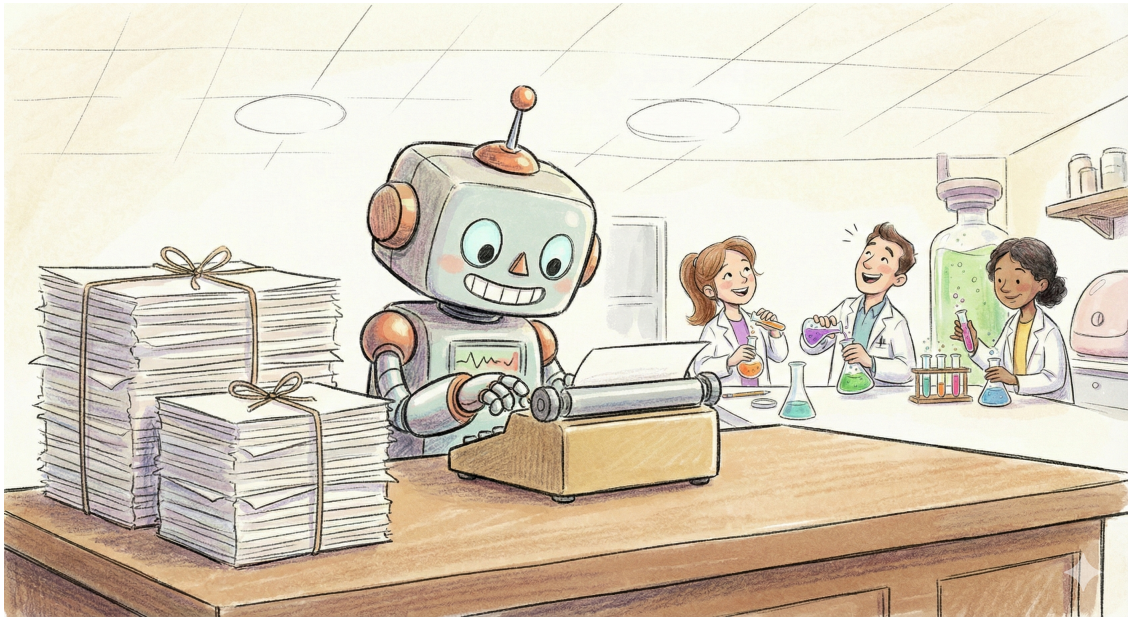
CHALMERS
UNIVERSITY OF TECHNOLOGY



Data-Driven Automated Reporting Solution for External Collaborations - LLM-driven KPI Definition

A Proof-of-Concept at AstraZeneca

Master's Thesis

Jakob Juul and Marcus Lorentzon

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's Thesis 2026

Data-Driven Automated Reporting Solution for External Collaborations - LLM-driven KPI Definition

A Proof-of-Concept at AstraZeneca

Jakob Juul: jakobjuu@chalmers.se

Marcus Lorentzon: marcuslo@chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden 2026

Data-Driven Automated Reporting Solution for External Collaborations - LLM-driven KPI Definition
A Proof-of-Concept at AstraZeneca
Jakob Juul
Marcus Lorentzon

Supervisor: Jesús Pineda, Department of Physics at Gothenburg University
Examiner: Giovanni Volpe, Department of Physics at Gothenburg University

Degree project report 2026
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Cover: A cartoon sketch of a robot AI processing text information from the pharmaceutical industry. Generated by Google Gemini's Nano Banana Pro model.

Typeset in LaTeX
Gothenburg, Sweden 2026

Data-Driven Automated Reporting Solution for External Collaborations - LLM-driven KPI Definition
A Proof-of-Concept at AstraZeneca
JAKOB JUUL and MARCUS LORENTZON
Department of Physics
Chalmers University of Technology

Abstract

This thesis presents a proof-of-concept, developed with AstraZeneca (AZ), that explores automating progress reporting for external collaborations by testing whether a large language model (LLM)-driven system can extract objectives from contracts and translate them into tailor-made key performance indicators (KPIs). Objective extraction is quite reliable, reaching several highs of accuracy around the 85%-mark, but converting objectives into KPIs that stakeholders judge as relevant, clear, actionable, and measurable, is substantially less solid. Fewer than half of the KPIs met each quality criterion on average, and 39% met none. Survey responses noted that KPIs were often unclear, overly generic, or poorly timed, and skewed toward simple counts (e.g., "number of models") that miss quality and impact.

From interviews conducted at AZ, a set of general KPIs, that were deemed meaningful to measure in a collaboration project, could be demonstrated. The final evaluation suggests that these KPIs (e.g., external engagement and budget coherence) outperform collaboration-specific KPIs generated directly from objectives. This underscores the difficulty of creating bespoke target measures in diverse contexts.

Despite these issues, the approach offers practical value. In principle, the pipeline should be better suited for agreements with explicit milestones (e.g., business or commercialisation contracts), where more clearly defined expected outcomes support better-formed KPIs. However, this cannot be conclusively established by the implementation in this thesis, due to limited data.

Ultimately, translating qualitative objectives into quantitative, decision-grade KPIs remains inherently difficult. Contemporary LLMs are capable across many aspects of automation, but evidently less reliable for high-judgement and context-specific KPI design that balances relevance, clarity, actionability, and measurability, at least by following the approach outlined in this thesis. Therefore, the most defensible near-term usefulness is in metadata extraction and recommendation, while still requiring a human-in-the-loop as a safeguard. In turn, this can improve customer relationship management (CRM) metadata completeness and enable collaboration health insights and automated reporting.

Keywords: AstraZeneca, KPI, LLM, GPT-4o, contracts, collaboration, automated, reporting.

Acknowledgements

Firstly, we would like to express our deepest gratitude to our AstraZeneca supervisors, Gaurav Gupta and Per Hillertz, for their guidance, trust, and day-to-day support throughout this project. Your insight, availability, and encouragement shaped both the direction and the quality of this work, and we are sincerely thankful for the opportunities to learn from you. Many thanks also go to the wider M&A IT team at AstraZeneca for your warm welcome, ongoing support, and for integrating us so well into the organisation. We are also thankful to the interviewees at AstraZeneca for taking the time to walk us through their processes and thoughts.

Moreover, we want to especially show our gratitude to Jesús Pineda, our university supervisor, whose commitment extended well beyond formal obligations. Even after his official appointment with the university had ended, he continued to provide thoughtful feedback and steady mentorship. His dedication and support were instrumental in the completion of this thesis.

We are also grateful to our examiner, Giovanni Volpe, for his role in the assessment process and for his perspective on the work. While less involved in the day-to-day development of the project, his input has been an important part of the overall evaluation and refinement of this thesis.

Finally, we would like to thank everyone who, in ways large or small, contributed to this project's progress and to our growth during this period.

Jakob Juul and Marcus Lorentzon, Gothenburg, January 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AI	Artificial Intelligence
AZ	AstraZeneca
BD	Business Development
CRM	Customer Relationship Management
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
PoC	Proof-of-concept
R&D	Research and Development

Contents

List of Figures

List of Tables

1
Introduction

Businesses across a multitude of industries depend on external partnerships to coordinate complex value chains, accelerate innovation, and meet predefined objectives. Increasingly, organisations operate within multi-party ecosystems where success relies on effective collaboration management [1]. In the pharmaceutical and life sciences sector, external partnerships and collaborations play a critical role in driving innovation, accelerating research, and gaining a competitive advantage [2]. However, the complexity and volume of such collaborations make it challenging to capture their true value and communicate their impact effectively [3].

Advancements in the field of machine learning (ML) have made it possible to automate not only tedious but also difficult tasks. Large language models in particular are one such domain of ML that has introduced novel capabilities beyond those previously thought to be attainable [4][5]. With such huge potential and broad application possibilities, it is of utmost interest for all companies, especially industry leaders, to leverage these tools to gain or retain a strategical market position. For example, in an article by Martín-Domingo, Fernandez Roblero, Efthymiou, *et al.* [6], the authors managed to use OpenAI's ChatGPT to extract KPIs with an accuracy of 71%, for airline emissions reporting. Models, such as ChatGPT, have achieved broad public visibility due to their demonstrated capacity to solve complex problems in text analysis and synthesis. This context creates a unique opportunity for AZ to leverage this technology to generate, process, and present data in support of decision-making.

1.1 Background

AstraZeneca is a leading global biopharmaceutical company originating from the UK and Sweden. AZ focuses on innovation through research, development, and marketing of pharmaceuticals [7]. The Gothenburg plant is one of the company's main centres of research and development (R&D), with 3100 employees ranging from formulation scientists to software developers [8]. The plant works on holistic drug development processes, from molecular biology to clinical trials of new drugs.

Chesbrough [2] introduced the concept of *open innovation* and describes it as purposefully allowing ideas, inventions, and knowledge to exit and enter an organisation in hopes of benefitting the innovative process of the organisation. He argues that

open innovation makes the process of innovating more efficient, reduces time to market, and betters the potential of innovative breakthrough compared to traditional closed innovation, which relies on internal R&D. Open innovation is presented with having the potential to innovate and advance the pharmaceutical industry, which generally has slow-progressing R&D[9].

Chesbrough [2] also highlights spin-offs as a form of open innovation. It is explained that a spin-off is when a company creates a legally separate business entity from the company while possibly maintaining some ownership. In this way, the separate business can commercialise innovations that are not part of the core business of the founding company. Some benefits of spin-offs are that they encourage entrepreneurial strategy, reduce risk for the founding organisation, and increase market adoption [2]. Wikhamn and Styhre [10] explains in an article how AZ collaborates with spin-offs by transferring internal projects to start-up companies backed by external venture capital. The study highlights several challenges facing AZ in this process, including internal decision-making difficulties, cultural barriers such as the 'not-invented-elsewhere' syndrome, and challenges in making internal projects appealing to external investors.

An example of how AZ works with collaborations is shown through the companies within the BioVentureHub. BioventureHub was launched in 2014 and has since hosted 57 companies at the heart of the AstraZeneca's R&D plant in Gothenburg [11]. The purpose of BioVentureHub is to create an open innovation environment that increases the competitiveness and energy of the life science sector using a public-private partnership model [11]. It allows start-ups and academic teams to work alongside AstraZeneca experts and access advanced laboratories and facilities, promoting collaboration and knowledge sharing.

In addition to the research taking place at the facility, AZ also actively participates in collaborations with organisations, such as research institutions and research-driven start-ups. AZ's relationships with these collaborators range from business partners, e.g. where AZ and a partner organisation develop a product together, to situations where AZ is investing in a company, to collective research with an academic institution. Managing the relationship with the collaborators in this ecosystem is crucial for their success in driving innovation, and therefore to keep AZ successful in the research and innovation oriented market of pharmaceuticals. Collaboration is at the heart of everything AZ strives to do.

1.1.1 Problem Description

Decision-makers at AZ need to constantly stay updated on the numerous different partners and the state of their collaborations. This includes analysing financials, as well as following scientific progression and product development, such as patents. This creates an enormous landscape of manual analysis that decision-makers need to consider. This analysis is time-consuming and prone to human error, especially considering that the data that needs compiling is heavily dispersed within AZ. These

data locations include CRM systems, publication databases, patent repositories, and investment trackers, as well as slide decks and emails. There is also a lack of accessibility for decision-makers to continuously keep up to date on the status of any given collaboration, including its health and progress.

In general, there is a lack of an intelligent system that can consolidate datasets into a coherent and evaluative framework. This makes it difficult for stakeholders to quantify both financial and non-financial outcomes of collaborations, identify opportunities, and benchmark against external activities. Therefore, the need to develop a system with automated capabilities, measuring the health and value of collaboration projects, has become increasingly prominent within the AZ business development team.

1.1.2 Scope of Implementation

In order to achieve the prospect of automation in the reporting purposes of collaborations, a multi-agent-system (MAS) is a potential solution that could be implemented [12]. A MAS divides tasks among multiple separate agents in order to manage a more complex goal [13]. This is the method deployed by Choi, Lopez-Lira, Lee, *et al.* [14] to retrieve financial insights, such as KPIs, from various reports. They developed a MAS, consisting of an extraction agent and a text-to-SQL agent, able to transform financial filings into structured data with an accuracy of 95%, showcasing the effectiveness of this approach.

KPIs are especially important for the solution proposed by this thesis, as they are vessels to coherently and concisely report the current state of a given endeavour. They are quantifiable measures that translate goals into trackable signals of progress and results. In collaborations, well-defined KPIs create a shared language for relevance, clarity, and accountability, linking objectives to data, timelines, and owners. They enable evidence-based decisions that ideally focus teams on what is truly impactful. Considering that as the end goal, developing an automated system similar to the one previously mentioned would, in the case of AZ's collaboration ecosystem, require several steps.

The first stage is the consolidation of all existing relevant data, which captures the nature of the collaboration. For a single collaborator with AZ, a multitude of information needs to be retrieved. Depending on the type of collaboration, these documents could include drafts of agreements, final contracts, non-disclosure agreements, and potentially significant metadata. To get a full picture of the collaboration, all information relevant to that relationship needs to be located, extracted, and then compiled, in order to eventually give a comprehensive report. This process could potentially be achieved with a specific agent that uses crawling techniques to search and gather relevant documents.

To establish what exactly the KPIs should be measuring, a separate agent could extract expected outcomes and objectives of a collaboration, based on what is stip-

ulated in the corresponding contracts of that partnership. Subsequently, another agent could define which KPIs best capture advances related to the identified objectives. In that way, performance measurements could be tailored to each collaboration.

Other agents could be tasked with quantifying the defined KPIs of a collaboration and retrieving relevant information to enable proper quantification of the KPIs. By contextualising the proposed KPIs with the continuous progress data on the corresponding collaboration, the end result would finally amount to bespoke valuable insights that could be reported to decision-makers.

A general overview of the problem and how it could be subdivided into separate tasks is illustrated in Figure 1.1. For the scope of this project, all of these procedures are not taken into account. Only the extraction of objectives and the definition of KPIs are considered. A further motivation for this can be read in Section 1.3.

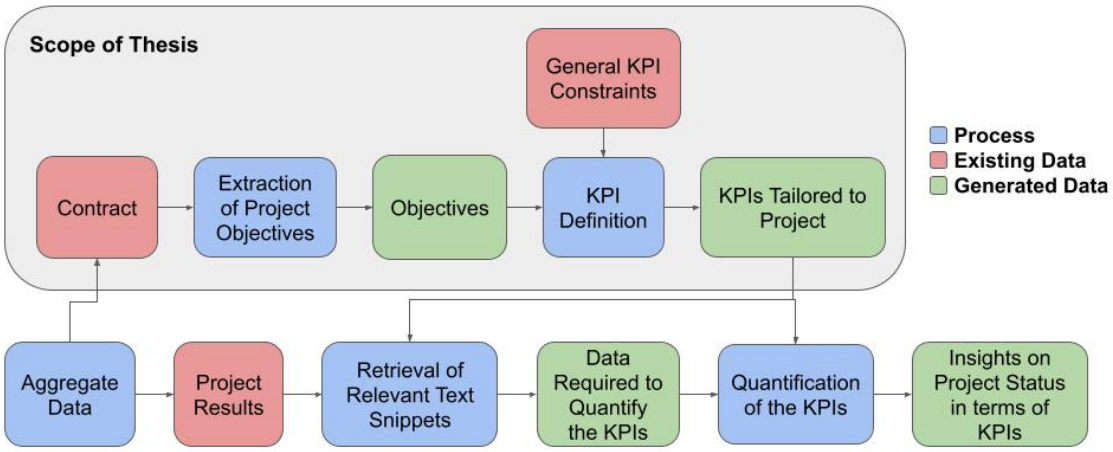


Figure 1.1: Schematic overview of how to address the problem. The scope of the project is defined within the gray box and is an important step of the bigger pipeline to evaluate collaborations. Blue boxes correspond to processes, such as tasks performed by an LLM. Red boxes are data that already exist within AZ in some shape or form, while the green boxes are data that are generated as a product of the developed system. The steps outside the gray box are left out of scope.

1.2 Purpose

This project strives to design and evaluate an agent that, based on contract agreements, can extract relevant KPIs for measuring the performance of a collaboration. This is one of the crucial steps of the overarching problem: taking agreements and

collaboration result data to automatically evaluate the health of a collaboration in terms of KPIs. The flowchart of the problem is presented in Figure 1.1.

When the LLM-driven system is developed, the performance will be evaluated based on real examples of contract agreements of different kinds, as well as dummy agreements, to see how well the agent is able to extract concrete objectives. Then the objectives will be used to define KPIs, which should reflect the KPIs that AZ actually wants to evaluate, and those that are feasible to extract from the existing data.

In doing so, the hope is that the system can introduce a way to increase transparency and business insight to meaningfully contribute to AZ and their partners, allowing them to enhance their collaborations and consequently advance their contributions to research and patient care. This project begins the development of the full system as a proof-of-concept (PoC) through the building of a key subsystem.

The project also aims to put the solution into the context of AZ and examine what value it may give the organisation. To do this, data management, collaboration project workflow, reporting process, and needs must be examined to accurately understand how such a model can be used effectively given the ecosystem, the needs of AZ, and what value can be created. An example is understanding what kind of KPIs capture value from collaborations.

This project seeks to investigate the following research questions:

- *How effectively can an LLM-based system extract and identify relevant KPIs from contract agreements of varying complexity to accurately report the health and performance of collaborations at AstraZeneca?*
- *Can this system be used efficiently in the pharmaceutical context of AstraZeneca?*

1.3 Limitations

As with all research projects, some limitations are inevitable. Conducting research in partnership with a large company comes with great benefits, such as experienced guidance and other resources, but also constraints related to organisational bureaucracy, which have shaped the projects scope. These limitations are presented here.

1.3.1 Scope-related

Due to limitations in time and data, the scope is to develop a system that solves one part of the general problem that the solution shown in Figure 1.1 strives to address. The project will not include implementations of every agent in the larger MAS and data analysis method in the system. It will also be assumed that the format of the documents is already processed in a suitable manner. Hence, the scraping and crawling of the relevant systems to extract documents will be left out of the project's scope.

1.3.2 Organisational

Due to strict requirements on data security, the proposed solution is restricted to utilise tools that are approved by AZ. This means that only the software, e.g. programs and models, that have been explicitly verified by AZ can be employed in our solution, thus ensuring that patient safety and company intellectual property are maintained.

1.3.3 Data access

Because the collaboration agreements analysed in this PoC are confidential, only a handful of agreements were supplied for the project. This hinders the possibility to make use of and extract more general insights. For instance, it confined the project to use pretrained LLMs rather than being able to train or even fine-tune any models. It also limits the ability to conclusively compare different kinds of documents. Finally, the proprietary nature of the few agreements that were provided restricts this report from presenting any specific examples.

2
Theory

This chapter gives a brief overview of the foundational concepts that are needed to properly digest this report. The theories are described but not explained in detail, so for a deeper understanding, it is recommended to explore the cited sources.

2.1 Deep Learning

Deep learning is a subset of ML and powers advanced artificial intelligence (AI) methods, such as computer vision, classification algorithms, and generative AI, including LLMs [15][16]. The concept leverages neural networks consisting of multiple layers of artificial neurons to learn non-linear representations of data [15]. The neurons are connected through a weight and bias, which together with a non-linear activation function define the behaviour of the neural network model [15][16]. The model is trained on data samples by minimising a loss function that measures prediction error based on the outputs of the forward pass [15][16]. Backpropagation computes the gradient of the loss function based on the weights and biases [15][16]. Using gradient descent, the learning rule is deducted, which updates the parameters in the direction that minimises the loss function most [15]. By applying this concept to a network of enough layer size and amount of layers, it is possible to model complex problems.

A drawback of deep learning is the so called black box behaviour that implicates low intuitive interpretability of the model output [15]. Complex neural networks are also computationally heavy, which means that a huge amount of energy is required for the increasingly sophisticated hardware to perform the calculations on a massive scale [15]. Often, a massive amount of labelled training data is needed for these larger networks to achieve optimal performance [15].

2.2 Large Language Models

During recent years natural language processing (NLP), and more specifically LLMs, have fundamentally improved their ability to process and generate text to perform various tasks at human level [4]. LLMs have gained abilities such as decision-making, reasoning, planning, and in-context learning due to the gigantic scale of the models [5]. The LLMs, which are probabilistic sequence models, achieve this by predicting

the next token [5]. All tokens are subword units that have a unique index. Each index is then mapped to a learned vector embedding. These embeddings are in turn numerically processed in the deep learning model, such as an LLM, to distinguish between contexts and generate outputs.

At the core of the LLM architecture is the transformer. The transformer replaced older methods, such as Recurrent Neural Networks, and enabled the models to process tokens in parallel instead of sequentially [17]. The transformer consists of multiple layers creating encoders and decoders, which can be seen in Figure 2.1. The encoder layers embed the input into a higher dimensional space, leveraging feed forward network layers. The decoder layers utilises these higher-dimension embeddings to create output sequences [17]. The transformer architecture needs to encode the position of a token and in that way account for the order of the words when processing them [17]. Traditional methods such as Recurrent Neural Networks process data sequentially, while transformers use multi-head self-attention [18][17]. Self-attention relates tokens to each other by weighting the relevance of other tokens to a specific one [19]. The model can then dynamically process information from multiple input positions [19]. This allows the model to consider positional data from all positions and provide contextual awareness [19]. The multi-head self-attention runs several attention "heads" in parallel, each with its own learnt projections of the input [18]. Every one of these heads computes attention weights over the sequence to produce an output focused on different aspects of the data [18]. The head outputs are then concatenated and passed through a final projection to form the layer's result. This design permits the model to process multiple representation subspaces and positions at once, capturing patterns that a single head would blur together, while keeping the computation efficient by using reduced dimensions per head [18].

The pretraining step in the LLM training requires a large amount of data and compute. The weights of the layers are optimised by training on billions of tokens, in the case of full-scale models. After pretraining, the model can be fine-tuned for specific tasks, e.g. processing a certain kind of document or to follow instructions and answer questions by training it on question-answer pairs [20]. To perform the correct tasks with better accuracy, it is possible to give input prompts in ways that guides how the output should be created [21]. For example, few-shot prompting includes providing some examples of how the model should respond when given the instruction [21]. In this way, structured output can be generated.

2.2.1 GPT-4o

Researchers and industry are racing to develop the most powerful and intelligent LLM systems [22]. During the last years, Google have developed Gemini while Anthropic have Claude. OpenAI, made the first mass-breakthrough beyond the AI community with their ChatGPT. GPT stands for *Generative Pretrained Transformer* and is based on a similar architecture to the one shown in 2.1. Since then, OpenAI have developed a range of models: GPT-3, GPT-3.5, GPT-4, and currently

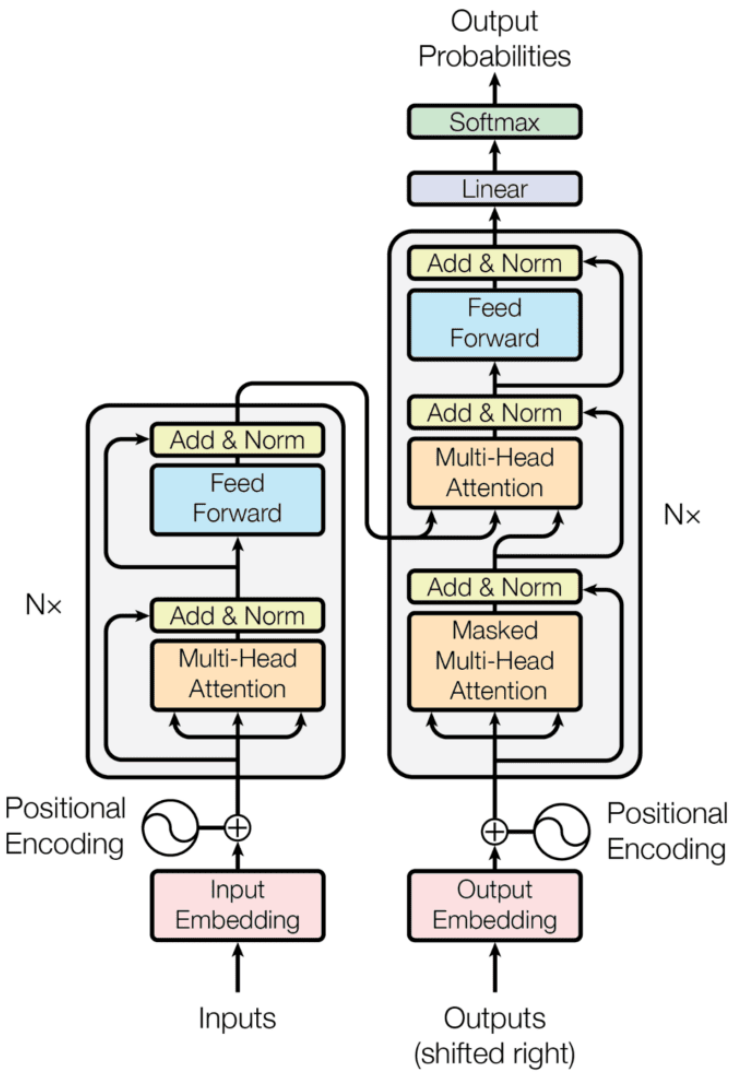


Figure 2.1: The images shows an example of a typical transformer architecture [18]. The architecture includes the encoder and decoder part, both with processes such as Multi-Head Attention, Positional Encoding, as well as a Feed Forward layers, among others.

GPT-5 just to list a few. GPT-4o, in particular, was released in 2024 to surpass the performance of the competitors [22] and is estimated to consist of over one trillion parameters. This is well above the number of parameters that the competing models had at the time. The "o" in GPT-4o stands for *omni*, which highlights the model's ability to accept prompts consisting of audio, images, and text [23].

In the study *Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency* [22], GPT-4o is tested on multiple exams to evaluate language and the model's ability to understand and solve complex problems. In the *The United States Medical Licensing Examination Step 1* (USMLE) test, the model achieved an accuracy of 83.5%, which is a lot better than GPT-3.5, which attained 51.67%. However, this was lower than GPT-4, which scored 90.00%. On *The Chartered Financial Analyst Level 1* (CFA) the GPT-4o performed with an accuracy of 85.6% and beat the other two models by more than 12 percentage points. When assessed on *The Scholastic Assessment Test* (SAT), the model obtained a 90.91% accuracy on reading and writing questions and 87.48% on mathematical questions.

In a benchmark study conducted by OpenAI, after the release of GPT-4o, it was clear how, at that time, the model outperformed previous models from OpenAI and competitors alike [24]. The benchmarking was evaluated on different benchmarking sets, such as HumanEval and MMLU. The results of the comparison can be seen in Table 2.1.

Table 2.1: Text evaluation accuracy (%) across benchmarks and models from study by OpenAI [24]. Blank fields indicate that there is no data for that measurement.

Benchmark	GPT-4o	GPT-4 (23-03-14)	Claude 3 Opus	Gemini Pro 1.5	Gemini Ultra 1.0	Llama 3 400B
MMLU	88.7	86.4	86.8	81.9	83.7	86.1
GPQA	53.6	35.7	50.4	–	–	48.0
MATH	76.6	42.5	60.1	58.5	53.2	57.8
HumanEval	90.2	67.0	84.9	74.9	84.1	84.1
MGSM	90.5	74.5	90.7	88.7	79.0	–
DROP (F1)	83.4	80.9	83.1	78.9	82.4	83.5

3
Interviews

Interviews were conducted to answer the research question regarding the organisational workflow of collaborations at AZ to understand how an automatic reporting system can be used. To answer this question, a qualitative approach is needed. The purpose of the interviews was to understand how people in the organisation working with collaborations report the progress of the collaboration, and what their stakeholders are expecting from them in terms of reporting. This is a way to put the automatic reporting system into the organisational context of AZ, and to understand strengths and weaknesses of the system in practice. The manner in which the interviews were conducted and analysed is described in Appendix A.

The answers of five interviews are presented in this chapter after being thematically grouped into sections. To keep the anonymity of each participant, they are referenced by their title and number, e.g. *Alliance Manager 1*.

3.1 Different Kinds of Agreements

According to *Principal Scientist 1*, collaborations take multiple forms: "It can range from collaborations around pharmacological equity to technical development of instrumentation, techniques, or scientific collaborations aiming to publish new scientific discoveries." The interviewee continues to explain how the collaborations differ. In terms of follow-up for pharmacological equity, there are usually agreed milestones that trigger payments or continuation of the project. Technical development has to do with providing feedback on new equipment and evaluating the usefulness of the data in the pipeline. Scientific collaborations have a more free form of structure, and the nature of the reporting is up to the principal investigator and their academic counterpart.

Alliance Manager 1 explains in the interview that academic agreements start out on shaking grounds but get more solidified as the project progresses and AZ continues working with them. The participant continues to say that on the other side, business agreements are "cast in stone from the very beginning and you need to deliver this to get money", meaning that business collaborations are more dependent on defined milestones, which are required for payments to take place. Academic collaborations do not have to stop if specified milestones are not reached.

As explained by *Vice President 1*, differences arise because commercialisation agree-

ments require significant effort to ensure that all parties are aligned. As they explain, "tremendous amounts of work that goes into that, because companies also come at things with different points of views, different cultures, different sizes, different stages of the life cycle of the company itself, different priorities. So, it is quite complex to navigate."

Principal Scientist 2 works in an academic collaboration and mentions that milestones are also set up and agreed upon in advance for this type of collaboration. Another person working within an academic collaboration is *Science Director 1*, who says "it's research, we don't know the value beforehand." Thus, even if the work process is structured in their case, the value and goals might not be clearly defined. When asked about clearly defined milestones the participant answered, "I don't know how it would work, right, but it's a bit maybe complicated. I don't know because it's research right? We have no idea the direction. So we can start with something and we can discover something else. We will change totally the milestone on what we saw." The statement paints a picture of the research as an exploratory effort. Additionally, the interviewee describes business collaborations as having much clearer timelines.

3.2 Reporting Data Workflow

Multiple interview participants testify that reporting to senior management and stakeholders is usually done solely through meetings and presentations on progress, and at vastly different intervals. *Principal Scientist 2* explains their process as follows: "As part of the post-doc proposal, both parties agreed to a set of goals (6-month goals, 1-year goals, 2-years goals) that are monitored throughout the project via recurring meetings (fortnightly meeting frequency), reports (quarterly status updates, 6-month progress report, annual written report)." The participant continues to describe that progress is reported to the head of the department. During these recurring meetings, progress, which often takes the form of collected data or new insights, is discussed in relation to the goals. In the same participant's current project, results from modelling analyses are literature studies, which correspond to the progress that is being reported. The outcome is written in the mentioned report, which is compiled by the external collaborator. This report is stored in a common SharePoint or Teams channel.

Moreover, *Allicance Manager 1* states that the frequency of reporting depends on how fast the project is moving and how often there is something to report. If there is nothing to report in quarterly meetings they will report bi-annually. The respondent also mentions that SharePoint is used for storing report data if not stored in folders of the alliance manager.

Science Director 1 reports in a similar fashion. Once a year, a one slide update is presented, including goals and key deliverables. Apart from that, there is one oral presentation per year with all data and progress. This reporting is directed at alliance management. Reporting data does not seem to be stored in any centralised

database, according to *Science Director 1*.

For commercialisation collaborations, *Vice President 1* states that there is often a joint steering committee consisting of senior leaders that handle the governance of the collaboration. Any reporting is ultimately targeted towards this committee. It is clarified that the reporting to the committee is qualitative and takes place as presentations in quarterly meetings.

Vice President 1 continues by saying that a system for structured and recurring reporting would be of high value to both parties of a commercialisation collaboration. However, the system would have to bridge the structural gap between the organisations to be efficient and actually counteract duplication of work. Some difficulties related to this, especially if the system relies on AI, is getting the partner organisation to trust the system with their sensitive data.

Nevertheless, through these interview responses, it showcased that progress reporting on collaborations is performed dynamically and in an unstructured manner, depending on the needs of the involved people. Instead of continuously and systematically reporting milestones so that stakeholders can keep up to date, progress is reported in a free format with a certain infrequency. This is a natural way of working on these projects, as they often are fundamentally dynamic, as research often is. However, it creates difficulties when considering an automated approach, as automation requires access to systematic and structured data.

3.3 Relevant KPIs

In what way the value and health of a collaboration can be measured in terms of KPIs is described in detail by *Alliance Manager 1*. The participant mentions how conducting scientific research can create opportunities for follow-up research, which is a form of value. Conducing successful research also improves the scientific assets of AZ, which in turn creates opportunities for recruiting talented post-docs or PhD-students, since they may be keen to join future projects and work. A person who has worked in an AZ collaboration has already acquired relevant experience to continue working effectively in AZ. The interview participant says that relationships and academic contacts related to previous research are important values, especially for future and follow-up research. Some of these thoughts from *Alliance Manager 1* are captured in the following statement.

> *You just don't randomly pick up a lab or a scientist to work with in an academic collaboration or even anywhere. It's not like you know, you just scan through [names and say], this name looks fantastic. No, you would have probably worked with him. That person would have presented somewhere. That person would be having an asset which is being referred to in multiple publications. [...] That creates a value for that person, and when you get into that, that becomes your KPI. [...] Sometimes you*

would find it very well written in the agreement, but sometimes not.

Naturally, other performance indicators of research-oriented collaborations are key deliverables and scientific publications, as exemplified by *Scientific Director 1* and *Principal Scientist 1*. *Alliance Manager 1* expresses the strong benefits of published scientific work in the following way.

> *It's a huge qualification for academic collaborations, because if we are to do a publication, we need to generate a lot of data ourselves, and that data may not be towards our goal of bringing new medicines to the market. That is the primary goal for all of us. Now, when you are working with an academic partner, the academic partner does 60% of the job, we put in the 40%, but you still become part of the publication. So that also adds value, and that also adds value to the people who work here. As you know, I've got a publication in say impact Factor 15, which is a huge thing. Anything around [impact] factor 10 is pretty good, so that adds the value again.*

Ultimately, *Alliance Manager 1* provided a list of KPIs deemed to be relevant for measuring the performance of collaborations in general. The list looks as follows:

- The deliveries of the collaboration within the time frame and budget
- External funding secured
- Number of post-docs and PhDs working on the collaboration
- External presentations
- Publications or joint publications
- No escalations to senior leadership
- Positive feedback during health check-ups

According to *Vice President 1*, who works on commercialisation agreements with companies, early in the collaboration process, the joint commercialisation committee will establish the goals, how to measure them, and the value that the partnership will bring. These criteria are critical to the success and efficiency of the collaboration, as is establishing principles of cooperation. The participant continues by pointing out that these collaboration criteria have to be explicitly determined, particularly in the case of financial KPIs: "Although the business performance is ultimately going to be the main goal for a commercial collaboration because the collaboration has to have a positive financial impact or it will not continue very long."

Although, complexities and differences in the partnership with each commercial collaborator, make measuring of KPIs inconsistent. *Vice President 1* says, "for every company there are similarities, but sometimes companies want to see the same kind of information in different ways and it would be good to try to be more consistent and try to harmonise. If there was some kind of tool to try to help with that, that would be helpful."

Finally, based on what has been expressed by the interviewees, a set of generally

applicable KPIs could be formulated. These are presented in Table 3.1 and become important for the prompting of the system developed in this project.

Table 3.1: The KPIs presented in the table are deduced from the interviewees' suggestions on measuring the value stemming from collaborations.

#	**KPIs**
1	Patents filed
2	Follow-up research
3	Impact Factor of scientific publication
4	Reputation of the journal of scientific publication
5	External engagement (e.g., event presentation)
6	External funding for projects
7	Alignment with company goals
8	Budget coherence
9	No escalations to senior leadership
10	Recruitment of Post-Doc from University
11	Collaboration-specific KPIs based on objectives

3.4 Pain-points and Needs

When the interviewees are asked about pain-points in the current workflow, potential improvements, and the role of new technologies, such as AI, most participants (*Principal Scientist 1*, *Principal Scientist 2*, and *Alliance Manager 1*) answer that they see no problem with the current reporting. However, *Science Director 1* replies that there is not much transparency in the system, making it difficult to find information on ongoing collaborations. The participant mentions that a common collaboration platform could improve these issues.

The responses regarding the role of new technologies included AI for note taking, a system for identifying new partners, a data base with past and ongoing projects, and AI for streamlining the processing from multiple collaborations.

In sum, the interview participants are satisfied with the workflow of reporting today, but some of them see opportunities to use new technologies to improve the transparency and accessibility of data related to collaborations. This could be a database that connects the set milestones of a collaboration to the most recent progress in terms of those milestones. This finding aligns well with the scope of this thesis and the identified problem to be addressed.

4

Implementation

In this part of the report, everything from how information was collected to how the end product was created is covered. All these steps are based on the premise that the purpose of a collaboration can be read in the contract of that project. Then, they further build on the assumption that that specific purpose can be condensed into quantifiable performance indicators. The implementation of the technical LLM pipeline will be presented, and that workflow could be considered in the following parts. Figure 4.1 shows a visualisation of the different stages in the implementation of this thesis including the evaluation steps, which are explained in Chapter 5.

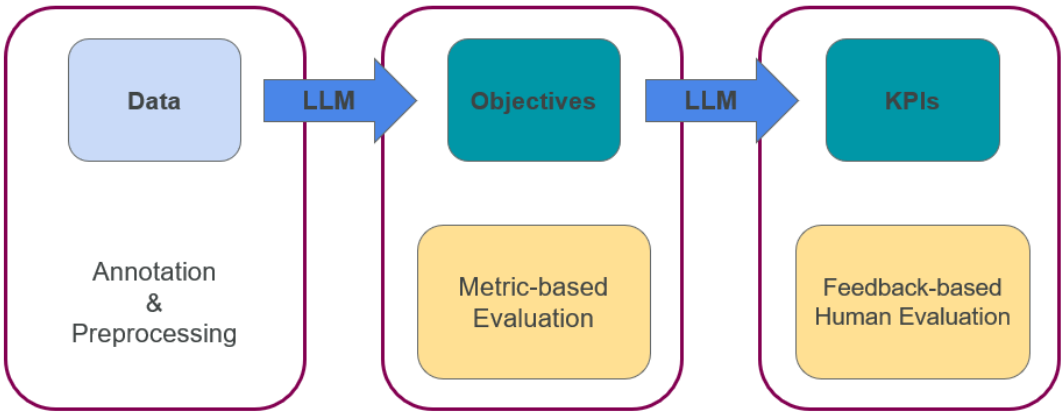


Figure 4.1: The figure visualises the evaluation pipeline, including the first step of data preparation and annotation. The objectives are extracted by the LLM and are subsequently evaluated. In the last step, the LLM defines KPIs from the objectives. These KPIs are evaluated by asking experts working on each respective project for feedback.

4.1 Input

As defined by the scope of the project, the relevant data needed to address the specified problem consist of legal contracts between AZ and their corresponding collaborator. These documents are gathered from AZ's CRM system for business development (BD).

4.1.1 Data Preprocessing

The agreements are confined to a PDF format and do therefore require suitable preprocessing to be used as input for an ML model. With the use of the PyPDF library in Python, the text can be extracted from the PDF files. Depending on whether the document has been scanned from a physical copy, a separate method was applied to extract the textual content; when scanned, the PDF file considers the entire content of a page as an image. Therefore, the words are not discernable by the PDF-reading program, which is why those documents have to be processed with optical character recognition software. There are a multitude of different solutions that do this, but to ensure the confidentiality of the data, only AZ-approved tools could be used. To avoid data leakage, Microsoft OneNote was used to prepare the text in these cases, even though the program's primary function is not to extract text from images. There were discussions about granting the project approved OCR tools from Amazon Web Services, but the access was never finalised. Ultimately, adopting this solution would likely result in much higher quality transcripts.

Moreover, the length of the documents had to be adjusted in some instances, due to token restrictions. Those files had to be shortened to fit the given requirements. Therefore, such contracts were cut to include only a certain piece of the whole text that was deemed to be relevant. This may be considered a form of selection bias; however, given the scarcity of AZ data, it was considered preferable to use as much accessible data as possible. An alternative that was evaluated was to process the entire document by passing segments of it through the system, which ensured that all the text would be processed. Nevertheless, this approach has its own separate drawbacks.

Before getting security clearance, part of the solution was evaluated on legal documents from a public database. Hence, the development of the system could proceed without having to wait for the proprietary data. When permission was granted to the agreements, access to a limited dataset, containing a handful of confidential contracts, was given.

4.1.2 Data Annotation

One of the most important and time-consuming tasks of the project was the annotation of the data. Since the intention was to evaluate the performance of the system, it required structured information, i.e. a labelled dataset. This is so that the output of the LLM can be directly compared to something that, according to us annotating, can be considered to be the objective truth. In NLP, this ground truth is often referred to as the gold standard. Since it is pivotal for gaining meaningful outcomes, it is usually performed by domain experts [25]. Even though we, as annotators, are neither legal nor pharmaceutical experts, we annotated the contracts based on our experience and sense of the topics. Due to the time required to complete this task, it was not feasible to let other people in the organisation perform the annotation. The labelling process involved going through each of the documents and identifying which statements in the contract could define the overarching purpose of the col-

laboration. Often, there were segments containing sections such as a research plan, which were particularly helpful in defining the project objectives. If a document did not outline any motives for the collaboration, it was disregarded.

Once an objective was identified, it was copied over, word for word, to a JavaScript Object Notation (JSON) file. JSON is one of the data standards in the field and was very useful for the purposes of structuring the annotations. The objectives were not subdivided into sub-objectives, although in some documents a sentence or bullet point could consist of what can be considered multiple objectives. Only the part of the text sequence that related to what could be considered an objective was placed in the gold standard JSON-document. In order to keep consistency, it was decided to start the annotated segment with an action-verb, when it was possible and logical. For instance, if the text in the agreement was "the goal of the project is to increase the research capabilities within the given field", then the objective that was annotated in the gold standard would be "increase the research capabilities within the given field".

Furthermore, no more information than the objectives that we had identified was annotated, even if that additional intelligence could have been valuable. An example of such information could be data that is useful in contributing to specific KPIs for a given collaboration, based on their agreement. For instance, this could be information such as time-duration, budget, reporting, research method, or number of samples. This was left out of scope due to time constraints and challenges in annotation and evaluation. It would have been too cumbersome to determine exactly which data points to target and then identify all of them from each document. There are simply too many relevant data points to extract in that case, and doing so puts the thesis down a quite different path. Sticking to extraction of objectives was therefore seen as the most sensible prospect.

Also, it was not clear to what extent the LLM could perform multiple tasks in one prompt, so to be sure that it would not have a negative impact on the extraction of the objectives only, it was omitted completely. An alternative would have been to dedicate a separate agent (instance of the LLM) to perform this task and compare its results. However, doing so would have expanded the thesis outside of the originally defined scope, which is why it was not investigated further.

4.2 System

The subsequent passages involve the implementation of the AI models that enable the extraction of objectives and definition of KPIs from legal contracts. The idea is to automatically identify important quantifiable metrics for any given collaboration.

To achieve this, a range of software libraries were employed in the development of the Python-based solution. In particular, Azure OpenAI was used to access the GPT-4o model. Although there are many alternatives through Microsoft Azure [26],

it was the only one provided by the AZ organisation. It is, however, still a highly functional model [22] that has been frequently investigated in research, making it suitable for what is sought to be achieved in this thesis. Pretrained LLMs, such as GPT-4o, serve as sufficient means for a PoC, but for a full-fledged future system it could be wise, depending on available resources, to train it, or at least fine-tune it, to the specific needs of AZ.

Another important library was LangChain, which is a common open-source framework for NLP. By implementing it, the system could be designed to allow future swapping of LLMs without difficulty [27]. With the LangGraph library, the construction of a larger system could be facilitated by connecting and coordinating specialised subtasks through a graph network approach. Thus, the shell of the pipeline described in the flowchart of Figure 1.1 could be implemented with greater ease and with options for extensibility.

The described system outlines one part of a proposed MAS. Thus, it strives to be a piece of a larger puzzle. This puzzle piece can, in terms of the graph representation, be broken down into the following constituents. The first node is defined as the objective extraction. It takes the content of an agreement and adds it to the prompt. The result is the collection of identified objectives. The second node corresponds to the KPI generation. This takes the output from the previous node, includes it in its prompt, and returns the generated KPI definition related to the identified objectives. Prompt-engineering is a substantial part of the system's development process, as it greatly affects the results from each invoke of the LLM. For that reason, the prompting has changed during the project to adjust the results and accommodate new needs. The prompts used to achieve the final results are presented below. The first one instructs the LLM to extract objectives from the provided text.

```
You are an expert at identifying big-picture goals, objectives, and
    targets. Your task is to provide insights from a document in the {
    state.get('domain')} domain.

Document:
{state.get('raw_text')}

Based on the provided text, you should answer what the purpose of the
    given collaboration/partnership is.
Do this by extracting the overarching objectives of the collaboration
    between the company and its counterpart, from the provided text. If
     none are present, then do not output any.
It is extremely important that each objective should be a separate
    point. Do not answer with super long sentences, but rather keep the
     extracted objectives concise and to the point. If the objective is
     longer than a sentence, it can most likely be subdivided into
    separate objectives.
```

```
It is also crucial that you make sure to quote the text directly, i.e.
     do not alter any of the excerpts!
If there are multiple objectives provided in a sentence, split them
    into different objectives. Never include more than one objective
    within an objective.
When possible, start the quotation of the identified objective by
    beginning with the first action-verb of that text sequence, e.g., "
    The aim is to measure..." becomes "measure...".

The output should follow the format below and thus be returned as a
    JSON array:
[
  {{
    "obj_id": "obj_1",
    "text": "quote of the full objective statement"
  }},
  {{
    "obj_id": "obj_2",
    "text": "quote of the next full objective statement"
  }}
]

Return ONLY the JSON array! It is vital that you do not respond with
    any other text.
```

For the generation of the project specific KPIs, the prompt was instead formulated as follows.

```
You are an expert at defining quantifiable key performance indicators
    (KPIs) from identified objectives for collaboration projects in the
     {state.get('domain')} domain.

Objectives:
{state.get('objectives')}

Instructions:
Define KPIs that can be measured grounded strictly in the objectives
    above and useful for decision-makers. There can be multiple KPIs
    per objective.
Return ONLY a valid JSON array. Do not include any explanations,
    comments, markdown code fences, or a leading JSON label. The first
    character must be "[" and the last must be "]".
Each array element must be a JSON object with keys: "kpi_id", "kpi", "
    relating_objectives".
"relating_objectives" must be an array of objects with keys "obj_id"
```

```
    and "text". Use the exact objective text for "text".
Assign sequential KPI IDs: kpi_1, kpi_2, kpi_3, kpi_4, kpi_5, kpi_6,
    ...
Preserve objective IDs if provided.
Use plain UTF-8 characters and standard JSON escaping. Do not use
    trailing commas.
If no relevant KPIs can be created from the objectives, return a
    single-element array with one object where "kpi" is "insufficient
    objective statements".
Example format (shape only): [ {{ "kpi_id": "kpi_1", "kpi": "a
    measurable KPI", "relating_objectives": [ {{ "obj_id": "obj_1", "
    text": "full objective text" }} ] }} ]

There are some typical KPIs that should always be included in the KPIs
     but formulated based on the objective. These KPIs are:
{KPI_constraint}

Apart from these, project specific KPIs based on objectives should
    also be included.
```

4.3 Output

The output of the first node is a string meant to be formatted as JSON. To assert that this truly is the case, it is processed through a Python function that extracts a correct JSON-object from the string. Sometimes, it is possible that the string may contain incomplete objects, such as when the context window is surpassed. At that point, this safety measure ensures that the output passed to the next step is in the correct format. Specifically, all complete objects are passed on, while the incomplete remainder is disregarded. Moreover, the output is also saved in a JSON-file containing all the processed contracts. This is so that the results of a given run can be evaluated at a later stage. The reason for keeping the results from each document in the same JSON-file is mainly because of convenience. By doing so, there are substantially fewer files to keep track of, and they can be grouped together based on which prompt they were initiated from and at what time. This also facilitates the annotation and evaluation process as all the gold labels can be collected similarly in one place.

In the example below, it is possible to examine how a JSON format of extracted objectives might look. The report ID refers to which contract the objectives belong to. The three consecutive dots mark a hypothetical continuation at that level in the data structure.

```
[
  {
```

```
    "report_id": "*first contract*",
    "objectives": [
      {
        "obj_id": "obj_1",
        "text": "*The text of the first objective in the first
    contract*"
      },
      {
        "obj_id": "obj_2",
        "text": "*The text of the second objective in the first
    contract*"
      },
      ...
    ]
  },
  ...
]
```

If the model is not able to define a KPI based on the presented objectives for some reason, e.g. due to the absence of identified objectives, it is instructed to answer with "insufficient objective statements".

When it comes to the output of the second node, it is fairly similar to the previous case, although the JSON-object contains the proposed KPIs instead. Linked to each KPI is the set of objectives that the KPI aims to quantify. The example below illustrates how this structure might look. The output is also stored in a JSON-file, to save the progress. The KPIs are later retrieved and incorporated into surveys that are sent out to people with the right expertise to receive human feedback.

```
[
  {
    "report_id": "*first conctract*",
    "kpis": [
      {
        "kpi_id": "kpi_1",
        "kpi": "*The first KPI defined from the first contract*",
        "relating_objectives": [
          {
            "obj_id": "obj_1",
            "text": "*The text of the objective, in the first contract
    , that this KPI corresponds to*"
          },
          ...
        ]
      },
      {
```

```
        "kpi_id": "kpi_2",
        "kpi": "*The second KPI defined from the first contract*",
        "relating_objectives": [
          {
            "obj_id": "obj_1",
            "text": "*The text of the objective, in the first contract
    , that this KPI corresponds to*"
          },
          ...
        ]
      },
      ...
    ]
  },
  ...
]
```

5
Evaluation Criteria

It is crucial that the results from the system can be trusted. For that reason it is necessary to validate the outputs at each step of the pipeline, by evaluating its performance therein. As LLMs have a tendency to hallucinate, i.e. fabricate information that is unsupported by their training data and reality, validation is made essential. Therefore, it is important to ensure that the produced results are traceable back to the source from which they were retrieved. These concepts are described in further detail in this part of the report.

5.1 Evaluations of Objectives with ML Metrics

The performance of the objective extraction is evaluated using multiple metrics. Many of these metrics are standard in the field of ML, and will be described in greater detail. For the output of the LLMs to be evaluated, there needs to be a reference. In this case, the reference is a golden label, i.e. the ground truth, which represents the text snippet in the input text that ideally should be extracted and outputted by the model. The process of labelling the objectives was described in Section 4.1.2.

To consistently compare the set of predicted objectives to the set of golden objectives, in terms of the ML metrics, an evaluation program was written. There, the golden labels of one document are compared to the output prediction from the LLM. The texts will be presented in sets of multiple objectives with no connection between which objectives in each respective set correspond to each other. To pair up the corresponding objectives a general purpose sentence embedding model is used, commonly called SBERT [28]. These models are LLMs fine-tuned for comparing sentences. The way this is done is by running both sentences through the backbone of the LLM to get the embedding vectors. The vectors are then analysed using cosine similarity, which is a measurement of the similarity of the two embedding vectors. A high score of a cosine similarity means the texts are semantically similar, even if the exact words differ. The process can compare every single objective in the objectives extracted from the LLM to the gold label objectives to see which ones match the best. The algorithm used to match these objectives was a greedy one-to-one matching algorithm, which takes the highest value in the cosine similarity matrix, matches the corresponding objectives and then removes the values in the matrix corresponding to these objectives by setting the specific row and column to $-\infty$. This process is repeated until there are no more objectives left from one of the lists. If there is

an unequal amount of objectives in the lists, some objectives will not be paired up and therefore ignored. If the objectives are very different they might still get paired up but will create a low evaluation metric, mirroring that the LLM performed poorly.

When choosing which model to use for creating embeddings of sentences to calculate the cosine similarity, there were multiple choices. To find the models that can best embed the context of the objectives, a test was conducted. Using public agreements from ResourceContracts.org [29] as dummy data, the full evaluation pipeline was employed to examine the average cosine similarity. The tested models and the corresponding metrics can be seen in Table 5.1, where *gte-large* has the highest cosine similarity. Although, when examining specific examples, it was made clear that the models *e5-base-v2*, *e5-large-v2*, and *gte-large* had a positive bias. They seemed to output too high cosine similarity, particularly for cases where the sentence clearly had differing semantic meaning. For that reason, *all-mpnet-base-v2* and *all-MiniLM-L6-v2* remained the two embedding models to be explored. Ultimately, *all-MiniLM-L6-v2* was selected for the cosine similarity functions of the evaluation pipeline, as it performed marginally better. More information on why this choice was made is presented in Appendix B.

Metric	**all-mpnet-base-v2**	**all-MiniLM-L6-v2**	**e5-base-v2**	**e5-large-v2**	**gte-large**
Precision	0.3948	**0.3999**	0.3885	0.3995	0.3951
Cosine Sim.	0.4875	0.4879	0.5671	0.5709	**0.5807**
Recall	0.4875	0.4961	0.4834	**0.4970**	0.4841
F1	0.4263	**0.4344**	0.4211	0.4341	0.4265

Table 5.1: The table shows a comparison of embedding models across metrics tested on dummy data. The highest score per metric is marked in bold. The reason the scores are significantly lower than one is because there are some predictions that are not in the golden labels, and some that have different semantic meaning.

Four different evaluation metrics are used when evaluating. The first is cosine similarity, which is also used to match the objectives in the greedy match algorithm. For two vectors, $\mathbf{a}$ and $\mathbf{b}$, the cosine similarity is calculated as follows:

$$\textit{Cosine Similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||||\mathbf{b}||}$$

Cosine similarity shows how similarly the texts are embedded by comparing the distance between the vectors in the latent space. The LLM should embed them similarly if the model works well for the text, assuming that the semantic meaning is similar.

Other metrics are precision and recall, which the F1-score is comprised of. Given a pair of sets of tokens, which in this case is one output prediction and one golden label, the token level precision is calculated as:

$$\text{Precision} = \frac{|\text{Predicted Output} \cap \text{Golden Label}|}{|\text{Predicted Output}|} \tag{5.1}$$

This quantifies the proportion of the predicted output tokens that also appear in the corresponding golden label tokens. Similarly, the recall metric calculates the proportion of the golden label tokens that also appear in the predicted output tokens. It is instead calculated as:

$$\text{Recall} = \frac{|\text{Predicted Output} \cap \text{Golden Label}|}{|\text{Golden Label}|} \tag{5.2}$$

The F1-score measures the harmonic mean of these two proportions and is defined as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

These metrics cover different aspects of the result. Cosine similarity assesses context and paraphrasing, whereas the F1-score evaluates lexical overlap by measuring whether the exact tokens in the reference set are present in the prediction. The F1-score combines recall and precision; predictions containing many tokens not present in the gold standard reference are penalised through lower precision, while predictions that omit many reference tokens are penalised through lower recall. By balancing these two components, the F1-score accounts for both overinclusive and underinclusive predictions.

At the end of the evaluation, when all the objectives have been matched and evaluated in one agreement, the average score was calculated for that document. The average gives an overall score for each and every agreement. The average score was calculated in two ways. One score consists of the average for all the matched objectives and ignores the unmatched ones. In the other case, the average score takes the unmatched objectives into consideration and applies zero-padding. This zero-padding means that the unmatched objectives that remain either in the golden objectives set or the predicted objectives set will be given a score of zero for all four evaluation metrics. For instance, if only half of the predicted objectives are matched with a golden objective, none of the evaluation scores will be more than 0.5; Half of them are considered incorrect and therefore they penalise the total score.

The reason for keeping both methods is that they show different sides of the result. The average matched score, without zero-padding, shows how well the objectives that are matched actually perform in terms of the evaluation metrics. In plain words, it reveals how many of the reference objectives were identified. In contrast, the zero-padded and penalised average considers the whole output and can therefore present a bad score if the gold standard set and prediction set differ in size a lot. Thus, it also captures to which extent the LLM has identified more or less objectives that the annotators have. The scores do therefore not solely reflect how well the model managed to extract the specific wordings of each objective. In the end, an average of all the agreement averages is calculated for both the matched and penalised metrics.

5.2 Evaluation of KPIs with Human Feedback

When evaluating the KPIs generated by the system of LLMs, a human evaluation method was implemented where the KPIs are manually validated based on a defined set of criteria. The reason for choosing this method is that the KPIs do not have a predetermined golden label, unlike the objectives, which can be retrieved directly from the agreements. The KPIs are something that logically need to be derived based on the objectives. They indicate the performance of something, and that something is determined by what the objective of a given endeavour is. In addition to that, there are an arbitrary number of different ways in which the same KPI can be defined and expressed. Moreover, multiple different KPIs can measure the same objective. These complexities make it infeasible to create labels in the same way as when evaluating objectives. The golden labels need to be an objective truth that reflect the ideal answer. A KPI generated by the system could be completely valid, but due to it not being exactly as in a potential ground truth set, it would be considered incorrect by the scoring system.

Another recurring issue would be the annotator bias. This, however, would be especially prominent given the subjective nature of producing a suitable KPI for a specific objective. There is no way of determining a ground truth, and therefore a case-by-case approach for validating is the only reliable option. The evaluation should ultimately address the usefulness of the system, which most automatic metrics typically are incapable of, and for which human evaluation still remains the undisputed dominant approach [30]. Therefore, evaluation by the means of human feedback from relevant people was considered the best alternative. Specifically, this was achieved by formalising surveys, one for each collaboration described by an agreement. These could then be sent out to the appropriate parties. To ensure relevance, the evaluators were selected by searching through the agreements to find lead-scientists, scientists, or principal investigators related to the project. The CRM system was also utilised in the search for appropriate assessors.

One survey was created for each agreement. The evaluators received the questionnaire by email based on their involvement in the project to which the agreement corresponded. Since these people had a direct link to the collaborations, their judgement regarding the model's proposed KPIs of the same project could be deemed more trustworthy. For the evaluators to better grasp the meaning behind this request, background information about the PoC was also provided.

Each survey consisted of all the KPIs of a project, generated by the LLM-based system. For each KPI, the evaluators had to reflect on it in terms of four criteria. They were asked to answer whether the KPI was relevant, clear, actionable, measurable, or neither. The criteria were explained to the respondents as explained in Table 5.2.

In addition, evaluators were also asked to comment on a KPI if they felt the need to express some other reflections on the KPI. Furthermore, general questions were posed at the very end, which can be viewed in Table 5.3.

Table 5.2: The table shows the meaningfulness criteria and their respective definition as it was described in the evaluation surveys.

Criteria	**Definition**
Relevant	Do the KPIs reflect what matters for this collaboration?
Clear	Are the KPIs clearly defined and unambiguous?
Actionable	Would these KPIs support decision-making and tracking?
Measurable	Are the KPIs possible to measure given existing data today, or data that could exist?

Do you consider that the KPIs cover the whole problem and measure everything that should be measured when evaluating the health of the collaboration?
Are there any KPIs you would add which are not included in the stated KPIs?
Do you have any general comments and/or recommendations?

Table 5.3: The table shows the general questions that were asked at the end of every survey. The questions do not pertain any single KPI but rather the composition of all proposed KPIs for a given collaboration.

The final general questions assess how well the predictions address the problem overall and whether the agreement's KPIs collectively provide complete coverage of all relevant aspects of what can and should be measured by KPIs.

6
Results

This chapter presents the results of each step in the developed system. Therefore, the outcomes of the evaluation are divided accordingly into the following sections: Objective Extraction and KPI Definition. These findings stem from the methodological process described in the previous chapters. Due to the confidential nature of the data used to generate the results, no concrete examples of objectives or KPIs derived from proprietary contracts can be presented. Instead, the illustrations shown are based on publicly available contracts. The KPIs generated from these public sources are unvalidated and are provided solely to demonstrate how the results would appear if the proprietary data could be disclosed. Because the survey responses refer to confidential AZ contracts, portions of their responses have been redacted.

6.1 Performance of Objective Extraction

The ability to identify and extract objectives from the documents has been evaluated in terms of the metrics specified earlier (see Section 5.1). To ensure that the results were not skewed in any direction, a multitude of embedding models were tried for the evaluation of the objective extraction. When comparing the performance metrics with differing backbone selection, the alternative yielding the highest score and limited standard deviation was ultimately picked, resulting in the use of *all-MiniLM-L6-v2*. Details regarding backbone selection are left to be read in Section 5.1 and Appendix B.

This objective extraction procedure was performed on both public data and proprietary data from AZ. The AZ data could also be subdivided based on whether the project was conducted with an academic or a business collaborator. The performance on each of these document types can be seen in Table 6.3. The row marked with *Business [long]* refers to the data where, for extremely long contracts, the whole document has been processed. This is in contrast to processing only the segments that were deemed to be more relevant. The reasoning behind these approaches can be read in Section 4.1.1.

Similarly to during the annotation process, the outcome is a set of identified text snippets considered to reflect the overall objective of a given collaboration project. The difference is that now the LLM is tasked with doing this. To illustrate the results of the objective extraction process, an example based on a publicly available

contract is displayed in Table 6.1. As such, it is possible to compare the two sets.

The indices show which segment from each set that is mapped to one another. Their numbering is based on the order the snippet was identified in the text. In the subsequent Table 6.2, the remaining snippets are presented. The matching algorithm is exhaustive, meaning that it maps elements from each of the sets until one is empty. In this case, one can observe that more objectives were identified by the annotators than by the LLM. The fact that it is the objective with index 0 also explains why the matched indices are staggered by one in Table 6.1.

Indices	**Prediction**	**Gold Standard**	**Metrics**
(1, 2)	closer economic and industrial integration of the Participants in sustainable value chain of raw materials	closer economic and industrial integration of the Participants in sustainable value chain of raw materials	P=1.000 R=1.000 F1=1.000 cos=1.000
(6, 7)	cooperation on skills, capacity building and competences necessary for the development of sustainable raw materials value chains, including the promotion of the most sustainable extraction and transformation practices, and circular economy	cooperation on skills, capacity building and competences necessary for the development of sustainable raw materials value chains, including the promotion of the most sustainable extraction and transformation practices, and circular economy	P=1.000 R=1.000 F1=1.000 cos=1.000
(2, 3)	cooperation to increase resilience of raw materials value chains	cooperation to increase resilience of raw materials value chains	P=1.000 R=1.000 F1=1.000 cos=1.000
(4, 5)	the development of open, resilient and competitive markets for raw, processed and recycled materials, allowing the EU to diversify its suppliers for materials necessary in particular to achieve the clean and digital transition and its open strategic autonomy	the development of open, resilient and competitive markets for raw, processed and recycled materials, allowing the EU to diversify its suppliers for materials necessary in particular to achieve the clean and digital transition and its open strategic autonomy	P=1.000 R=1.000 F1=1.000 cos=1.000

Continued on next page

Indices	Prediction	Gold Standard	Metrics
(5, 6)	promoting the alignment of sustainable raw materials value chains developed between the EU and the Argentine Republic with internationally agreed principles and guidelines for environmental, social and governance (ESG) standards	promoting the alignment of sustainable raw materials value chains developed between the EU and the Argentine Republic with internationally agreed principles and guidelines for environmental, social and governance (ESG) standards	P=1.000 R=1.000 F1=1.000 cos=1.000
(7, 8)	facilitate closer cooperation on research and innovation along the raw materials value chain, including advanced exploration, earth observation, innovative extractive, processing, refining and recycling technologies	facilitate closer cooperation on research and innovation along the raw materials value chain, including advanced exploration, earth observation, innovative extractive, processing, refining and recycling technologies	P=1.000 R=1.000 F1=1.000 cos=1.000
(0, 1)	identifying and jointly developing innovative and sustainable and responsible raw materials value chain projects by facilitating business opportunities, deploying financial support, investment de-risking instruments	identifying and jointly developing innovative and sustainable and responsible raw materials value chain projects by facilitating business opportunities, deploying financial support, investment de-risking instruments	P=1.000 R=1.000 F1=1.000 cos=1.000
(3, 4)	developing the Argentine Republic's sustainable raw materials value chains in its environmental, social and economic dimensions as a lever for a sustainable and inclusive economic growth, the creation of local added value, quality employment, the development of local industrialization and domestic revenue mobilisation	developing the Argentine Republic's sustainable raw materials value chains in its environmental, social and economic dimensions as a lever for a sustainable and inclusive economic growth, the creation of local added value, quality employment, the development of local industrialization and domestic revenue mobilisation; thereby increasing the competitiveness of the Argentine economy	P=1.000 R=0.895 F1=0.944 cos=0.988

Continued on next page

Indices Prediction	Gold Standard	Metrics

Table 6.1: Comparison between text segments identified as objectives, from the resource agreement [29] between the EU and Argentina. In the left column, the text corresponds to predictions by the LLM, while the right column displays what was deemed to be the most correct text when annotating the dataset.

Index	Gold Standard
(0)	deepen cooperation in the field of sustainable raw materials value chains that support the clean energy and digital transition

Table 6.2: The table lists a set of annotated objectives that have not been paired with an LLM output, for the EU-Argentina resource agreement [29].

The performance of the first stage can be summarised in Table 6.3 for all types of agreements. The scoring is done with the same metrics as previously. Each metric is separated into two distinct measurable scores; one in which all identified objectives from each respective set have been paired, and another in which the remaining unpaired objectives are considered as well. The unmatched case could be considered as a penalised version of the matched case (see Section 5.1). This is because by taking more unidentified objectives into account, the denominator increases, which pushes the quotient down. This explains why the penalised result is constantly lower for all metrics than its matched counterpart. Notable is how consistently the score is the best in the matched case for academic documents. For the penalised scores, the resource agreements fair better. When it comes to *Business* compared to *Business [long]*, the former performs better than the latter, significantly in the penalised cases. The only time the roles are reversed is for matched cosine similarity, where *Business [long]* scores slightly higher than *Business*. Nevertheless, both achieve results that are lower in general compared to *Academic* and *Resource*.

Document type	No. of files	Precision		Recall		F1		Cosine Similarity	
		Match.	*Pen.*	*Match.*	*Pen.*	*Match.*	*Pen.*	*Match.*	*Pen.*
Academic (AZ)	15	**0.867**	0.569	**0.830**	0.548	**0.839**	0.553	**0.891**	0.580
Business (AZ)	3	0.728	0.568	0.709	0.545	0.708	0.547	0.699	0.532
Business [long] (AZ)	3	0.708	0.318	0.690	0.288	0.676	0.293	0.723	0.275
Resource (Dummy)	7	0.807	**0.642**	0.810	**0.633**	0.793	**0.626**	0.881	**0.688**

Table 6.3: The average performance in terms of the evaluation metrics is showcased by document type. The underlying backbone that is utilised for this is the *all-MiniLM-L6-v2*. The best scores per metric are highlighted in bold.

6.2 Performance of KPI Definition

The results are based on the surveys consisting of quantitative yes or no questions, which aim to rate in terms of predefined criteria, as well as qualitative comments. Hence, this section is divided into a part on quantitative and qualitative results, respectively. In total, 101 KPIs were evaluated, spread over seven evaluators who responded to the survey.

6.2.1 Quantitative Survey Results

In total, there were eight answers to the surveys. One person was involved in two separate projects, thus responding to two different surveys. For both these questionnaires, the respondent answered by marking all proposed KPIs as irrelevant, except for two and three instances, respectively. The other surveys did not reflect responses as one-sided as this.

In Table 6.4, each agreement has had the KPIs analysed by calculating the average score over all KPIs in terms of the meaningfulness criteria (see Section 5.2). The *Average* column shows the average score for the specific criteria over all the KPIs in all documents. For instance, the KPIs of agreement number eight were 8% relevant on average. The average relevance over all agreements was then calculated to be 41%. This is performed for all meaningfulness criteria. *Not Applicable* considers whether that box has been checked in the survey or not. If selected, it signified that none of the other four criteria were satisfied for a proposed KPI. *Applicable* is calculated by $Applicable = 1 - (NotApplicable)$, assuming that all KPIs not marked as *Not Applicable* can be considered *Applicable*.

Table 6.4: Averages over all KPIs within agreements by meaningfulness criteria. The numbers correspond to different agreements that survey respondents gave feedback on.

Criteria	1	2	3	4	5	6	7	8	Average
Relevant	0.30	0.67	0.54	0.46	0.17	0.53	0.54	0.08	0.41
Clear	0.10	0.67	0.62	0.38	0.00	0.40	0.54	0.31	0.38
Actionable	0.10	0.67	0.23	0.46	0.00	0.67	0.46	0.08	0.33
Measurable	0.10	0.67	0.69	0.46	0.00	0.67	0.69	0.62	0.49
Not Applicable	0.70	0.17	0.23	0.54	0.83	0.20	0.23	0.23	0.39

In Table 6.5, the evaluation has instead been divided into the different categories of KPIs deduced from the interview process (see Table 3.1). This is done to better examine which of the general and specific KPIs the respondents consider relevant, clear, actionable, measurable, or not applicable. Collaboration-specific KPIs are grouped into one category as they differ between each project.

Table 6.5: The table presents which traits of the meaningfulness criteria the evaluators found the KPIs to have. The scores are ratios of how many KPIs that were selected as having a given trait. KPIs are grouped by categories described in Chapter 3. The best value per criterion is highlighted in bold.

Category	Relevant	Clear	Actionable	Measurable	Not Applicable
No. of patents filed	0.50	0.38	0.38	0.38	0.50
Follow-up research	0.50	0.50	0.50	**0.75**	0.25
No. of scientific publications	0.50	0.50	0.50	0.50	0.50
Impact factor of publication	0.63	0.63	0.50	**0.75**	0.25
Reputation of publication journal	0.33	0.33	0.17	0.33	0.67
External engagements	0.75	0.38	0.38	0.50	0.13
External funding	0.29	0.57	0.43	0.57	0.43
Alignment with company goals	0.50	0.00	0.50	0.00	0.50
Budget coherence	**0.88**	**0.75**	**0.63**	0.63	**0.00**
No escalations to leadership	0.00	0.38	0.13	0.25	0.63
Recruitment of Post-Doc	0.25	0.38	0.38	0.63	0.38
Collaboration-specific KPIs	0.30	0.20	0.23	0.5	0.4

6.2.2 Qualitative Survey Results

As part of the human evaluation, the participants were asked to write qualitative comments. Table 6.6 shows the answers to the general questions about the KPIs collectively. The responses reflect mixed reviews regarding the quality of the KPIs. They also provide some insight on the project's relation to KPIs and their measurement possibilities. Table 6.7 instead presents the comments that the respondents made on the KPIs while quantitatively evaluating them. It works as a compliment to the scoring, as it specifies some cases of why a given KPI is considered good or bad.

Table 6.7: The table presents the written answers corresponding to the different kinds of KPIs. Every KPI category is not present as all types of KPIs did not receive qualitative feedback. Specific information relating to the projects have been redacted. If a KPI category is not displayed in the left column, the previous category still applies.

KPI	Answers
Number of patents	– We don't expect any patents to be filed from our collaboration, thus not checking the 'relevant' box. A more appropriate measure along the same lines would be scientific publications. – Not a goal but at least clear, here number has more value as it goes though an evaluation process – It´s not likely that any patents will be filed, but the KPI is valid.

Continued on next page

KPI Category	Answers
Follow-up projects	– It would be an indirect measure of the success of the first project, but wouldn't necessarily tell the whole story. The project could have been great but run its course. Or funding might have run out even though there's a want from the people involved to continue.
Impact factor of publication	– Number of publications, which would strike me as the first obvious KPI, seems to be absent as a KPI alltogether.
Reputation of publication journal	– We always use impact factor as the metric for journals we publish in. I don´t know what a reputation score is.
Alignment with company (%)	– Don't think it either relevant or in any way quantifuable/measurable.
No escalation to senior leadership	– not sure what this means/refers to – Define escalation - I assume concerns or issues escalated? – It´s a rather odd KPI; does it mean you are successful in performing your research without ethical or data integrity misconduct? Or what?
Number of publications	– Good evaluation of academic collaboration, often takes time though and may not be completed within contract time. – Number of research results" isn´t really a well defined metric. What constitutes "a result"? If the KPI is supposed to capture "Number of publications in peer reviewed journals" it´s a relevant KPI.
External engagement	– External engagements" is rather vague. "Number of posters and presentations at external conferences/workshops" would be a better KPI defined KPI, in my opinion
Collaboration-specific KPI	– The time scale for getting on this would be years/decades, so not really measurable in practice I would say. – Too wide of a scope. – Not sure if this can be actionable since we cannot influence what [REDACTED] – It's unclear what [REDACTED] means – we included this as a QC check and conditional element to the collaboration

Continued on next page

KPI Category	Answers
	– i think there is something here but how do you define "accuracy and consistency" – Number of algorithms does not necessarily capture impact completely. Number of houses: three sheds and a palace... – Same as in KPI 1 – Number not good measure, perhaps functional integration of new QC/analysis modules from collaboration into internal pipeline. – Again, would focus on integration or maybe. Again, this is relevant but easy to check box without impact. – did the collaboration result in the development of a useful FM for feature extraction. Number has little meaning – I presume you could count the number of models tested, but that by itself has no meaning. The important thing is if the models are relevant in the context. – Measurable, but like the "number of mathematical models" KPI the number itself has limited relevance, it´s the impact that is important. – I am uncertain about what data this KPI is supposed to be calculated from. Is it peer reviewer feedback from journals the work is submitted to? Scoring on these aspects on research grant requests submitted based on the data? Or what?

Table 6.6: This table presents answers to the general broad questions of the survey.

Questions	Answers
Do you consider that the KPIs cover the whole problem and measure everything that should be measured when evaluating the health of the collaboration?	– No – To some extent – Largely covers it – Some happen sooner (data generation) and some later (publications, etc) so KPIs might not be time bound – No
Are there any KPIs you would add which are not included in the stated KPIs?	– Number of publications, or total impact/citations etc. – Not sure – This is a collaboration in R&D, specifically early development - KPIs in relation to new target identification, positive governance interactions would seem salient – Again, focus on integration of methods vs PoC or number of algorithms. One can develop 100 algorithms in an afternoon but none of them are useful... – Yes, several. E.g. "Number of AZ projects supported by the new models developed"
Do you have any general comments and/or recommendations?	– The suggested KPIs seems to have a more late stage product focus than basic, early science. – You are on the right track, explore how better define quality and impact of new algorithms or digital deliveries – Identifying measurable KPIs that show relevant impact on how we do our business is challenging in general. In my experience we tend to prioritize "measurable" over "relevant"...

7
Discussion

This chapter aims to dissect the meaning behind the results and try to package a recommendation on how to maximise the utility of the implemented system as well as how to apply it in a larger setting. However, the analysis of the outputs may be limited to general discussions, and the presented examples may have been altered to not reflect confidential information.

7.1 Objective Extraction

To begin with, the objective extraction capabilities are quite good. Even if processing large quantities of text is what an LLM is designed to do, it is still quite impressive that the same task can be accomplished multiple times to a fairly high level of accuracy, given the variation in the data it is exposed to.

For the largest set of documents, namely the ones with academic collaborators, the results were the best for the matched case. For the unmatched, the performance metrics were, while not the best, still very similar to those of the business agreements. In the unmatched category, the resource agreements (dummy data) performed the best instead.

What is not presented in the results section of this report is the ratio of unmatched objectives to the exhausted set of objectives. The example shown in Table 6.2 infers that the exhausted set is the one with the predictions. N.B., this is not always the case, as it can sometimes be the other way around. Although, from looking directly at the data, it can be concluded that in the vast majority of instances, the unmatched objectives left are from the prediction set, i.e. there is a surplus of candidates that leads to these low scores. Albeit, in some cases, it can also be that the remaining objectives after matching are from the gold standard set. This entails that, at times, objectives simply are not identified to the extent that they ideally would. Regardless, the objectives that have been identified are most often correct, indicated by the matched metrics. This reasoning can be applied to both academic, business, and resource collaborations.

The performance on the *Business [long]* agreements was substantially worse across all metrics, in comparison to the *Business* where only a selection of relevant excerpts were taken into account as business agreements. As expected, the preprocessing procedure has a massive effect on the results. The view that the discrepancy stems from

a selection bias may be correct, although it seems even more likely that simply too many candidates for objectives were identified when processing the extremely long documents in full. Piece-wise processing of the long contracts has without question produced more objectives than in the gold standard. Their length likely affects the LLM's ability to sift out the correct objectives. A most possible cause for this is that the system forces itself to extract objectives from the provided excerpt, even when that piece of text does not necessarily contain the relevant information. The LLM is prompted with finding the overarching objective of the collaboration, which naturally is skewed when the whole frame is narrowed down.

The fact that more candidate objectives were found explains why the penalised metrics become much worse than otherwise. Similarly, one would guess that the matched metrics would increase, or at least stay the same, as the certainty of finding correct objectives rises with more candidates available. This turns out to be quite the opposite. One explanation could be that with more identified objectives, the risk of multiple candidates having similar enough embeddings to be incorrectly paired increases. Similarly, there is the chance that objectives may be rephrased in slight variations, which might lead to different behaviours in the matching algorithm. Potential duplicates also affect the penalised score negatively, as they are seen as separate objectives. Nevertheless, for these reasons, this processing approach will not be recommended to be used in a full-fledged future system.

However, regardless of the results, there are a multitude of reasons why the findings can be scientifically problematic. Firstly, there is a probability that the LLM's ability to identify objectives surpasses that of our own in some cases, given that we are not domain experts annotating the data. The obvious solution is to let experts perform the labelling. Still, the notion of what constitutes an objective is not always clearly defined. There are instances where it is up to interpretation whether the text at hand should be considered as an objective or not. Thus, variations in which approach is pursued heavily affect the evaluation score without necessarily leading to substantially altered effective results.

As the work presented in this report is for a PoC, it means that some simplifications and deviances from an optimal case are allowed. Therefore, from a purely functional aspect, a tendency of the model to extract additional candidate objectives to the gold ones can be a benefit, even if it may be reflected with a lower score in terms of the evaluation metrics. As long as the candidates are still extracted from the document, and are not hallucinations, it means that more KPIs are proposed to the human evaluators. Thus, additional potentially valuable insights are therefore presented.

7.2 KPI Definition

In general, the quantitative results are very mixed. Most of the respondents answered that only some of the KPIs presented are *Not Applicable*. However, in three

of the surveys, respondents answered that less than half of the KPIs were applicable. On average 61% of the KPIs were considered applicable, which is not a good result, since it suggests that 39% of the presented KPIs were neither relevant, clear, actionable, nor measurable. One can argue that all four criteria need to be met for a KPI to be usable in practice. This is something that rarely happens based on the survey results. If we consider the respondent who answered two surveys completely negatively as an outlier, and disregard those responses, the average applicability increases to 79% instead. This can be interpreted in two ways: Either the respondent was particularly negative towards the presented KPIs, or the specific projects of this respondent simply have a different way of measuring performance that does not correspond to the other projects, and how the LLM defines the KPIs in this case.

Furthermore, when looking at the different performance for each criteria in Table 6.4, there is a similar mediocre result. None of the averages surpass even 50%. This result is also worsened by one respondent scoring two surveys very negatively. The most important criteria is arguably relevance, since it is a constraint that must be met for the KPI to hold any fundamental value. Similarly, measurablility is important, as an unquantifiable KPI is not much of an indicator. One respondent phrased that they even tend to prioritise measurablility over relevance, as seen in Table 6.7. Regardless of their ordering, these two criteria are the ones with the highest scores with 41% and 49%, respectively. In total, the average scores indicate that the respondents, who ultimately are the people most involved in the projects, do not consider most of the KPIs useful. This is a strong signal that this stage of the developed solution is not sufficiently effective to create useful KPIs.

Looking at each individual KPI in Table 6.5, there are more insights from the survey responses. It is clear that most of the general KPIs, i.e. the ones that were suggested from auxiliary interviews, are the best performing KPIs compared to the collaboration-specific ones that were based on the contracts. *Number of patents filed* scores a relevance of 50%. This KPI's relatively low score can be explained by three evaluators pointing out that patents are not an expected outcome in any of their respective projects. Therefore, the KPI can be useful, but not in all cases. This finding is helpful as it implies that the general KPIs might not be as general as assumed.

Impact factor of scientific publication has a higher relevance score of 63%, while *External engagement* and *Budget coherence* have even higher, with 75% and 88%, respectively. In particular, the first could be specified more clearly according to the feedback from the surveys. *No escalations to senior leadership* has a relevance score of 0%. Many respondents expressed confusion regarding this KPI, saying that it is odd and unclear. Hence, an improved explanation of its meaning could be beneficial. One of the purposes of the LLM is to take these general KPIs and tailor them to specific projects. The findings demonstrate that this ability is suboptimal. This can best be summarised by the evaluators expressing that some of the general KPIs were unclear or unspecific, as seen in Table 6.7. The reason could be related to not enough context being provided to the LLM for it to produce sufficiently meaningful outputs. See Section 4.1.2 for more information about why additional context is left

out.

Further problems are showcased when looking at the collaboration-specific KPIs, which are based directly on the extracted objectives. As seen in Table 6.5, the scores for the specific KPIs are significantly lower than for the general KPIs. This suggests that the model has failed to translate the objectives into useful KPIs, in terms of the meaningfulness criteria. Although, there is one positive remark from the surveys, stating that the one of the proposed KPIs is actually one use for quality control. Even though the rest of the comments indicate problems with the KPIs, this is still a sign that useful results are achievable.

When it comes to the problems, they include that the KPI does not consider the time scale. In some cases, it could take years or decades to measure the KPIs, according to some respondents. This removes the main purpose of this case, which is to continuously measure collaboration health while it is ongoing. Other comments say that the KPIs are unclear, too broad, or that they need clearer definitions of things like "accuracy" and "consistency". For some instances, the model will simply translate an objective to a KPI by adding "Accuracy of ...", which does not always create a valuable KPI and causes unclarity. Another problem that was presented by the reviewers is quite alike: If the objective states "develop models of X", then the KPI may be defined as "number of models of X developed".

Since the term *model* encompasses a wide range of applications, the number of models alone does not constitute a meaningful or comparable measure. Furthermore, the proposed KPI focuses exclusively on quantity and does not account for model quality, making it largely non-actionable. One respondent highlighted this limitation by noting that the “number of algorithms does not necessarily capture impact completely,” comparing it to counting “three sheds and a palace” as equivalent houses. This analogy illustrates that substantial differences in value can exist within the same nominal category, which are differences that a well-defined KPI should capture. Additional responses reinforce this concern, noting that “one can develop 100 algorithms in an afternoon but none of them are useful,” and that the suggested KPIs seem to have a stronger "late stage product focus than on basic, early science” (see Table 6.6).

In addition to the comments provided on specific KPIs, the surveys also included general remarks, which are presented in Table 6.6. When asked whether the defined KPIs adequately cover the complete problem, the responses were mixed. Two respondents answered “no,” one indicated that the KPIs “largely cover it,” and another stated that they do so “to some extent.” These qualitative responses are consistent with the mixed quantitative results, where, in most cases, approximately half of the KPIs were evaluated as meeting the specified criteria, while the remaining half were not.

This variability suggests that the perceived value of the KPIs differs substantially between evaluators. It remains unclear whether these differences stem from incon-

sistent model performance or from divergent interpretations of what constitutes a good KPI in terms of relevance, clarity, actionability, and measurability. Additionally, perceptions may vary depending on the specific project under evaluation. One respondent further noted that KPIs may manifest at different stages of a collaboration, distinguishing between continuous indicators of ongoing progress and later-stage KPIs that retrospectively validate project outcomes. While retrospective KPIs can clearly demonstrate measurable value—such as scientific publications, continuous KPIs are necessary to assess the current health of a collaboration, which aligns more closely with the operational needs at AZ.

In the second general question, some of the evaluators respond that multiple KPIs are missing, giving examples of ones they believe should be included. This suggests once again that defining KPIs is quite a subjective task with no clear answer. This is further strengthened by one respondent stating that they themselves are not sure which KPIs they would include. If the people who work on the project are uncertain about which KPIs to measure, it is very unlikely that an LLM can extract KPIs that fully satisfy the needs of people involved in the project.

Another problem that is seen in the KPIs is the exact definition of numbers. In some cases, the LLM includes numbers in the proposed KPI, e.g. "number of scientific publications within 12 months." Here, the specific quantity of 12, which discloses the time duration, is unmotivated. It is a subjective quantity that the LLM generates without a clear reason for why, since instructions on that matter have not been provided. This can without a doubt strongly affect the usefulness of the KPI.

While the assessment of general KPIs shows promise, particularly when only those KPIs with high relevance score are considered, the responses suggest that KPI definition is, in many cases, too subjective to be reliably extracted by an LLM without additional contextual information. For example, aspects related to time dependencies and project specific complexity are difficult to infer from the agreements alone. Moreover, effective KPI definition may require a deeper understanding of the underlying science, research objectives, and methodological context than is typically contained in contractual documents or what an LLM can derive from them.

7.3 Value of System for AZ

This thesis has from the start had the ambition to be of help to AZ, exploring what needs they have, which capabilities they possess to go about the task, and what solution may suit this endeavour. The PoC that has been developed throughout this project has aimed to address a piece of this. The subsequent text goes into further detail regarding what value the solution and its findings bring as well as the potential value they can bring.

7.3.1 Analysing Different Kinds of Agreements

A key insight from the conducted interviews is that there are fundamental differences between varying kinds of collaborations. Academic agreements tend to be open with more unspecified goals. Even if there is a project plan, the objective is often to keep conducting science and investigating what it might lead to. In contrast, business and commercialisation agreements tend to have clearer objectives and milestones that need to be fulfilled for the project to continue and for payments to be made. In commercialisation agreements, there are also distinct financial goals that need to be met for the collaboration to continue. One can assume that commercialisation is a concept that is present in a later stage than academic research, which often takes place years before any actual revenue may even be created. This means that varying types of agreements include different kinds of objectives, which may or may not be more clearly stated.

Because objectives and aims are more clearly defined in business and commercialisation agreements, the methods proposed in this thesis are likely to perform more effectively in such contexts. This hypothesis comes with the caveat that KPIs have not been evaluated for business agreements, due to organisational restrictions. Although, it seems reasonable that with clearer targets, the LLM is required to make fewer assumptions when defining KPIs, which reduces the risk that the extracted KPIs diverge from what stakeholders intend to measure. In commercialisation contracts, relevant metrics, such as financial performance and revenue generation, are more explicitly stated, making them more inclined for automated KPI definition.

However, for such improvements, the LLM's ability to extract relevant pieces of information from business agreements needs to improve quite dramatically, considering the results that were presented in Table 6.3. If the stringency of commercial contracts comes at the price of the documents being so long that they reduce the LLM's ability to isolate the important segments, there might not be much advancement to be made without either bettering the LLM or formulating the contracts more concisely and cohesively. Then again, too few contracts have been evaluated to know whether this is the rule or the exception.

Lastly, even if this research mainly addresses a small collection of academic contracts and a handful of business agreements, with mixed results, there might still be an additional use-case for AZ on other kinds of documentation. This can of course be evaluated further, assuming access to this data is provided.

7.3.2 Alternative Utilities of LLM Text Extraction

As shown by the results of Table 6.3, the LLM can extract the objectives from the agreements with a quite high degree of certainty. The prevailing issues occur mainly in the subsequent step of the designed system, where objectives are translated into KPIs. The second stage is still promising, but more has to be done for it to produce the results that are sought. Considering that the first part of the pipeline is fairly

successful, it allows for that solution to be utilised in other applications to similar problems in the meantime.

One challenge related to AZ's data management of agreements is the insufficient use of metadata in the CRM system. Metadata links agreement and collaboration to relevant information, such as lead scientists, budget, important dates, and collaboration partners. Throughout this thesis, the need for a more robust and consistently applied metadata structure has been repeatedly identified. The reason for this is that such metadata is critical for conveying project context and the broader operating environment, thereby enabling a clearer understanding of AZ's collaboration ecosystem and supporting well-informed insights.

During the interviews, it was noted that addressing metadata deficiencies represents one of the first steps toward automating assessments of progress and collaboration health. Stakeholders expressed a need to answer high-level questions, such as how many projects are currently ongoing with a specific university or how research funding is distributed geographically over time. While these questions do not require advanced ML techniques, they do rely on comprehensive and structured data collection. This, in turn, requires that relevant information is consistently captured and stored within the CRM system.

One potential solution is to define mandatory metadata fields that must be completed when creating new CRM entries. A likely reason for the current prevalence of missing data is that completing metadata fields is not prioritised by users. Manual data entry is time-consuming and is therefore often overlooked when fields are optional. Furthermore, the lack of clear incentives or demonstrated value for completing metadata reduces user motivation to contribute to and maintain a comprehensive database.

An alternative approach to addressing this issue is by leveraging the objective extraction method presented in this thesis. The system would be expanded to extract, not just objectives, but all metadata relevant to understanding and analysing the AZ collaboration ecosystem, thereby supporting data-driven decision-making. Given that the model showed promising results when extracting objectives from agreements, it is reasonable to expect that other metadata elements, such as lead scientists or project timelines, could also be extracted when they are explicitly stated in the documents.

Together, these two solutions could complement each other well and ensure that correct metadata is stored in the database. By implementing the objective extraction as a sort of recommender system on mandatory data fields, the process of completing the database can be facilitated. When a user is tasked to fill in a data field, the system presents suggestions retrieved from the corresponding document for that particular field. The user, knowing what the correct entry should be, could either select a suitable suggestion or provide a manual entry if none of the recommendations are appropriate. This introduces a direct human validation step, ensuring that

the stored data is verified for correctness. At the same time, data on the accepted and rejected recommendations could be collected over time and used to fine-tune the model, potentially improving extraction performance and recommendation quality in the long run.

Building on the proposed solutions for metadata extraction and validation, it becomes clear that access to structured data is a prerequisite for automated reporting. Limited access to data has been a significant constraint throughout this project, and ensuring that relevant information is available as standardised metadata in a centralised database would address many of the identified limitations. Ultimately, it would enable more reliable automation of reporting and insight generation.

7.4 Sources of Errors and Problems

There have been multiple shortcomings and causes for problems throughout this project. Most of these are highlighted in the limitations (see Section 1.3), however, this part tackles in what way these causes for concern can become issues.

7.4.1 Lack of Data

The big constraint of this project is the limited access to data. This caused the scope of the implementation to only process a small number of agreements, and validate the final output on even fewer documents, namely 8 out of the 15 academic ones. This restricted data availability makes it difficult to fully assess both how well the system processes the documents to both extract objectives from the agreements and how well these objectives are translated into KPIs.

The small sample size introduces a significant risk of bias in the evaluation. This is especially evident in the human assessment of the generated KPIs, where only seven individuals completed the survey. With such a limited number of responses, each distinct interpretation can disproportionately influence the results. For instance, misunderstandings of the survey questions or differing interpretations of of the evaluation criteria may affect the outcome substantially. Additionally, some of the projects may for some reason not be well suited to KPI-based assessment, which can lead to poor results even if the underlying model performs adequately in general.

Beyond the quantity of data, the project would also benefit from access to a broader range of agreement types. Analysing a more diverse and representative set of agreements would make it possible to assess in which contexts the proposed implementation is effective and where it is not, given that differing documentation may vary considerably in how well they align with the system. Even if three types of contracts were processed, only academic agreements were fully evaluated (both extraction and generation). As mentioned in Section 3.1, academic agreements tend to be open-ended, which could cause the objectives to be less clearly defined compared to business or commercialisation agreements. This general characteristic could there-

fore risk influencing the system's ability to extract objectives and define meaningful KPIs negatively.

The shortage of accessible data also made it infeasible to develop a method for measuring the ongoing progress and health of collaborations. One key reason for this, as highlighted in the interviews, is the lack of structured and consistently reported project data. In scientific collaborations in particular, the projects are often managed and reported in a dynamical and informal manner, with substantial trust in the researchers to advance their work. For the full system proposed in this thesis to function, from input of agreements and progress data to collaboration health assessment, structured and regularly updated data is required as a foundation for analysis. As AI-based systems are inherently dependent on the quality of their input data, the feasibility of automated reporting ultimately hinges on the availability, structure, consistency, and completeness of the underlying information. Consequently, critical obstacles to establishing a high-performing LLM-driven solution lie in current reporting practices, the nature of open-ended research projects, and the conceptual differences between objectives and performance indicators.

7.4.2 Subjectivity of KPIs

When it comes to objectives, they are clearly defined. KPIs, however, can be translated from the objective in multiple ways. Objectives are qualitative, while KPIs are a quantitative measurement. Taking that into account, broader objectives specified in the agreement between the collaborating parties still have to be translated into measurable performance. To ensure consistency and comparability, broader objectives must be translated into KPIs according to predefined principles. Without such constraints, measurability and comparability across projects become difficult. This reasoning prompted the introduction of the general KPIs. As the result indicate, these general KPIs were evaluated more favourably than the specific KPIs derived directly from objectives.

The inherent flexibility in KPI definition presents a fundamental challenge. For any given objective, there may be numerous valid quantitative interpretations. Due to the black-box nature of the LLM, it is not fully possible to understand why a certain formulation was made. For example, when defining a KPI such as "scientific publications within 12 months", it remains unclear why a 12-month time frame was chosen rather than any other valid duration, like 8 months or 3 years. In theory, an infinite number of KPI formulations are possible. This variability impedes the system's possibilities to consistently generate high-quality, contextually appropriate measurement indicators. A possible improvement could therefore be to constrain the model more strongly through better prompt engineering or predefined KPI templates.

Furthermore, to define KPIs that are genuinely useful for project stakeholders, contextual knowledge from within or even outside the agreement, other than the objec-

tives, is evidently required. For instance, whether a project aims to apply for patents may not be formally articulated in the contract, but it is nevertheless understood by the individuals involved. Hence, some general KPIs are not universally applicable across projects. The LLM cannot infer such implicit intentions when they are not documented, which represents a limitation of the current system design.

The current implementation focuses only on extracting objectives, however, relevant information for KPI development may reside outside of these. Budgetary or financial details, for example, may appear elsewhere and could provide valuable input for defining meaningful performance indicators. Expanding the system to incorporate additional structured information, both within and beyond the extracted objectives, could be key to improving its performance by generating more relevant, clear, actionable, and measurable KPIs.

7.5 Future Research

Suggestions for future research include exploring the alternatives proposed in Section 7.4. These include testing the implementation on a larger and more diverse set of agreements, as well as expanding the system to incorporate information beyond objectives when defining KPIs. Additionally, establishing clearer frameworks, stricter prompt-engineered constraints, or predefined KPI structures would be encouraged.

Furthermore, the exploratory nature of the current implementation shaped the system design presented in this report, however, this does not mean that it represents the optimal solution. For that reason, developing an alternative baseline system for direct comparison could be of value in assessing the PoC. Such benchmarking would provide a more rigorous evaluation of the proposed approach. Ultimately, once sufficient performance and reliability are achieved, the development of a full-scale system, such as the architecture illustrated in Figure 1.1, would constitute the next step.

7.6 Recommendations for AZ

As discussed in Section 7.3.2, there is a clear need for more thorough metadata reporting to enable ecosystem-tailored insights, which account for the core stakeholder requirement. To enhance the transparency regarding collaboration progress and health requires integrating structured metadata collection with continuous reporting practices.

Therefore, one suggestion is an updated CRM system that not only captures the metadata associated with project initiation (e.g. name, project lead, partner name, location) but also continuously captures progress-related information. This could include mandatory metadata fields, automatically extracted agreement data where

possible, and continuous KPI related metadata as discussed in Section 3. In this way, the general KPIs, which were deemed the most useful according to the survey results, could be captured when relevant.

The new CRM system would enable capturing value of a collaboration on multiple time horizons. As highlighted in the surveys and discussed in Section 7.2, some KPIs measure ongoing progress, whereas others can first be measured once a project has concluded. Capturing both of these values in the CRM system would allow decision-makers to not just follow the progress of current collaborations and the perceived current value, but also monitoring historical project outcomes in terms of these long term KPIs, such as publications or recruited post-docs. In the end, this would create actionable insight and give AZ decision-makers the collaboration overview they seek.

Therefore, another suggestions could be to incorporate the continuous and retrospective KPIs presented in Table 3.1 into the CRM metadata structure and ensuring they form part of regular reporting and follow-up processes. The collaboration-specific KPIs could be defined by using a similar system as described in this thesis, provided that the definitions are clear and the system is improved (see Section 7.5).

Moreover, objective extraction (see Section 4) showed promising results and could be incorporated as a form of recommender system for facilitating extensive metadata completion of an expanded CRM system, while keeping a human-in-the-loop. More about that can be read in Section 7.3.2.

Lastly, the interviewees expressed a need for clearer and more systematic progress overviews. Although the approach introduces additional manual tasks, they may be justified considering the outcome is valuable for project investigators, decision-makers, and AZ as an organisation, while opening up for further automation in the future. Implementing a structured reporting framework would also reduce dependence on the largely unstructured formats currently used, such as emails and PowerPoint presentations. Such formats hinder automated analysis and limit AZ's ability to generate appropriate data-driven insights. Effective automation ultimately depends on structured and consistent data, particularly in this case of efficient progress reporting of the AZ collaboration ecosystem.

8

Conclusion

This work set out to explore whether an LLM-driven solution could first extract objectives from collaboration agreements and then translate them into useful KPIs for continuous progress reporting on collaboration. The outcomes are mixed. On the one hand, objective extraction performed reliably: the model consistently surfaced aims and goals as written. On the other hand, turning those objectives into KPIs that stakeholders deem relevant, clear, actionable, and measurable proved substantially harder. Quantitatively, less than half of the generated KPIs met each criteria on average and 39% of the defined KPIs did not meet any, according to the the responding evaluators. Qualitatively, respondents often found KPIs unclear, insufficiently specific, ill-suited, or focusing on counts that do not capture impact or quality.

At a criterion level, relevance and measurability were the strongest dimensions but still below where they need to be for dependable decision-making, echoing a tension respondents articulated explicitly: in practice, "measurable" sometimes gets prioritised over "relevant." The general KPIs compiled from interviews performed better than the collaboration-specific KPIs derived directly from objectives. This suggests that while a predefined library of general KPIs can provide value, automatic specialisation to project context is not yet reliably achieved by the current approach.

These results reflect deeper structural issues. Academic agreements are often open-ended, and the science evolves dynamically. Objectives are qualitative and not distinctly defined, while KPIs must be quantitative with precise definitions, measurement windows, and data sources. Without strong priors, the LLM will invent specifics that may not align with stakeholder expectations or project realities. It is therefore unsurprising that general KPIs outperformed specific ones based on the objectives, and that perceived usefulness varied across respondents and projects.

Despite these limitations, the work points to practical value for AZ. Firstly, robust objective extraction can be leveraged to improve metadata completeness in CRM by capturing information such as leads, timelines, budgets, and partner details. This would enable ecosystem analytics and automated reporting. Secondly, the pipeline may be better suited to agreements where objectives and milestones are explicitly defined, such as business or commercialisation agreements. In these contexts, LLM assumptions could be constrained by clearer contract language, potentially defining more useful KPIs. Thirdly, the findings motivate a hybrid design: a curated catalogue of general, continuous KPIs for ongoing monitoring, complemented by

retrospective KPIs and limited, template-driven specialisation for project-specific metrics. Tight prompt constraints could improve clarity and comparability.

Ultimately, the central lesson is that KPI definition is a complex task. Useful KPIs require both structured data and shared meaning. The current LLM approach helps reveal objectives and highlights gaps, but it cannot replace the domain alignment needed to agree on what matters, how it is measured, and when. Future work should expand data coverage, include other agreement types, explore hardened constraints, and broaden extraction beyond objectives to critical metadata. With these steps, AZ can progress toward a system that supports continuous visibility into collaboration health while preserving the nuance and rigour that meaningful measurement demands.

9

Bibliography

[1] R. Adner and R. Kapoor, "Navigating the leadership challenges of innovation ecosystems", English, *MIT Sloan Management Review*, Jun. 2021, Online article. [Online]. Available: https://sloanreview.mit.edu/article/navigating-the-leadership-challenges-of-innovation-ecosystems/ (visited on 01/14/2026).

[2] H. Chesbrough, *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Boston, MA: Harvard Business School Press, 2003, Accessed: 2025-09-15. [Online]. Available: https://www.sustanciainfinita.com/wp-content/uploads/2017/03/LIBRO-Henry-Chesbrough-Open-Innovation.pdf.

[3] B. Rake, K. Sengupta, L. Lewin, A. Sandström, and M. McKelvey, "Doing science together: Gaining momentum from long-term explorative university–industry research programs", *Drug Discovery Today*, vol. 28, no. 9, 2023. DOI: https://doi.org/10.1016/j.drudis.2023.103687.

[4] A. Wang, Y. Pruksachatkun, N. Nangia, *et al.*, "Superglue: A stickier benchmark for general-purpose language understanding systems", in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] H. Naveed, A. U. Khan, S. Qiu, *et al.*, "A comprehensive overview of large language models", *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, p. 106, Aug. 2025, Article 106.

[6] L. Martín-Domingo, J. Fernandez Roblero, M. Efthymiou, and M. Ali, "Extracting airline emission kpis from sustainability reports using large language models (llms)", *Transportation Research Interdisciplinary Perspectives*, vol. 33, p. 101 599, Sep. 2025. DOI: 10.1016/j.trip.2025.101599.

[7] AstraZeneca. "Om oss på astrazeneca". Accessed: 2025-09-15. (2025), [Online]. Available: https://www.astrazeneca.se/om-oss.html.

[8] AstraZeneca. "Astrazeneca i göteborg". Accessed: 2025-09-15. (2025), [Online]. Available: https://www.astrazeneca.se/om-oss/verksamheten-i-sverige/goteborg.html.

[9] O. Gassmann and G. Reepmeyer, "Organizing pharmaceutical innovation: From science-based knowledge creators to drug-oriented knowledge brokers", *Creativity and Innovation Management*, vol. 14, no. 3, pp. 233–245, 2005, Accessed: 2025-09-11. DOI: 10.1111/j.1467-8691.2005.00346.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-8691.2005.00346.x.

[10] B. R. Wikhamn and A. Styhre, "Managerial challenges of outbound open innovation: A study of a spinout initiative in astrazeneca", *R&D Management*, vol. 49, no. 4, pp. 652–667, 2019, Accessed: 2025-09-11. DOI: https://doi.org/10.1111/radm.12355. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1111/radm.12355?saml_referrer=.

[11] A. BioVentureHub. "About astrazeneca bioventurehub". Accessed: 2025-09-15. (2025), [Online]. Available: https://www.azbioventurehub.com/about.html.

[12] S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on llm-based multi-agent systems: Workflow, infrastructure, and challenges", *Vicinagearth*, vol. 1, Oct. 2024. DOI: 10.1007/s44336-024-00009-2.

[13] D. Acharya, K. Kuppan, and D. B. Ashwin, "Agentic ai: Autonomous intelligence for complex goals – a comprehensive survey", *IEEE Access*, vol. PP, pp. 1–1, Jan. 2025. DOI: 10.1109/ACCESS.2025.3532853.

[14] C. Choi, A. Lopez-Lira, Y. Lee, *et al.*, "Structuring the unstructured: A multi-agent system for extracting and querying financial kpis and guidance", *Proceedings of The 2nd Workshop on Financial Information Retrieval in the Era of Generative AI (SIGIR FinIR'25)*, vol. 1, 2025, Accessed: 2025-09-24. [Online]. Available: https://arxiv.org/pdf/2505.19197v3.

[15] IBM. "What is deep learning?" Accessed 2026-01-11, IBM Think. (2026), [Online]. Available: https://www.ibm.com/think/topics/deep-learning.

[16] B. Mehlig, *Machine Learning with Neural Networks: An Introduction for Scientists and Engineers*. Cambridge: Cambridge University Press, 2021.

[17] M. Shao, A. Basit, R. Karri, and M. Shafique, "Survey of different large language model architectures: Trends, benchmarks, and challenges", *IEEE Access*, vol. 12, pp. 188 664–188 706, 2024. DOI: 10.1109/ACCESS.2024.3482107.

[18] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[19] S. Srivastava, *A deep dive into the self-attention mechanism of transformers*, Medium — Analytics Vidhya, 10 min read, Sep. 2024. [Online]. Available: https://medium.com/analytics-vidhya/a-deep-dive-into-the-self-attention-mechanism-of-transformers-fe943c77e654 (visited on 01/12/2026).

[20] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) Workshop*, OpenAI, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[21] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language models are few-shot learners*, 2020. arXiv: 2005.14165 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2005.14165.

[22] S. Shahriar, B. Lund, N. R. Mannuru, *et al.*, "Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency", *Applied Sciences*, vol. 14, p. 7782, Sep. 2024. DOI: 10.3390/app14177782.

[23] IBM. "Gpt-4o". Accessed 2025-11-27, IBM Think. (2025), [Online]. Available: https://www.ibm.com/think/topics/gpt-4o (visited on 11/27/2025).

[24] OpenAI. "Hello GPT-4o". Accessed 2025-11-27. (2024), [Online]. Available: https://openai.com/index/hello-gpt-4o/ (visited on 11/27/2025).

[25] L. Wissler, M. Almashraee, D. Monett, and A. Paschke, "The gold standard in corpus annotation", Jun. 2014. DOI: 10.13140/2.1.4316.3523.

[26] Microsoft, *Ai model catalog*, https://ai.azure.com/catalog/models, Accessed: 2026-02-03, 2026.

[27] V. Mavroudis, "Langchain", *Preprints*, Nov. 2024, Alan Turing Institute; contact: vmavroudis@turing.ac.uk. DOI: 10.20944/preprints202411.0566.v1.

[28] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[29] *Resourcecontracts.org: Petroleum and mining contracts repository*, https://www.resourcecontracts.org/, Accessed: 2025-10-12, Natural Resource Governance Institute, the World Bank, and the Columbia Center on Sustainable Investment, 2026.

[30] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: Current trends and best practice guidelines", *Computer Speech Language*, vol. 67, p. 101 151, 2021, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2020.101151. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082030084X.

[31] E. Dahlin, "Email interviews: A guide to research design and implementation", *International Journal of Qualitative Methods*, vol. 20, 2021, Original work published 2021. DOI: 10.1177/16094069211025453. [Online]. Available: https://doi.org/10.1177/16094069211025453.

[32] S. K. Ahmed, R. A. Mohammed, A. J. Nashwan, *et al.*, "Using thematic analysis in qualitative research", *Journal of Medicine, Surgery, and Public Health*, vol. 6, 100198, Aug. 2025. DOI: 10.1016/j.glmedi.2025.100198. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949916X25000222.

[33] V. Braun and V. Clarke, "Thematic analysis.", in Jan. 2012, pp. 57–71, ISBN: 978-1-4338-1003-9.

A

Interview Method

The interviews were conducted in two steps. First, each participant was invited to a 25 min meeting, where they were informed about the purpose of the project, to understand what insights were being searched for, and where the interview questions were briefly presented. This gave the participants the opportunity to ask questions related to the project, as well as to ask for clarifications of the interview questions. These meetings were recorded and transcribed using automatic transcription for all but one case, which was done manually. After the introductory meeting, the participants were asked to answer the questions by email. This approach was chosen so that the participants could have time to reflect on the questions, discuss with their colleagues, and formulate clear and concise answers. By providing participants with time to think through their process, they could generate more valuable responses.

Email interviews can work well for exploratory studies, because they allow planned follow-up questions and learning across multiple interviews, thus helping researchers to think critically during data collection and notice any unexpected fields of further exploration [31]. Hence, they are useful for developing new ideas and can sometimes be better than in-person methods depending on the type of study and participants. This method was also beneficial because of the time constraint of the project, where conducting conventional, semi-structured interviews would create longer meetings with in-depth discussions of each question and more time needed for transcription and analysis of the content. Using email interviews for data collection frees up time to simultaneously develop and evaluate the system, while also respecting the work schedule of the participants.

The challenges or short-comings of the email interview method could be that respondents don't put the same amount of effort into written answers because they want to get it done quickly and there is less opportunity for follow up questions or clarifications compared to conventional interviews [31]. One participant expressed in the interview that, due to time constraints, they did not want to answer the questions in writing, but instead wanted the material from the interview to be used. In this case, the interview was recorded and transcribed for analysis.

Examples of questions the participants were asked were specific questions, such as; *Can you give specific examples of KPIs that are measured throughout the project, or KPIs that should be measured for clear follow ups?*; as well as more open ended questions such as *In a perfect world, how should collaboration research information be managed and do you see a role for AI or other new technologies in that future?*

In some cases, new questions arose after receiving responses via email. In those circumstances a follow-up email was sent, asking the participant either to elaborate on a topic, or to add further questions that might offer value for the project.

A.1 Sampling and Participant Recruitment

The prerequisite for being chosen as a participant was that they had to be actively involved in a collaboration with a party outside of AZ. These could be principal investigators of collaborations with academia, or scientists involved in a collaboration with non-academic businesses. In this way, it was guaranteed that the participants had experience working in and reporting on collaborations, which was a necessity to understand the problem. By choosing participants involved in different projects, it was ensured that the experiences of participants covered different kinds of data reporting processes used in the organisation. The participants were then found by leveraging the internal network of colleagues and looking through active collaborations taking place at the moment, as well as identifying responsible principal investigators.

In total, seven people were interviewed. Thematic saturation was achieved, meaning that, given the amount of interviews, no new essential data was gathered in the last interview, compared to what had already been established from previous interviews.

A.2 Thematic Analysis

To systematically and scientifically analyse the content from the interviews, a thematic analysis approach was used. Using thematic data analysis it is possible to identify, organise, and interpret patterns or themes that present themselves as a result of the interviews [32]. Braun and Clarke introduced the six steps of thematic analysis in 2006 [33]. The first step is to familiarise oneself with the gathered data, where the researcher reads and re-reads the data, noting ideas, patterns and early insights. The next step is to generate initial codes to identify meaningful features. The third step is searching for themes by grouping codes into themes. The themes are then reviewed in the fourth step. In the fifth and sixth step, the themes are defined and presented in the report. By following these clear and structured phases, researchers can capture nuance and complexity while maintaining transparency in the methodology. This way of working makes thematic analysis a robust, widely applicable approach [32].

When interviews were analysed, these steps were taken by reading and understanding the written answers and grouping them into themes, by noticing similarities in responses and drawing conclusions based on patterns of answers. The aim of these conclusions was to answer the second research question stated in Section 1.2.

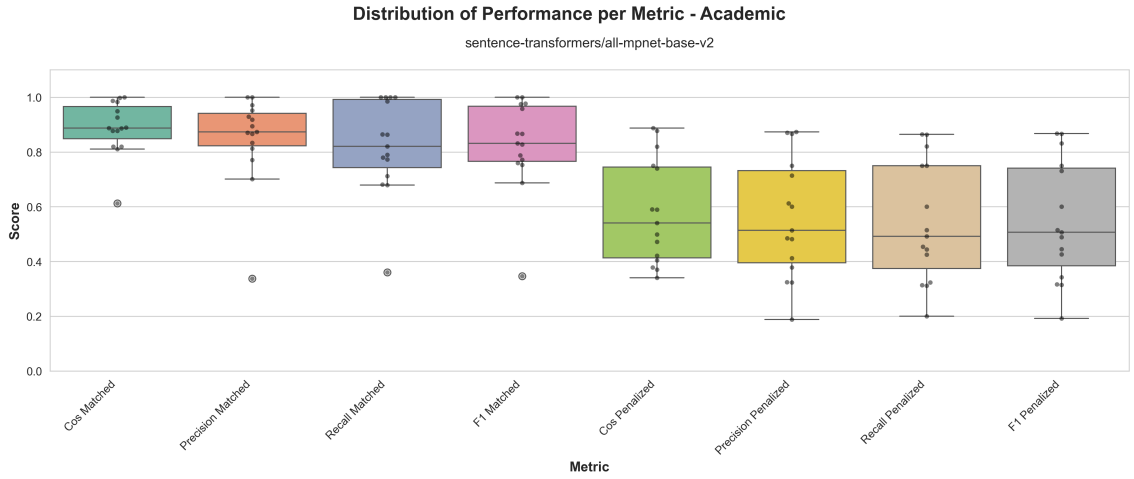
B

Backbone Performance Distributions

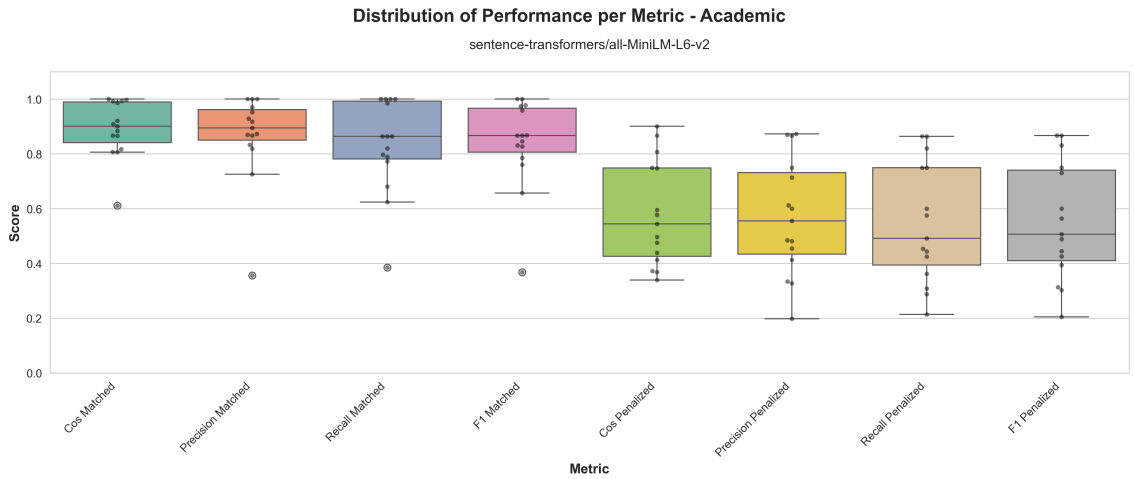
Here, the box plots of the performance distribution for all document types are presented. The distributions showcase the performance in terms of the defined metrics. The variations can be seen for each backbone and document.

A box in the box plot diagram corresponds to the interquartile range (IQR), which is defined as $Q3 - Q1$. $Q3$ is the 75^{th} percentile, while $Q1$ is the 25^{th}. The points that can be found below some of the box plots are outliers, determined by the fact that they are below $Q1 - 1.5\text{IQR}$.

Ultimately, the *all-MiniLM-L6-v2* was selected as the embedding model for the evaluation of objective extraction, as the mean score was slightly better for almost all metrics, with marginally lower variance as well. The scores also reflects a realistic span which is consistent with the F1-score, unlike the models that were disregarded due to positive bias.

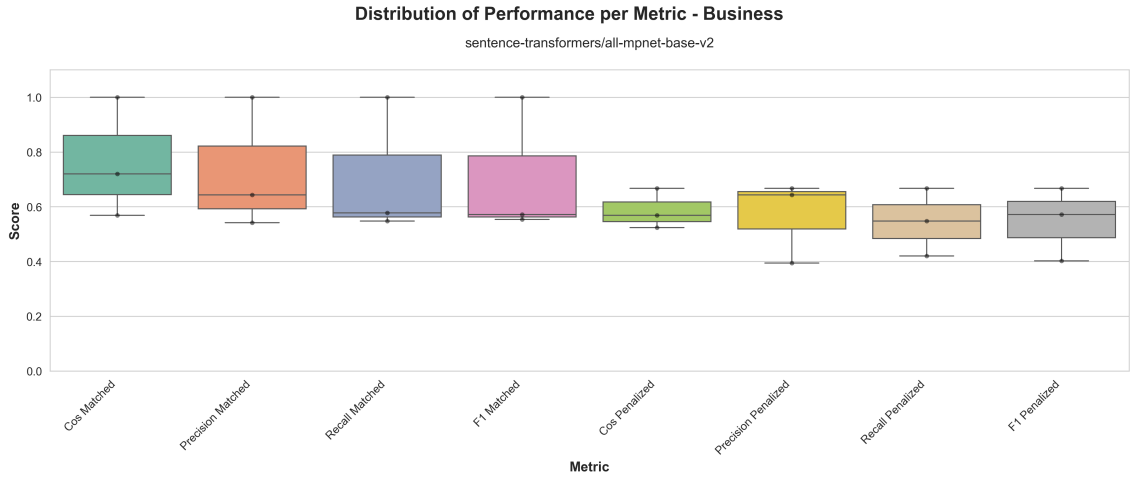


(a) Backbone: *all-mpnet-base-v2*

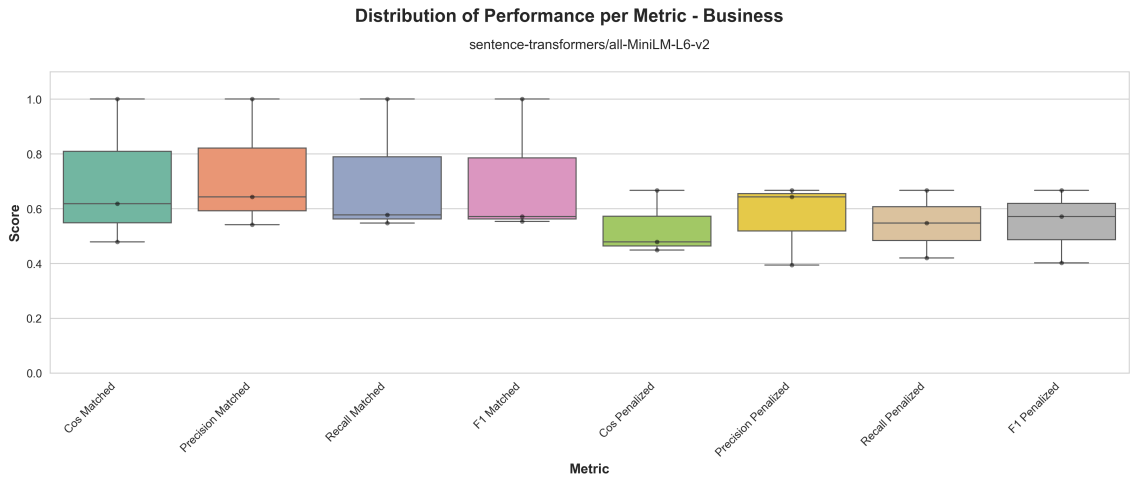


(b) Backbone: *all-MiniLM-L6-v2*

Figure B.1: Distribution of academic agreement scores across performance metrics for different embedding models.

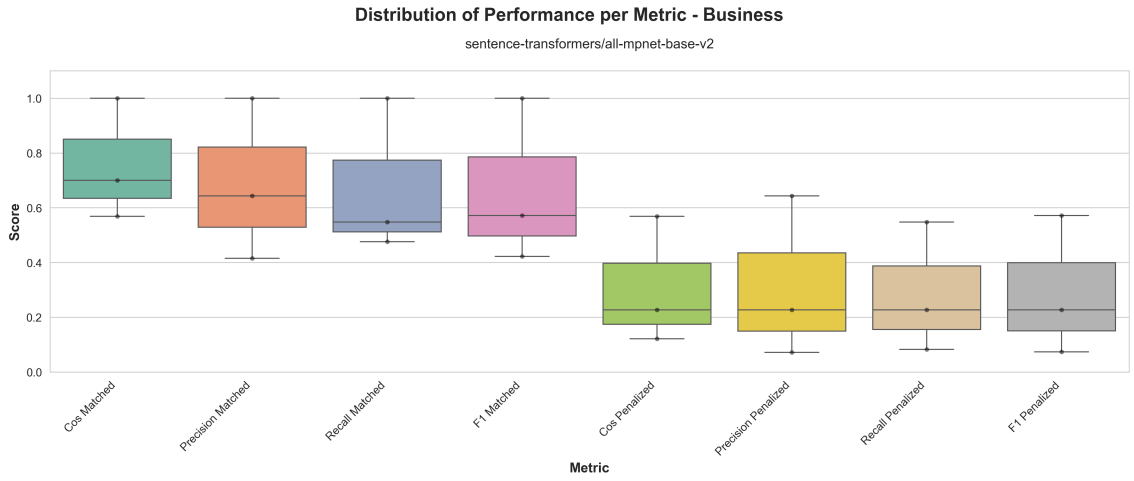


(a) Backbone: *all-mpnet-base-v2*

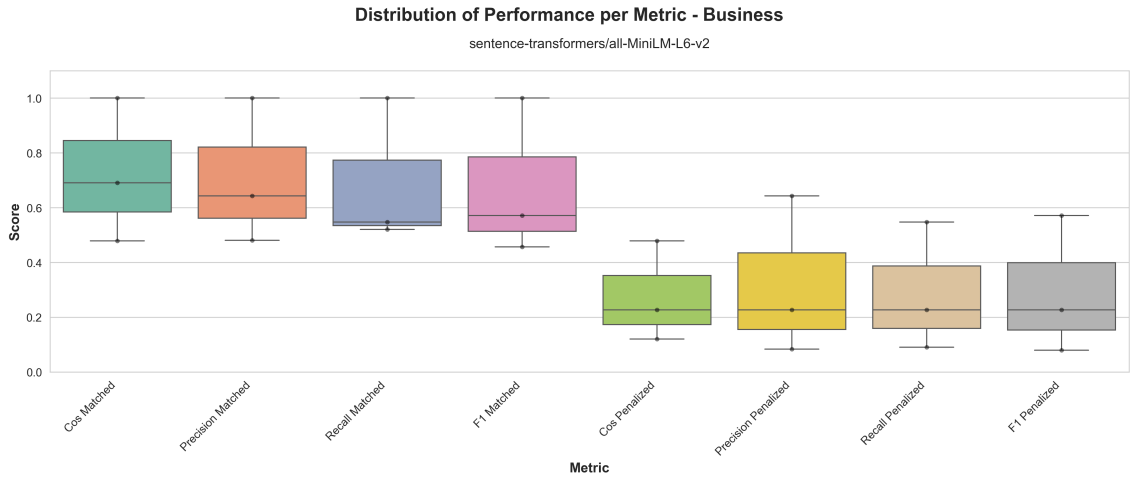


(b) Backbone: *all-MiniLM-L6-v2*

Figure B.2: Distribution of business agreement scores across performance metrics for different embedding models.

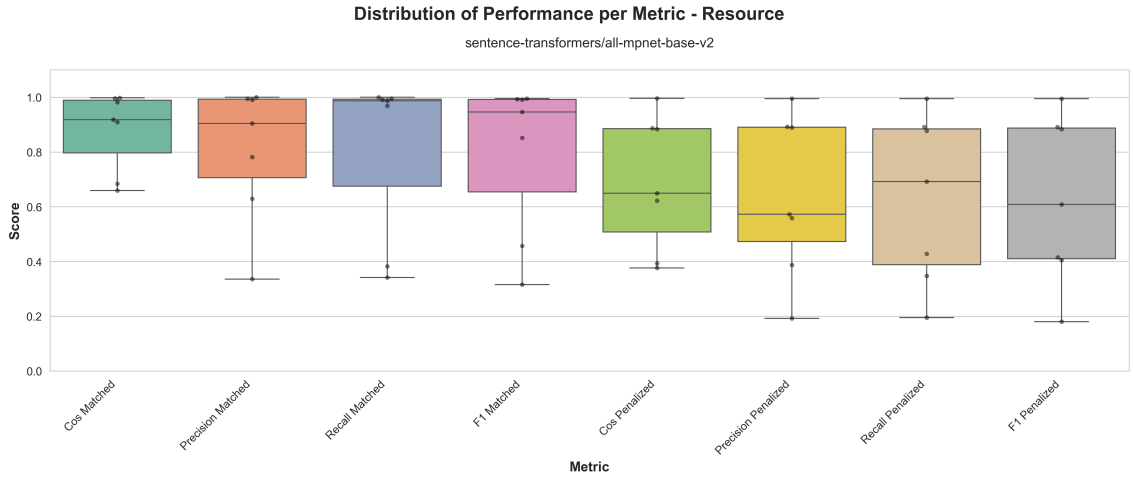


(a) Backbone: *all-mpnet-base-v2*

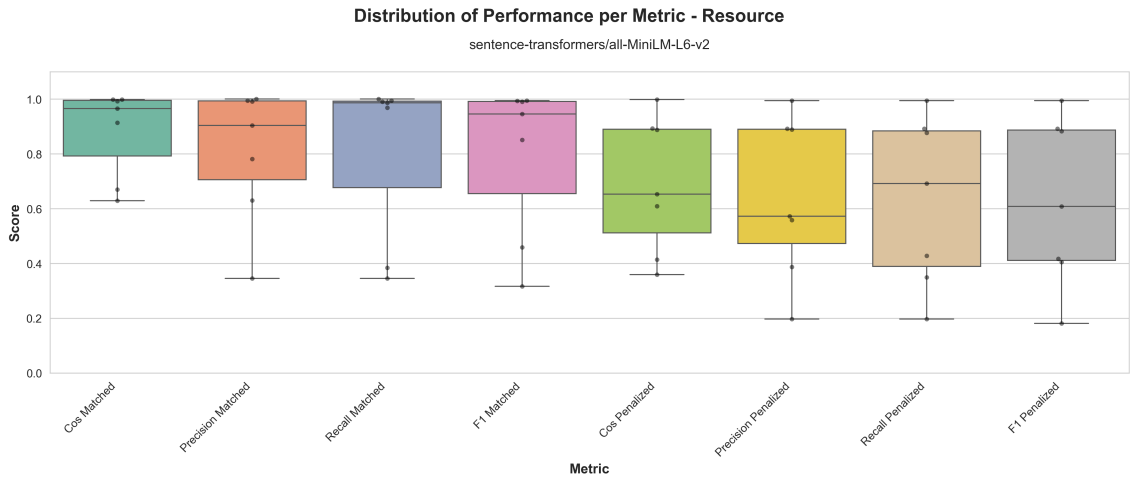


(b) Backbone: *all-MiniLM-L6-v2*

Figure B.3: Distribution of business agreement scores across performance metrics for different embedding models. These business agreements are, however, referring to the long versions (with complete objective extraction).



(a) Backbone: *all-mpnet-base-v2*



(b) Backbone: *all-MiniLM-L6-v2*

Figure B.4: Distribution of resource agreement scores across performance metrics for different embedding models.

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY