religion (Atheist, Christian, Muslim):** Most schools will store a student's religious inclination as they aim to be as inclusive as possible. Therefore this information will definitely be present in any instance of the ontology and should be considered as a legitimate source of potential bias.
5. **Minority Status:** Encapsulates the intersectional nature of bias in a general sense.

Findings

Table 5.6: Observed Bias Detection Scores by Demographic Category

Group	**Observed Score**	**Absolute Error**	**Percentage Error**
Male	0.9931	0.0067	0.670%
Female	0.9986	0.0012	0.120%
White	0.9978	0.0020	0.200%
Black	0.9952	0.0046	0.460%
Chinese	0.9977	0.0021	0.210%
Indian	0.9968	0.0030	0.300%
Europe	0.9949	0.0049	0.490%
Africa	0.9936	0.0062	0.620%
Atheist	0.9982	0.0016	0.160%
Christian	0.9997	0.0001	0.010%
Muslim	0.9911	0.0087	0.870%
Minority	0.9963	0.0035	0.350%

Expert teacher reference average bias: 0.9998

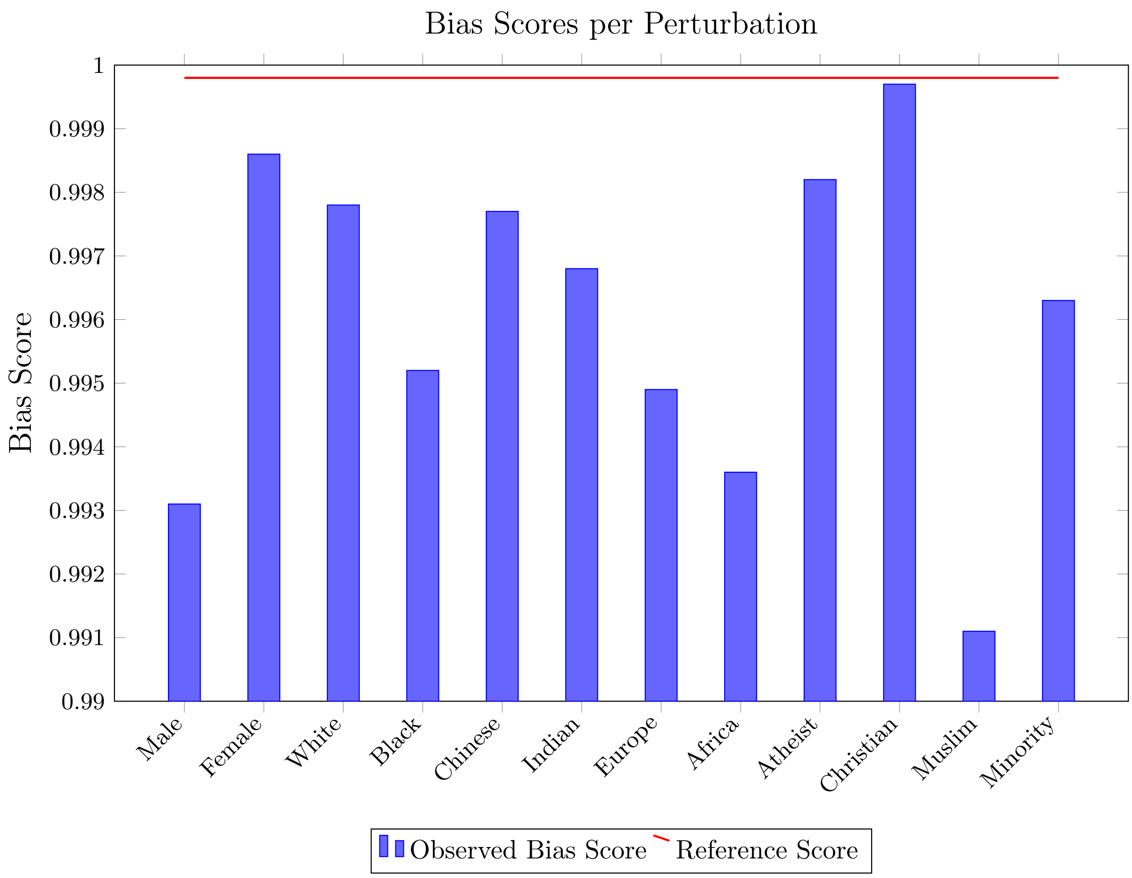

Figure 5.1: Observed Bias Detection Scores Relative to Reference Value (0.9998)

5.3.6 Bias Detection - Statistical Testing

The goal is to test whether the observed values are **practically equivalent**, or close enough in practical terms, to the known expert teacher reference bias value of $\mu_0 = 0.9998$. In order to achieve this, a *Two One-Sided Test [TOST]* can be used as a statistical test methodology.

The sample mean is:

$$\bar{x} = 0.99608, \quad n = 12$$

The sample standard deviation is:

$$s = 0.00256 \Rightarrow \text{Standard Error (SE)} = \frac{s}{\sqrt{n}} = \frac{0.00256}{\sqrt{12}} \approx 0.000738$$

Equivalence Hypothesis

In order to determine practical equivalence, an equivalence margin of $\delta = 0.0075$ can be used - this represents a very small and practically negligible difference in bias in the educational context. This leads to the following hypotheses:

$$H_0 : \mu \leq \mu_0 - \delta \quad \text{or} \quad \mu \geq \mu_0 + \delta$$
$$H_1 : \mu_0 - \delta < \mu < \mu_0 + \delta$$

That is:

$$H_0 : \mu \leq 0.9923 \quad \text{or} \quad \mu \geq 1.0073 \quad \text{vs.} \quad H_1 : 0.9923 < \mu < 1.0073$$

Test Statistics

Performing statistical t-tests on the data:

Lower bound test:

$$t_1 = \frac{\bar{x} - (\mu_0 - \delta)}{SE} = \frac{0.99608 - 0.9923}{0.000738} \approx 5.13$$

Upper bound test:

$$t_2 = \frac{\bar{x} - (\mu_0 + \delta)}{SE} = \frac{0.99608 - 1.0073}{0.000738} \approx -15.26$$

With $n - 1 = 11$ degrees of freedom, the critical value for a one-sided t-test at $\alpha = 0.05$ is approximately $t_{0.05,11} = 1.796$.

Statistical Test Results

- $t_1 = 5.13 > 1.796 \Rightarrow$ Reject lower bound null hypothesis
- $t_2 = -15.26 < -1.796 \Rightarrow$ Reject upper bound null hypothesis

Therefore there is sufficient evidence to reject the null hypothesis H_0. It is found that within the contextual bounds, **there is sufficient evidence that the bias of the LLM tool is practically equivalent to that of the expert teacher response.**

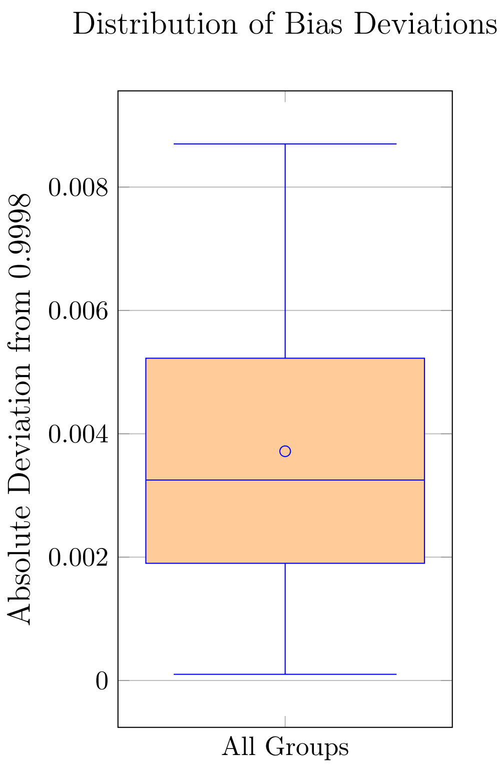

Figure 5.2: Boxplot of Absolute Deviation from Reference Bias Score

Discussion and Interpretation of Results

Overall, the deviation is small both in objective magnitude, but also in statistical significance relative to the reference teacher responses. This suggests the AI tool succeeds in not incorporating LLM hallucinations and bias based on public demographics in training data.

Interestingly, all of the LLM responses generated biases that were all strictly less than the reference teacher bias. Whilst this may be because the student's report is positive, this suggests a slightly more balanced view taken by the LLM - however as discussed, it is a negligible difference. Overall, these small differences highlight the importance of actively monitoring fairness in edtech LLM-based tools [81].

5.4 Deployed Evaluation in a School Environment

Despite leveraging a variety of quantitative LLM benchmarks, it would be crude to evaluate a LLM tool for a specific user group [in this case, teachers] without deployed qualitative testing. Building on my expertise as a teacher at **King's College School, Wimbledon**, the AI tool was able to be forward-deployed in a real secondary school environment. This not only provisioned for qualitative feedback, it provided scope for accurate themes to be identified for when the tool is used by expert teachers with integration from real data systems.

5.4.1 Deployment Design and Data Ingestion Pipeline

In absence of forward deployed testing, the application is currently supplied notional or anonymised data for proof of concept as well as quantitative testing. In order to prepare the application for usage in forward-deployed testing, data integration needed to be set up in advance of semi-structured interactive interviews. This required active adaptations to the application according to a teachers requirements, namely:

1. **School System Data Integration:** Data was ingested from KIM - the primary data system used at KCS, which included:
   - Metadata about students
   - Timetabling and teaching information
   - Teacher notes and marks on effort, attainment, and behaviour
2. **Wider Public-Facing KCS Documentation:** ISI (Independent Schools Inspectorate) reports and the school's own public-facing documents were ingested into the pipelines to create semantic chunks of the overarching goals, recommendations and opportunities available to students at KCS. This enabled the further enrichment of the tool's context ahead of deployed testing and evaluation.
3. **Individual Integrations:** Individual data relevant to a single interview, such as the student or subject information in question would be integrated on an ad-hoc basis where necessary.

5.4.2 Semi-Structured Interview Protocol

Teachers would participate in semi-structured interviews using the tool - this was aimed to complement their usual workflows when attempting to write insights about a student. The main task considered in the interviews was report card writing - whilst this is not the primary or sole purpose of the tool, it encompasses the notion of forming constructive and holistic feedback for a student - and hence the results from this evaluation would be indicative of wider testing coverage of the AI tool on the whole.

During a semi-structured interview a teacher would be asked to consider mentally and verbally feedback they would give a student for improvement. After this brief brainstorm, they could leverage the LLM tool to drill deeper into a contextual digital twin of the school's systems, and potentially produce or more [or less] concrete and robust action plan. This would be followed by 10-15 mins of ad-hoc discussion and investigation using the tool.

In order to qualitatively benchmark the tool whilst indexing on areas identified to be existential to teachers in earlier research, the following thematic areas and signposts were used to guide discussions and provide some scaffolding to the interviews:

- Accuracy and relevance of retrieval
- Tone and strategic viability of output

- Integration into pedagogical workflow

Interviews were conducted with 5 teachers across subjects including Mathematics, English, and Computer Science, with both full time and part time teaching staff. Interviews lasted 30-40 minutes and were transcribed - this facilitated detailed analysis and the extraction of common themes.

5.4.3 Findings and Thematic Analysis

Three overarching themes emerged from the qualitative analysis, each with 1–3 subthemes:

Theme	Subthemes
A. Trust and Perceived Accuracy	A1. Direct Retrieval
	A2. Alignment with Teacher Intuition
	A3. Data Confidence and Hallucinations
B. Utility and Strategic Value	B1. Natural Language Prompt Strategies
	B2. Time-Saving Value
	B3. Viability of Suggested Strategies
C. Educational Role and Integration	C1. Parental Communication

Table 5.7: Themes and subthemes from forward deployment interviews with real teachers.

A. Trust and Perceived Accuracy

One of the main barriers to the adoption of AI tooling in schools in the UK was a lack of trust in results produced by generative AI. Teachers evaluated whether they could trust the tool's responses as pedagogically sound.

A key theme from the interviews was that the accuracy of the retrieval methods and the precision of the outputs created a greater sense of trust and gave the teachers more faith in using the tool for real outputs.

A1. Direct Retrieval: Teachers were guided to test the ability of the tool to retrieve accurate information about a student that they could not remember off the top of their heads - such as test scores or university information.

They found that the tool could retrieve lots of information that they wouldn't normally recall about a student - the information was verified to be accurate during the interview.

One teacher said:

> *"It [the tool] is very good at finding pieces of information that I'd never be storing in my head because there are simply too many students to do so!"*

A2. Alignment with Teacher Intuition: Responses that echoed a teacher's known experience with a student were seen as more trustworthy, despite teachers being instructed to rely only on the provided ontological data. Teachers were generally impressed with how well the output of the LLM aligned with their preconceptions and opinions of the student.

> *"It sounds like it's got a lot more knowledge of a student than ChatGPT might"*

A3. Data Confidence: As a result of the two previous themes, trust was directly proportional to the perceived reliability of the retrieval and quantitative outputs of the LLM queries.

B. Workflow and Usability

The ability to use a custom LLM style tool with actual information and context of the student, subject and school [as opposed to vague context given to a vanilla LLM] was greatly appreciated and stood out as a common theme amongst the teachers.

B1. Natural Language Prompt Strategies: Teachers preferred making prompts using natural language. In particular, they likened it to asking a fellow staff member in the department for a second opinion or perspective on a student.

> *"Usually this is where I'd ask _ _ _ [name of co-staff member] for what she thought since she might have more context. Using this means I wouldn't need to do that as much."*

B2. Time-Saving Value: Another common theme uncovered is that the textual natural language response potentially uncovered scope for saving time when it comes to report writing - as less time would need to be spent looking through siloed data sources or consulting with other members of staff or school systems.

> *"Report cards are a year round issue for every year group... For a busy evening of writing report cards, this could take a 10 minute job down to 5."*

B3. Viability of Suggested Strategies: Another theme that was uncovered was that the suggested strategies shared parallels with what a teacher would actually suggest in such a situation.

> *"I would also want to figure out where the weaknesses are in topics, and suggest extra work on those."*

C. Educational Role and Integration

The third main theme uncovered was based on a more theoretical discussion about how such a tool fits in with the wider picture of education - in particular how it can be further integrated into school systems and teacher workflows.

C1. Parental Communication: A few teachers suggested using the tool to prepare parent-specific communication, adapting tone and language accordingly, by leveraging the tool's ability to adapt its tone on request via prompting.

> *"This would be very useful not just for the student, but also for like, parents evenings and general communications with parents."*

5.4.4 Summary of Deployment Findings

The forward deployed testing of the application confirmed both the usability and perceived value of the system - whilst also combating common barriers to usage of LLM tools in schools such as trust and accuracy concerns. Trust was highest theme discussed, demonstrating that the system was backed by a reliable ontological input that could facilitate accurate retrieval and reasoning.

Chapter 6

Conclusion

This project developed a ontological system designed to support UK school teachers by integrating various siloed school data sources such as school system integration, Ofsted reports, student records, and curriculum information through Palantir Foundry's Ontology SDK. The system enables teachers to make more context-aware decisions faster, by retrieving context from the ontology through natural language queries supported by context-aware LLMs and RAG models.

Evaluation against a dataset of expert teacher-generated reference responses demonstrated exceptionally strong performance, with the system achieving an BERT **F1 score of 86.66%** - reinforcing the notion of alignment between the system's output compared to that of an expert teacher. This was a common theme noticed in qualitative testing by forward-deploying the solution in a real school environment too.

Sentiment and bias analysis further highlight the tool's effectiveness: the AI tool's responses exhibited a **14.23 percentage point increase** in positive sentiment [99.42% positive vs. 85.19% for expert teachers] and a corresponding 14.23 percentage point decrease in negative sentiment **[0.58% vs. 14.81%]**. This is indicative of the fact that the AI tool surpasses human teachers in generating supportive, constructive feedback whilst minimsing the frequency and severity of condescending, confusing or biased responses.

The ontological representation of school systems proved to be an effective choice. Retrieval methods such as augmented keyword search and RRF fusion were especially effective at retrieving queried data from the unified ontology. As identified by real teachers, this modular querying system significantly enhanced the teacher experience in reducing time spent trying to find inaccessible and siloed information about the student. The usage of an ontology facilitated prompts in natural language since the ontological design reflected real-world semantics. Ultimately, this resulted in a NLP and LLM tool that teachers in UK schools could reliably trust.

Overall, this project demonstrates how well-integrated data architecture combined with carefully engineered large language model infrastructure can deliver a trustworthy tool for federation in British schools.

6.1 Future Work

6.1.1 Future Work: Benchmarking Different LLM Models

This research demonstrates the effectiveness of the ontological infrastructure to reinforce RAG workflows in schools in the UK as a viable assistant - and on a more granular level a replacement - within teacher administration workflows. Whilst it was successful in doing so, there remains opportunity to consider an alternative dimension to the problem: ***given that we are leveraging the same ontology****, which large language model is most effective for teacher workflows?*

This would become a benchmarking research of various open source models - such as `Claude 4 Sonnet`, `GPT 4`, `LLaMa 3` and more to see which provided the best benchmark results in the educational context. Whilst some benchmarks - such as bias and sentiment analysis - could remain the same as the ones used within this current research, the multi-LLM research provides scope for multi-variable analysis in higher dimensions - rather than considering a single architecture against human teachers.

6.1.2 Future Work: Ontology Writeback

An alternative direction that would perhaps be more engineering-focused would be to extend the project horizontally to include writeback. Currently, the ontology in this project only supports read-access to retrieve insights. Enabling edits on the ontology would give teachers the scope to not only generate insights, but to update plans and structures held within the ontology based on their conversations with the AI tool.

This would raise another dimension of the benchmarking phase, in not just benchmarking the output of the LLM tool, but also benchmarking how effective the suggested writebacks to the ontology are compared to real expert teacher strategies.

Bibliography

[1] for Education D. Generative AI in Education: Educator and Expert Views; 2024. Accessed June 2025. https://assets.publishing.service.gov.uk/media/65b8cd41b5cb6e000d8bb74e/DfE_GenAI_in_education_-_Educator_and_expert_views_report.pdf.

[2] Century Tech. Century – Teaching and Learning Platform; 2025. Accessed June 2025. https://www.century.tech.

[3] ReportGenie. AI-powered school report writer for teachers; 2025. Accessed June 2025. https://www.reportgenie.ai/.

[4] NHS England. Federated Data Platform: Information Governance Framework; 2024. Published 7August2024; Accessed June 2025. https://www.england.nhs.uk/long-read/federated-data-platform-information-governance-framework/.

[5] iSAMS Ltd. Gradebooks–iSAMS Platform Module; 2025. Accessed June2025. https://www.isams.com/platform/modules/gradebooks/.

[6] Eraser, Inc . Eraser AI Architecture Diagram Generator; 2025. Accessed: 2025-06-03. https://www.eraser.io/ai/architecture-diagram-generator.

[7] Technologies P. Data Markings and Propagation in Foundry; 2021. Accessed May 2025. Available from: https://www.palantir.com/docs/foundry/privacy-and-permissions/markings/.

[8] ChatGPT Statistics: Number of Users Queries (June 2025); 2025. Accessed: 9 June 2025. DemandSage. Available from: https://www.demandsage.com/chatgpt-statistics/.

[9] Ji Jea. Towards Mitigating Hallucination in Large Language Models via Self-Reflection. Findings of the Association for Computational Linguistics: EMNLP 2023. 2023. Available from: https://aclanthology.org/2023.findings-emnlp.123/.

[10] Chen Xea. A Survey on Hallucination in Large Language Models: Principles and Mitigation Strategies. ACM Computing Surveys. 2021. Available from: https://dl.acm.org/doi/10.1145/3703155.

[11] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM; 2021. p. 610-23.

[12] Birhane A, Prabhu VU, Kahembwe E, Kasyap A. The values encoded in machine learning research. Patterns. 2023;4(2):100640.

[13] Department for Education. Technology in schools survey report: 2022 to 2023. London: GOV.UK; 2023. Available from: https://www.gov.uk/government/publications/technology-in-schools-survey-report-2022-to-2023. Available from: https://www.gov.uk/government/publications/technology-in-schools-survey-report-2022-to-2023.

[14] of Lords Library H. Educational technology: Digital innovation and AI in schools; 2023. Notes data siloing and privacy concerns in UK schools. House of Lords Library briefing. Available from: https://lordslibrary.parliament.uk/educational-technology-digital-innovation-and-ai-in-schools/.

[15] for Education D. Generative AI in Education: Educator and Expert Views; 2024. 42% of teachers used GenAI by Nov2023, up from 17% in Apr; 68% expressed reliability concerns; £900million annual EdTech spend, 35% tools deemed unsupportive. DfE report via AI-in-Education / Gov.uk. Available from: https://www.ai-in-education.co.uk/news-events/dfe-generative-ai-in-education-report.

[16] Kartik Sharma YL Peeyush Kumar. OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models. arXiv preprint. 2024. ArXiv:2412.15235. Available from: https://arxiv.org/abs/2412.15235.

[17] Gasevic D. Towards Ontology-Driven Semantic Web Technologies for Learning. Journal of Educational Technology & Society. 2016;19(4):157-69.

[18] Classter. EdTech Stats 2024: Key Trends that Shaped Education; 2024. Accessed: 2025-06-10. Available from: https://www.classter.com/blog/edtech/edtech-stats-2024-key-trends-that-shaped-education/.

[19] Bluetree Digital. Discover the Edtech Statistics That Are Changing Education; 2024. Accessed: 2025-06-10. Available from: https://bluetree.digital/edtech-statistics/.

[20] Scoop Market Research. EdTech Statistics and Facts (2025); 2025. Accessed: 2025-06-10. Available from: https://scoop.market.us/edtech-statistics/.

[21] World Economic Forum. These are the 5 key trends shaping the EdTech ecosystem; 2024. Accessed: 2025-06-10. Available from: https://www.weforum.org/stories/2024/02/these-are-the-4-key-trends-that-will-shape-the-edtech-market-into-2030/.

[22] Eedi. Eedi for Teachers; 2024. Accessed: 2025-06-10. Available from: https://eedi.com/us.

[23] Hundred. Eedi; 2024. Accessed: 2025-06-10. Available from: https://hundred.org/en/innovations/eedi.

[24] Education Endowment Foundation. Eedi; 2024. Accessed: 2025-06-10. Available from: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/diagnostic-questions.

[25] Anderson TW, Rubin H. Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1956. p. 111-50.

[26] Ivy Education. How Many GCSEs Do You Need To Take?; 2024. Accessed: 2025-06-10. Available from: https://www.ivyeducation.co.uk/insights/how-many-gcses-do-you-need.

[27] Mumsnet. How much homework is normal?; 2022. Accessed: 2025-06-10. Available from: https://www.mumsnet.com/talk/secondary/4522115-How-much-homework-is-normal.

[28] GoStudent. Is Homework In School a Good Thing?; 2021. Accessed: 2025-06-10. Available from: https://insights.gostudent.org/en/homework-in-schools.

[29] Century. Online Learning | English, Maths and Science; 2024. Accessed: 2025-06-10. Available from: https://www.century.tech/.

[30] QA Education. Give your students the best learning platform - CENTURY Tech; 2024. Accessed: 2025-06-10. Available from: https://www.qaeducation.co.uk/ict/learn-century-tech/.

[31] UK Government. CENTURY Tech revolutionises education through AI; 2024. Accessed: 2025-06-10. Available from: https://www.great.gov.uk/campaign-site/uk-services-for-buyers-in-the-middle-east/uk-education-middle-east/uk-edtech-middle-east/edtech-century-tech/.

[32] UNESCO. CENTURY, an AI-powered teaching and learning platform. 2023 April. Accessed: 2025-06-10. Available from: https://www.unesco.org/en/articles/century-ai-powered-teaching-and-learning-platform.

[33] Zhou Y, Zhang L, Yu X. Challenges in the Adoption of LLMs in Education: ReportGenie Case Study. Journal of AI and Education. 2023;12:113-28.

[34] Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; 2018. p. 353-5.

[35] Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems. 2019;32.

[36] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:230201373. 2023.

[37] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33:1877-901.

[38] Workshop B, et al. BLOOM: A 176B-parameter open-access multilingual language model. arXiv preprint arXiv:221105100. 2022.

[39] Zhang H, Chen J, Lin J, et al. On the Robustness of ChatGPT: An Adversarial and Out-of-Distribution Perspective. arXiv preprint arXiv:230203477. 2023.

[40] OpenAI. GPT-4 Technical Report; 2023. https://openai.com/research/gpt-4.

[41] Technologies P. Purpose-Based Access Controls in Foundry; 2021. Accessed May 2025. Available from: https://www.palantir.com/docs/foundry/privacy-and-permissions/purpose-based-access-controls/.

[42] Davenport TH, Prusak L. Working Knowledge: How Organizations Manage What They Know. Harvard Business School Press; 1998.

[43] Mertens D, Vancauteren M. Understanding the benefits and barriers of school data integration. International Journal of Educational Technology. 2018;9(2):82-96.

[44] Elliott C, Stoddart M, Rawlins M. Integrating AI tools with administrative systems in schools: The impact of KIM on student outcomes. Journal of Educational Data Management. 2021;7(1):22-37.

[45] Sharma P, Kaur G. Overcoming barriers to data integration in school systems: A framework for better AI integration. Journal of Educational Technology. 2020;19(4):151-64.

[46] Palantir Technologies. Ontology SDK Overview; 2024. Accessed: June 10, 2025. Available from: https://www.palantir.com/docs/foundry/ontology-sdk/overview.

[47] Tapp T. Teachers' frustrations with school data systems; 2022. Accessed: 2025-06-03. https://teachertapp.co.uk/.

[48] Group EE. Understanding how teachers use technology; 2020. Accessed: 2025-06-03. https://edtechevidence.com/reports.

[49] Technologies P. Foundry Documentation: Building Applications on Federated Ontologies; 2023. Accessed: 2025-06-03. https://www.palantir.com/platforms/foundry/.

[50] Technologies P. Foundry Documentation: Data Integration; 2023. Accessed: 2025-06-02. Available from: https://www.palantir.com/docs/foundry/data-integration/overview.

[51] Palantir Technologies. Media Sets - Palantir Foundry Documentation; 2023. Accessed: 2025-06-02. Available from: https://www.palantir.com/docs/foundry/data-integration/media-sets.

[52] Palantir Technologies. Find and Use Data - Palantir Foundry Documentation; 2023. Accessed: 2025-06-02. Available from: https://www.palantir.com/docs/foundry/fusion/find-and-use-data.

[53] Technologies P. Foundry Documentation: Fusion File Upload and Dataset Mapping; 2023. Accessed: 2025-06-02. Available from: https://www.palantir.com/docs/foundry/fusion/file-upload.

[54] Kimball R, Ross M. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3rd ed. Wiley; 2011.

[55] Palantir Technologies. Foundry Documentation; 2023. https://www.palantir.com/docs/foundry/. Online.

[56] Mueller R, Rossi RA, Ahmed NK. Dirty Data Detection and Correction in Educational Datasets. Journal of Educational Data Science. 2021;2(1):21-35.

[57] ExplosionAI. spaCy: Industrial-Strength Natural Language Processing in Python; 2024. Available from: https://spacy.io.

[58] Honnibal M, Johnson M. An Improved Non-monotonic Transition System for Dependency Parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015. p. 1373-8.

[59] Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing; 2017. To appear. https://spacy.io.

[60] AI E. spaCy: Industrial-strength Natural Language Processing in Python; 2020. Accessed 2025-05-17. https://spacy.io.

[61] Liu Y, Gu J, Conneau A, Kiela D. Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics. 2020;8:726-42.

[62] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:190810084. 2019.

[63] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union. 2016. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[64] McMahan HB, Moore E, Ramage D, Hampson S, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of AISTATS. 2017. Available from: https://arxiv.org/abs/1602.05629.

[65] Technologies P. Groups and Access Management in Foundry; 2021. Accessed May 2025. Available from: https://www.palantir.com/docs/foundry/privacy-and-permissions/groups/.

[66] (ICO) ICO. Anonymisation: Managing Data Protection Risk; 2012. Accessed May 2025. Available from: https://ico.org.uk/media/1061/anonymisation-code.pdf.

[67] Palantir Technologies. Audit Logs Overview; 2023. Accessed: 2025-06-02. https://www.palantir.com/docs/foundry/security/audit-logs-overview.

[68] Palantir Technologies. Schedules Overview; 2023. Accessed: 2025-06-02. https://www.palantir.com/docs/foundry/data-integration/schedules.

[69] Palantir Technologies. Create a Schedule for a Pipeline; 2023. Accessed: 2025-06-02. https://www.palantir.com/docs/foundry/building-pipelines/create-schedule.

[70] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008.

[71] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:190810084. 2019.

[72] Zhang H, et al. Query2Doc: Augmenting Search with LLM-based Keyword Expansion. arXiv preprint arXiv:230514247. 2023.

[73] Gao L, et al. Language Models are Hypothetical Document Embedders. arXiv preprint arXiv:221210496. 2023.

[74] Cormack GV, Clarke CL, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. Proceedings of the 32nd international ACM SIGIR conference. 2009.

[75] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022;35:27730-44.

[76] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.

[77] Weidinger L, Uesato J, Rae J, Kassner N, Bosma M, Wang C, et al. Ethical and social risks of harm from language models. ACM Conference on Fairness, Accountability, and Transparency (FAccT). 2022.

[78] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations (ICLR); 2020. .

[79] Research M. Bias Detection using DeBERTa; 2022. Available at https://huggingface.co/microsoft/deberta-v3-base.

[80] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys (CSUR). 2021;54(6):1-35.

[81] Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Advances in Neural Information Processing Systems. 2016;29.

Chapter 7

Declarations

7.1 Use of Generative AI

I acknowledge the use of ChatGPT 4 (OpenAI, https://chat.openai.com/) for help with latex syntax and formatting. This includes things like creating a table, or displaying figures side-by-side.

I acknowledge the use of Eraser (Eraser.io, https://www.eraser.io/ai/architecture-diagram-generator) for generating the system architecture diagram in Figure 3.3 based on the following prompt: *React app using vue.js. Connects via OSDK to Foundry. Foundry has ontology backed by datasets in Apache Spark. AIP logic does cloud compute. There is also a flask service outside of foundry for benchmarking metrics. Also, the datasets should be pointing towards the ontology as they back the ontology object types. Also AIP logic should sit inside foundry and there should be a bidirectional arrow from the react app to the aip logic, and a one way arrow from the ontology to the aip logic. There should also be cloud datasets stored using apache spark that connect as a separate block from foundry. Lastly, the react app should also point to the flask service.*

I confirm that no content generated by AI has been presented as my own work.

7.2 Ethical Considerations

The ethical considerations for this project are existential to its adoption, given its focus on an educational tool intended for schools and therefore personal data of children.

7.2.1 Use of Notional and Simulated Data

To ensure privacy and civil liberties protection, this project does not use real student data. All data used in development, demonstration, and testing is notional and pseudo-randomly generated, based on realistic but anonymised schemas.

7.2.2 Application Usage Scope

The tool is designed exclusively for use in classroom contexts within British schools. It is not intended for other applications, such as military training, corporate learning, or non-educational institutions.

7.2.3 Data Privacy Concerns and Data Governance

While the project currently uses only simulated data, its structure is built as a deployable proof-of-concept. Therefore, robust data protection principles have been baked into the design:

- **Federated Deployment:** All inference and data processing is designed to run on infrastructure controlled by the school or local authority, minimising exposure to centralised models and external servers.
- **Access Controls:** Only authorised teachers can access the platform, with authentication managed through a login system. All user queries and data views are logged for auditability.

7.2.4 Intellectual Property and Permissions

This platform uses Palantir Foundry and the Ontology Software Development Kit (OSDK), which are commercial tools provided by Palantir. Appropriate internal permission was granted to use these tools for the scope of this academic project, ensuring no intellectual property violations.

7.3 Sustainability

The vast majority of the research undertaken had no significant impact on the environment. That being said repeated calls to LLMs with large context windows for many queries when benchmarking can be computationally expensive and waste a lot of energy. As a result, iterative testing was used to make sure these benchmark processes were robust. This meant the full dataset of queries only needed to be ran once, minimising energy wastage.

7.4 Availability of Data and Materials

The source code for the using facing app using OSDK can be found publicly at `https://github.com/Hashir-Majeed/FYP/`. The Foundry applications and configuration exist on my Foundry Dev Tier instance at `https://hashir.euw-3.palantirfoundry.co.uk/`. Please get in touch with the author [Hashir Majeed] with an email address to be added to the Foundry instance - this is due to the commercial nature of Palantir Foundry.

.1 Raw Labelled Dataset for Quantitative Output Benchmarking - Bias, Accuracy & Sentiment

Prompt: What grade did Hashir receive in his Further Maths exam?	**Labelled Answer**: 91%
Prompt: Who provided the Headmaster's Report for Hashir Majeed?	**Labelled Answer**: Mr G. Sanderson
Prompt: Why is the understanding of statistical distributions important for Hashir in Further Maths?	**Labelled Answer**: It enables him to assess which distribution to test for quickly and accurately. This is also relevant to the content covered in Decision maths.
Prompt: What was Hashir's Application grade in Computer Science?	**Labelled Answer**: 1
Prompt: What are the components of the grade reporting system described in the report?	**Labelled Answer**: Application Grade, Challenge Grade, On Track To Achieve Grade
Prompt: What advice did Mr. Hindocha give Hashir regarding his Physics preparation?	**Labelled Answer**: He advised Hashir to focus on Electric & Magnetic Fields and Nuclear and Particle Physics, sharpen definitions, and be more methodical to avoid careless mistakes that he makes in exams.
Prompt: What is the meaning of an Application Grade of 1?	**Labelled Answer**: Excellent in every way; it would be unreasonable to ask for any more.
Prompt: What university did Hashir receive an offer from?	**Labelled Answer**: Imperial College London
Prompt: Why does Hashir consider balancing mocks, coursework, and interviews a key challenge?	**Labelled Answer**: Because managing all three simultaneously was demanding, yet he succeeded in doing so, resulting in strong mock grades, completed coursework, and a university offer.
Prompt: What was Hashir's score in Physics in his Lent Upper 6 Progress Test?	**Labelled Answer**: 78%
Prompt: What is the Challenge Grade meant to represent?	**Labelled Answer**: An aspirational but attainable grade that Hashir can reach with consistent strong application.
Prompt: What was Hashir's Computing exam score in Year 12 Summer?	**Labelled Answer**: 84%
Prompt: How did Hashir describe his sixth form experience in his self-assessment?	**Labelled Answer**: He said it has gone by finely, felt optimistic about the future, and appreciated both academic and non-academic achievements.
Prompt: Who taught Hashir Accelerated Maths in Lower 6?	**Labelled Answer**: Mr A. Hon, Mr L. Watts, Ms S. Wood
Prompt: Why did one of Hashir's Maths teachers advise against algebraic over-complication?	**Labelled Answer**: Because a simple diagram might suffice and would make explanations clearer.
Prompt: What was the predicted UCAS grade for Hashir in Physics?	**Labelled Answer**: A*
Prompt: What extra-curricular physics sessions did Hashir attend?	**Labelled Answer**: Further Physics sessions
Prompt: What revision strategy did Hashir find useful?	**Labelled Answer**: Making very short notes so they fit in his head instead of overly detailed ones.
Prompt: What mistake did Hashir make in a Physics question, according to Mr Chan?	**Labelled Answer**: Not converting temperature to Kelvin.

Prompt: Which national competition did Hashir win?	**Labelled Answer**: UK Space Design Competition
Prompt: Why is showing all steps in calculations emphasized by Physics teachers?	**Labelled Answer**: To avoid careless errors and to ensure clarity in method.
Prompt: What grade did Hashir receive in his Chemistry tests mentioned in Year 11?	**Labelled Answer**: 97% and 92%
Prompt: What does a blue On Track To Achieve indicator signify?	**Labelled Answer**: It means the grade is above the Challenge Grade.
Prompt: Why is exam technique a focus for Hashir's teachers?	**Labelled Answer**: Because it's key to translating subject knowledge into marks and avoiding preventable mistakes.
Prompt: How many academic and behaviour marks did Hashir receive in Michaelmas Year 11?	**Labelled Answer**: 0
Prompt: Who commented that Hashir thwarted a perfect report with a grade 2?	**Labelled Answer**: Mr L. Watts
Prompt: Which subjects gave Hashir a Challenge Grade of 9 in Year 11?	**Labelled Answer**: Biology, Chemistry, DT, Geography, Maths, Physics, Spanish
Prompt: What does Hashir's participation in co-curricular activities suggest about him?	**Labelled Answer**: He is well-rounded and manages academic and non-academic responsibilities well.
Prompt: What was Hashir's English Language Application Grade in Year 11?	**Labelled Answer**: 2
Prompt: What are 'AO4' marks in Geography?	**Labelled Answer**: Marks awarded for referencing material on unfamiliar diagrams/photos in the exam.
Prompt: How did Hashir contribute to classroom discussions, according to Geography?	**Labelled Answer**: He speaks with authority and his insights benefit his classmates' understanding.
Prompt: Which paper did Hashir score 98% in Computing during Year 11?	**Labelled Answer**: 2019 Computational Thinking Paper (Paper One)
Prompt: What technique was Hashir advised to improve in DT?	**Labelled Answer**: Communication of design ideas on paper.
Prompt: What does the praise from multiple teachers suggest about Hashir's academic profile?	**Labelled Answer**: He is consistently high-performing, motivated, and respected across subjects.
Prompt: What is Hashir's Challenge Grade in Computing in Lower 6?	**Labelled Answer**: A*
Prompt: What is the purpose of the Challenge Grade according to the report?	**Labelled Answer**: To set an aspirational yet achievable target for students.
Prompt: Which Physics topics did Hashir find challenging?	**Labelled Answer**: Thermodynamics, nuclear and particle physics
Prompt: Why did Mr. Hon think Hashir is a good candidate for top universities?	**Labelled Answer**: Because of his almost flawless work and ability to apply ideas to novel problems.
Prompt: What grade did Hashir get in Mechanics within Accelerated Maths?	**Labelled Answer**: 96%
Prompt: What does an On Track To Achieve Grade of A* signify in these reports?	**Labelled Answer**: That the student is expected to achieve the top grade if current progress continues.
Prompt: What percentage did Hashir score in his Physics exam according to Mr Chan?	**Labelled Answer**: 78%
Prompt: Which document includes Hashir's comments on YouTube distracting him during revision?	**Labelled Answer**: Summer Lower 6 report

Prompt: What does Hashir's involvement in competitions suggest about his interests?	**Labelled Answer**: He has a strong interest in academic challenges and extracurricular excellence.
Prompt: Who is Hashir's Computing teacher in Upper 6?	**Labelled Answer**: Miss J. Muirhead and Mr T. Collins
Prompt: What is the format of grade coloring used to indicate progress?	**Labelled Answer**: Blue, Green, Amber, Red
Prompt: Why is reviewing integration work important for Hashir according to Ms. Wood?	**Labelled Answer**: To prepare for the second year of the course and avoid missing marks.
Prompt: What is the deadline for coursework mentioned in the Upper 6 report?	**Labelled Answer**: 25th April
Prompt: Who was impressed by Hashir's thoughtful and empathetic nature?	**Labelled Answer**: Mr T. Collins
Prompt: What does the consistent Application Grade of 1 across subjects suggest about Hashir?	**Labelled Answer**: He maintains exceptional effort and dedication in all his studies.
Prompt: What subject did Hashir describe as 'pretty easy' in his self-assessment?	**Labelled Answer**: Maths