Artificial Intelligence for reporting

Proof-of-Concept to explore use of AI to improve reporting speed and consistency

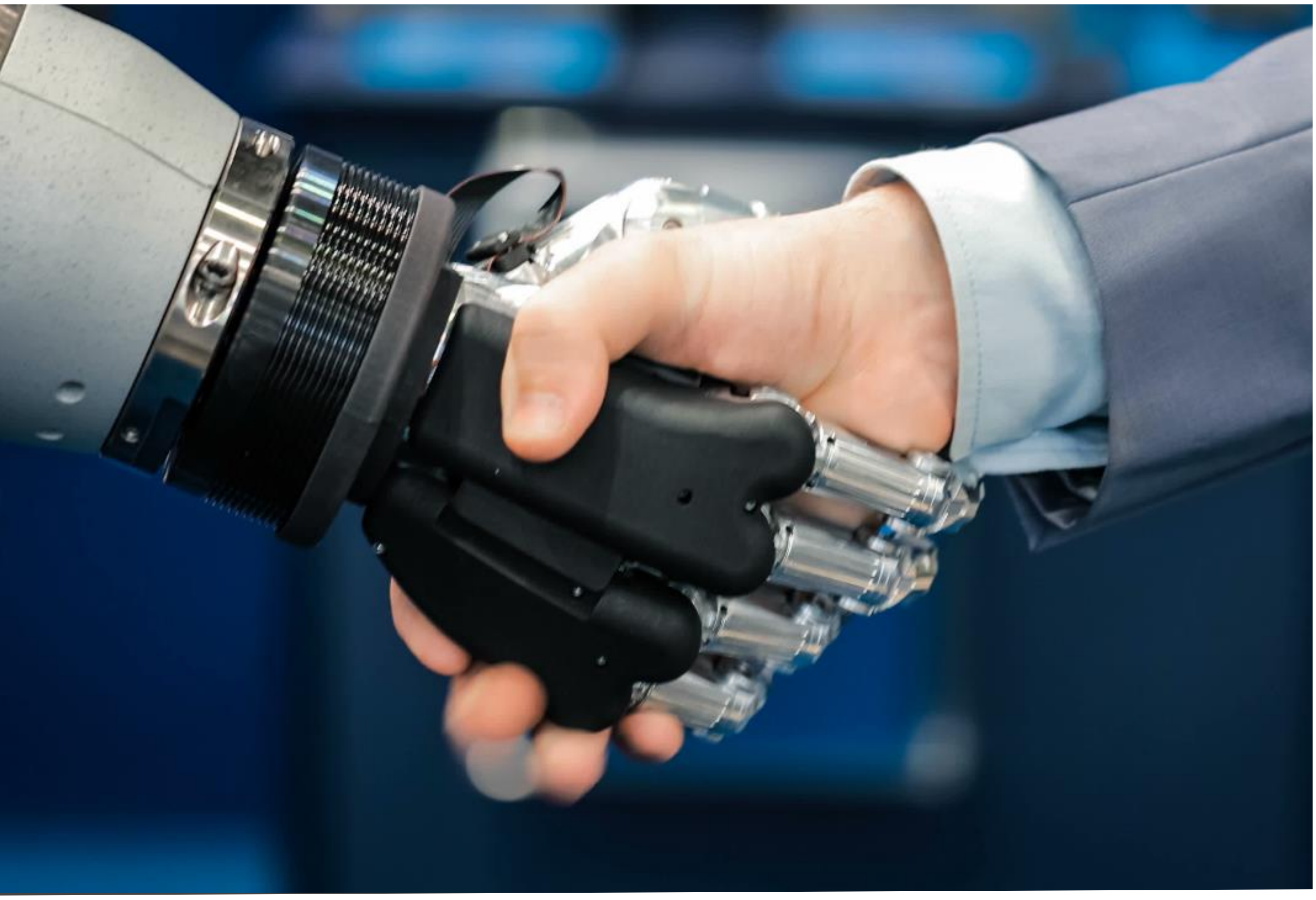


Table of Contents

1. Introduction

Increasing demands around customer expectations, technological capabilities, regulatory requirements, demographics and changing economics are reshaping the banking industry. Banks worldwide need to re-tool to win in this era of new challenges. They need to radically innovate and transform themselves for the future.

One of the biggest challenges that banks face is the high costs of reporting. It is traditionally labour-intensive, as it not only involves correct interpretation of reporting requirements, but also accurate sourcing and retrieving of corresponding data to populate the report. This high cost burden, to fulfill today's complex regulatory, risk and financial reporting requirements with widened scope, is expected to persist, or even increase in future.

Project Artificial Intelligence Reporting (AIR) envisions the digitisation of reporting for the financial services sector. It is aimed at reducing costs and mitigating key reporting challenges, so resources can be re-allocated to value-adding activities.

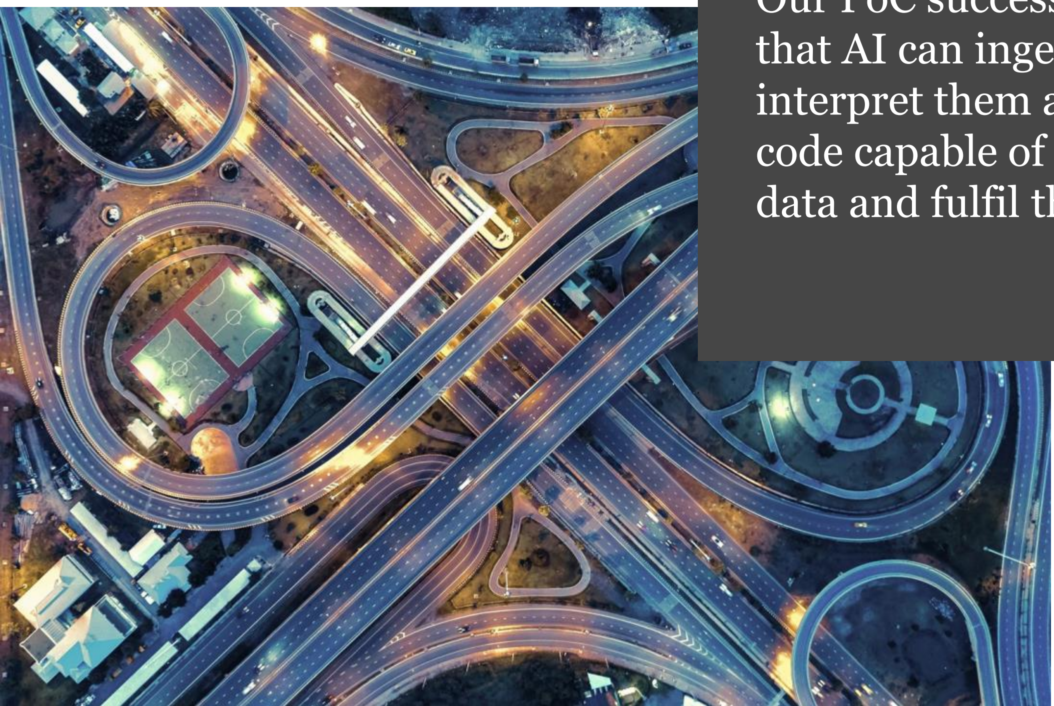
Leveraging innovations in artificial intelligence (AI) to automate reporting may potentially remove some major roadblocks in reporting, such as ensuring correct and consistent interpretation of reporting requirements, accurate sourcing and usage of data in fulfilling the reporting processes within a bank.

From an industry point of view, reporting for the financial services sector is further complicated by challenges such as the lack of consistency in regulations across borders, inadequate common industry interpretation, and the absence comparability across reports both within a bank and across multiple banks.

We initiated a Proof-of-Concept (PoC) for one of the aspects of a reporting process - interpretation of reporting requirements. We tested the concept of using AI to translate report requests into a machine-readable format that is scalable. We also identified the necessary conditions for subsequent pilot of Project AIR for participating banks.

Our PoC successfully demonstrated that AI can ingest English requests, interpret them accurately, produce a code capable of extracting the right data and fulfil the requests. The encouraging outcome reaffirmed the feasibility of digitalising various reportings for financial industry players.

We believe, with greater collaboration between regulators and banks, an industry-wide solution using AI can be developed, potentially saving a significant amount of time and costs. This whitepaper elaborates on our PoC journey and the key learnings.



Our PoC successfully demonstrated that AI can ingest English requests, interpret them accurately, produce a code capable of extracting the right data and fulfil the requests.

2. Background

Reporting has been a pertinent problem for banks globally. Banks are required to produce complex reports that typically consume a significant amount of time and resources. In recent times, some regulators are requesting for even more complex and comprehensive reporting, and have also increased granularity of data submissions.

This trend is likely to persist and become even more complicated over time[1], as seen in the revised requirements proposed by:

- The Markets in Financial Instruments Directive II (MiFID II) in the United Kingdom and European Union, which expanded the scope of existing transaction and instruments reporting[2], and
- The Monetary Authority of Singapore via Notice 610 (MAS 610), which now mandates banks to report granular multi-dimensional details of balance sheet and off-balance sheet information[3].

The bulk of the reporting costs is incurred in interpreting requirements, rewriting the rules into business texts, and then translating into codes to retrieve the necessary data. All of these activities are labour-intensive to begin with, and are further complicated by banks' matrixed organisation structure and intricate system and data architecture.

Consider this: The MiFID II rules were described in 1.7 million paragraphs covering over 30,000 pages. To date, the implementation has cost the banking industry over €2.5 billion (about S$4 billion). Further, to comply with MAS 610, financial institutions (FIs) now need to report 340,000 data points across 67 reports within a period of 12 months - an increase of 8,000% from 4,000 data points required earlier.

Reporting remains an industry-wide challenge in Singapore, as interpreting reporting requirements accurately and sourcing the correct data required continue to be costly and labour-intensive. While technology vendors do provide solutions for selected regulatory reporting purposes, there is no common solution that can meet the complete range of reporting requirements.

[1] Butler, Tom, Paul North, and John Palmer. "A New Paradigm for Regulatory Change and Compliance." A Whitepaper by the RegTech Council (2018).

[2] PricewaterhouseCoopers. "MiFID II Transaction Reporting: Detecting and Investigating Potential Market Abuse." (2017).

[3] Monetary Authority of Singapore. "Proposed Revisions to MAS Notice to Banks 610 and MAS Notice to Merchant Banks 1003". (2017).

2.1 Key reporting challenges

There are four key challenges in reporting across the financial services industry and within banks:

Lack of consistency across borders (i.e., there are no common reporting standards across borders). Banks have to cope with disparities in reporting requirements in each country of operation.

Lack of common industry interpretation of reporting requirements (including regulatory reporting). Each bank relies on multiple clarification sessions with regulators (e.g. during the exposure draft stage for new reporting requirements), consultants and internal subject matter experts and resources to interpret the reporting requirements. There are significant duplicated efforts by banks across the industry to independently interpret new industry-wide reporting requirements. This is neither cost nor time-efficient.

Lack of comparability and consistency. The lack of a common industry interpretation is exacerbated when data is required for consolidation and comparison across banks, as in the case of statutory reporting and/or regulatory reporting. There is a lack of comparability by analysts when comparing financial disclosures, and a lack of consistency when such information is consolidated for analysis by the regulators.

Inconsistent interpretation and definition. Reporting requests may arise from various functions from within a bank. Given that banks do not commonly define a common framework for interpreting and defining these requests, the activity of interpreting requests/requirements is generally a manual activity that relies heavily on the judgement of the person handling the request. This gives rise to different results when similar requests are produced by different parties. Banks typically resolve this through manual reconciliation efforts.

3. Future of reporting: Our vision and expected benefits

We believe, through digitalisation, reporting will no longer be repetitive or subjective, thereby minimising overall efforts and costs. With greater collaboration between the industry and regulators, the use of AI in reporting will benefit the financial services sector players in the following ways:

- **Data can be autonomously mined and retrieved** from multiple data sources and platforms to auto-populate reports, reducing manual efforts required
- **Automation of the end-to-end process of reporting will reduce manual interventions** such as mappings, checks, reconciliations and reviews. In turn this will cut down the resource-intensive efforts, and channel them into deeper focus on data quality and report analysis
- **New regulations and changes to regulations need not be a burdensome process.** With regulator-issued, machine-readable and machine-executable codified regulations, there will be significant reduction in the need for regulatory interpretations and therefore, reduced risks of misinterpretation and increased overall accuracy of reporting
- **Improved quality of submitted reports** as a result of automated data quality checks, variance reasonableness analysis and validation checks
- **Consistent industry returns** as standard definitions and requirements are set within systems to organise and classify data, therefore improving the comparability of results for deeper analysis by investors and regulators

We believe the ideal strategic solution would be a scalable AI reporting solution that will codify the entire reporting process, from report interpretation o report generation, across the entire financial services industry in Singapore.

This end-to-end AI reporting solution will streamline the reporting process with a common interpretation of reporting requirements, resulting in significant reduction in manual efforts for business users. Regulators will benefit from an improved and comparable standard, and consistency of reporting in the financial services industry. Banks will also benefit from potential time savings and reduced resources needed for reporting.

4. Proof-of-Concept (PoC)

4.1 Objectives

The key objective is to prove that AI can effectively digitalise the current reporting process.

Project AIR aimed to demonstrate, AI could:

- ingest reporting requests in natural language (English) despite varying levels of clarity and completeness;
- learn how to explicitly interpret the request functionally (i.e., figuring out what is needed), by leveraging an external, industry ontology and custom definitions and synonyms;
- transform the request into an SQL code that can extract the right data to produce correct results from a standardised data model; and,
- produce a technology-agnostic and user-friendly representation of the request (pseudocode) that can be easily translated into other coding languages and understood by users.

4.2 Scope

The scope of the PoC was focused on one aspect of the reporting process - the interpretation of reporting requirements (Step 1, Figure 1). As a by-product of this scope, the PoC did digitise a small extent of the remaining reporting process (e.g. data extraction & classification stage) (Figure 1).

Figure 1: Typical high-level reporting process

To adequately test the AI's ability to interpret reporting documents, we chose to focus on the following 14 key reporting metrics and dimensions for this PoC, taking into consideration banks' loan exposures.

- Accrued Interest
- Amortisation Type
- Borrower Name
- Borrower Income
- Credit Rating
- Collateral
- Country of Incorporation
- Industry
- Interest Rate
- Loan Principal Amount
- Loan-To-Value (LTV) Ratio
- Maturity Date
- Outstanding Balance
- Payment Schedule

3 key deliverables were developed as part of the PoC:

- **Conversational interface (Chatbot):** A web-based chatbot interface for the requestor to key in a request and thereafter, review and confirm the output generated.
- **Natural Language Processing (NLP) model:** It was based on current research and datasets, which are customised to perform well in a banking data environment. This model accepts the request input in the form of english sentences, and then processes and maps the sentence against the standard vocabulary available in the ontology to perform data retrieval. This model does this instantaneously, with reproducible results meeting the needs of the request.
- **Code generator:** Once the request parameters are mapped against the respective meaning (as applicable) within the ontology definition, this code generator then generates a pseudocode. This is supported by technical code such as SQL to verify the accuracy of the pseudocode. In addition, the actual results are also generated (based on the PoC dataset).

4.3 Key activities

In the process of the PoC, we identified two key streams of activities:

Business stream

In sharing business domain knowledge, our risk and regulatory team and the industry advisors identified realistic requests and common interpretation of expected results to ensure the realism of business requirements. The following steps were taken:

- Based on the selected 14 reporting metrics and dimensions, we identified a list of 50 reporting requests (AIR50) that the solution should fulfil
- Co-designed with the industry advisors, these AIR50 reporting requests were kept realistic and where possible, took reference to actual reports that banks submit, such as Comprehensive Risk Assessment Framework and Techniques (CRAFT), MAS 637, MAS 649, and MAS 656
- Used four levels (Very Hard, Hard, Medium, Easy) to categorise the complexity of these requests. Additional requests and fragmentation for some requests were used to expand the list of reporting requests to 75 (AIR75)
- Adopted the industry standard for Financial Industry Business Ontology (FIBO)[4], key attributes of a loan and definitions as baselined for data definitions in scope for the PoC
- Expected interpretation and result outputs for the AIR50 reporting requests were agreed upon to baseline the training plan and confirm the final output the solution was meant to produce

Technical stream

Technical skills to ensure the functional interpretation of requests was done accurately, and to quantify the request complexity and learning progression of the model, following steps were taken:

A limited data model was developed, to have sufficient data for the 14 key reporting metrics and dimensions.

- The AIR75 reporting requests were converted into SQL.
- Each of the reporting requests were then analysed for their complexity, and proxied by the SQL-statement token count.
- A learning progression training plan was designed to teach the model to progressively handle increasing complexity of request conversion from text to SQL.
- The training process involves increasing the complexity of the requests progressively, and having the training plan emulate the complexity distribution of the PoC request.
- Standardised, technology-agnostic queries in pseudocode are produced, which include the conditions fulfilled and executed. This is supported by technical code such as SQL to verify the accuracy of the pseudocode.

[4] FIBO is an ongoing industry initiative that had not been completed yet at the time of this PoC.

4.4 Limitations

Scope

Our PoC scope is limited to loan exposures, selected reporting metrics, dimensions, and ontology. Hence, we recognise that the results of the PoC excludes the various nuances associated with other reporting metrics and dimensions that may be key for a loan product.

The PoC does not preclude, and cannot predict the potential challenges and difficulties which may arise from an expanded scope. Furthermore, the specific scope and timeline for this PoC is built on the acceptance of selected assumptions, with regard to the system output, data quality, data formats etc.

Data model

While the test database was realistic, it does not represent the entirety of a bank system architecture, where voluminous sets of data may not be stored in integrated infrastructure. The result analysis was also limited by the data model used. For instance, the loan principal and collateral value in the database are static. As these are components to derive the loan-to-value (LTV) ratio, more complex trend analysis cannot be obtained.

Furthermore, the use of an incomplete industry ontology which has not yet been widely adopted by the banking industry has also resulted in some limitations. Firstly, the mapping of the data elements to the ontology terms will have to be refreshed when FIBO is updated, as currently the data model is not directly integrated with FIBO. Secondly, as the focus of the PoC was only loans. The reporting metrics and dimensions selected in the PoC is therefore limited to some loan-related data elements, and does not encompass other metrics and dimensions that may be key for other products (e.g. derivatives).

Industry representation

The industry advisers from international and regional banks are not representative of the wider banking industry. Where this effort is expanded to the wider banking industry, it is necessary to consult a range of industry advisers in order to have sufficient representation of the banking industry.

5. Learnings from the PoC

5.1 Standardised industry definitions are essential

From the PoC, we noted that standardised industry definition for reporting terms are key. In the PoC, we leveraged the FIBO industry ontology as much as possible for common definitions. However, FIBO and other industry standards are still being developed and some terms are yet to be defined industry-wide, leaving room for interpretation.

For example, in looking at a borrowers' industry, Singapore Standard Industrial Classification does not have a standalone term to refer to the tourism industry. Instead, "tourism industry" is derived from a combination of industry sub-sectors, such as "administrative and support services" for tours and "water transportation" for cruises. This might be different for every bank, which results in multiple interpretations and differing reporting outputs. Therefore, a fully developed industry definition and aggregation criteria would be key to resolve any doubts in terminology and ensure consistency.

In addition, we observed that certain requests could be perceived as open-ended, hence ambiguous. For instance, the term 'average' in the context of deriving an average LTV of a loan can have different meanings (i.e., this average LTV could be based on simple or weighted LTV, depending on the corresponding reporting needs). In such cases, additional clarification with business users would then be required to determine the actual intent of the request.

To work around such challenges and ensure consistency, we worked closely with our industry advisors to reach a consensus[5] on the interpretation for the requests in scope for this PoC upfront. With this agreed interpretation, we then trained the NLP model based on the correct expected output to prove the functionality of the model.

5.2 Implicit knowledge has to be included

As the model ultimately needs to be trained first in order to mimic a reporting specialists' thought process, it was crucial to incorporate implicit knowledge in the model. Implicit knowledge here refers to the business understanding that a reporting specialist would apply when deriving the request results. This is necessary in order for the solution to replace the need for a reporting specialist in the results generation process, and to ensure its scalability and effectiveness.

For instance, some requests implicitly require the data "as of today" to produce the right request results. However, if the request does not explicitly mention "as of today", the NLP model will not be able to answer the requests accurately. In these cases, the implicit knowledge that a reporting specialist would have (i.e., when there is no "date" mentioned in the request, it is inferred that we take data "as of today") has to be imparted to the NLP model. Imparting implicit knowledge as a part of NLP model training process leads to higher accuracy for all the requests that contain an implicit knowledge component.

The fact that it is possible to impart implicit knowledge in this loan-specific NLP model reassures that an NLP model can also handle requests that contain implicit knowledge. This is a positive indication if the model needs to be further scaled to handle more complex requests and/or other banking products/ domains in the future.

5 Consensus was based on whether the assumption/ approach applied is logical and acceptable at the current juncture of proving the concept, considering the limited scope and timeline set.

5.3 An effective training plan is required to fulfill typical business requests

As of June 2020, state-of-the-art models were based on datasets (e.g. Spider[6]) with only limited SQL-statement complexity. In our PoC, we had business requests that were initially about 2 - 3 times more complex than the current state-of-the-art models (Figure 2) when measured by the number of tokens.[7] Generally, more complex requests would translate to difficulty in achieving high levels of accuracy. Apart from request complexity, our PoC also deals with higher levels of functional complexity (i.e., there is a greater number of table joins and conditions needed and a wider variety of required SQL functions such as the date comparison functions used). All of these contribute to the high token count.

A comparison of the request complexity

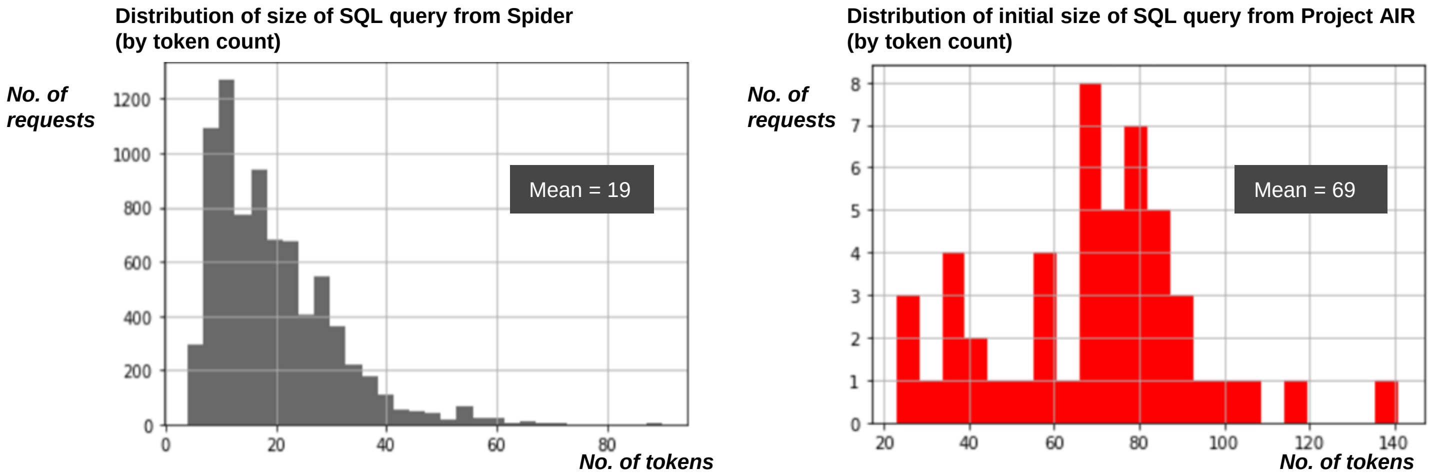


Figure 2: A comparison of the request complexity (proxied by token count) between Spider and Project AIR PoC. Initially the average business request of Project AIR PoC had 69 tokens, whilst Spider (the training dataset for current state-of-the-art Machine Learning models) has an average of 19 tokens. Source: Spider dataset[8] (left) and Project AIR (right).

In order to ensure that the NLP model can still produce accurate results for complex business requests, an effective training plan was developed to supplement the available training dataset. This is done via a progressive approach (i.e., from easy to very hard requests) to identify which step of the training is most effective, so that it can be reiterated to teach the model efficiently.

Two other helpful training methods were the rephrasing and fragmentation of reporting requests. Rephrasing is used in cases where human requests have not been worded clearly for the NLP model to ingest and provide an output. Fragmentation is used to break down a complex request into several simple requests, to further improve the NLP model's ability to generate the correct answers eventually.

Through our PoC, we concluded that with an effective training plan, complex natural language requests can be successfully ingested, interpreted into SQL code and accurately answered.

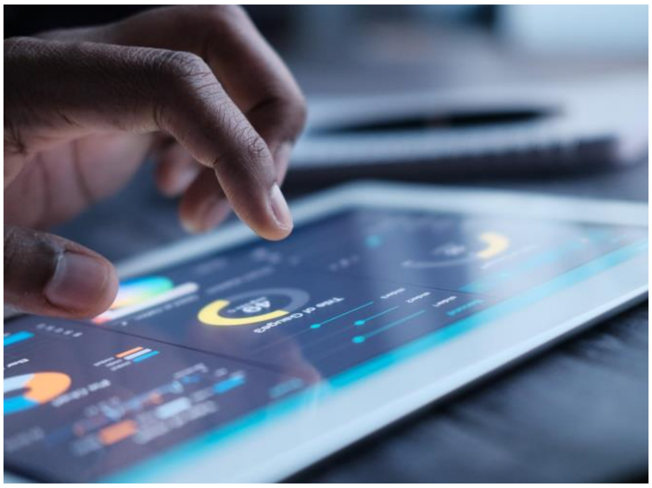
[6] Yu, Tao, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma et al. "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task." arXiv preprint arXiv:1809.08887 (2018).

[7] While there are 4 levels of request complexity based on token count (Very Hard, Hard, Medium, Easy). The AIR50 reporting requests generally fell into the Very Hard and Hard categories.

[8] Ibid.

5.4 Vital to refine the NLP model to suit the language of business domain

The language scope used in the Spider dataset is written in common-use English. In specific domains such as economics, medicine and in this case the banking industry, words could typically be used with different meanings. For instance, the word “default” would generally be defined as a preselected option. In our PoC, specific to loans, “default” would typically be referring to the inability to fulfil an obligation. Hence, refining the NLP model with appropriate training examples to suit the domain language would be necessary to ensure that the PoC requests can be answered accurately.



5.5 SQL-query functions needed for reporting requests to be unlocked

The current models are only able to interpret natural language requests into SQL code using a subset of SQL functionality. For instance, these models do not have a date-time SQL functionality, which is key for some of our PoC requests. The NLP model was thus enhanced to unlock and expand access to more SQL functionalities.

One functional workaround used during the PoC was to reword the request and SQL query in a way that will enable the use of SQL functionalities beyond what is available within typical NLP models. In the above example of managing date-time SQL functionality, we rewrote the question to reflect a specific date range (e.g. restating “next 3 months” to “period between 2020-01-01’ to 2020-03-31”). This helped us to manage questions that would otherwise be unsolvable within the current scope with robust precision.

5.6 Scaling the NLP model through innovative capabilities

Changing data model structure

Typical data model structures/databases in banks are designed to store information using minimal data. While the retrieval of data is often sufficient to address business requests, there are also requests that may require calculation. These requests usually appear as long and complex SQL statements as the calculation logic needs to be embedded in the SQL statement. However, we noted that it is possible to incorporate this calculation logic in the data model, which leads to shorter SQL statements that need to be predicted by the NLP model. In our case, this is done either by enriching the database (described below) or via python as a post-processing step to combine several simpler requests to get to a hard request outcome.

An example of the incorporation of this logic is to have daily records in the database as opposed to a record that only exists when a change happens. This means creating an unconventional data model structure/database whereby data is replicated where required, and interim Virtual Data Tables (VDTs) are generated to contain this data in the format that is easy for the NLP model. By using a standardised SQL-query that will transform data into easier query-able structure (record per day, versus records per change), it is easier to retrieve data and allows for simpler SQL-queries.

This capability allows for a NLP model that is scalable to a wider scope while still being compatible with the data model in banks. By experimenting with ideas such as interim VDTs, the scalability of the NLP model is tested and we can get an initial sense of the model’s flexibility and compatibility for industry wide implementation in the future. There might be some tweaks needed when we expand the scope to include the wide-ranging and voluminous set of bank data that may not necessarily be stored in integrated infrastructure.

Pseudocode

In addition, we developed a pseudocode (i.e., plain language description similar to SQL but for business users’ consumption) that can be translated into query languages apart from SQL. This enables organisations to implement other query technologies to translate the pseudocode into their preferred query language.

This innovation also allows business users to clearly understand how the request has been interpreted, and how the results can be expected to follow. It ensures the interoperability of our solution between systems, provides clarity to business users and is designed to scale across banks with different operating environments.

6. PoC's overall benefits demonstrated: Reflections

Given the limited scope and duration of this PoC, not all of the expected benefits that we outlined earlier were explicitly proven. However, this PoC was able to demonstrate some aspects of two aforementioned benefits in Section 2.2:

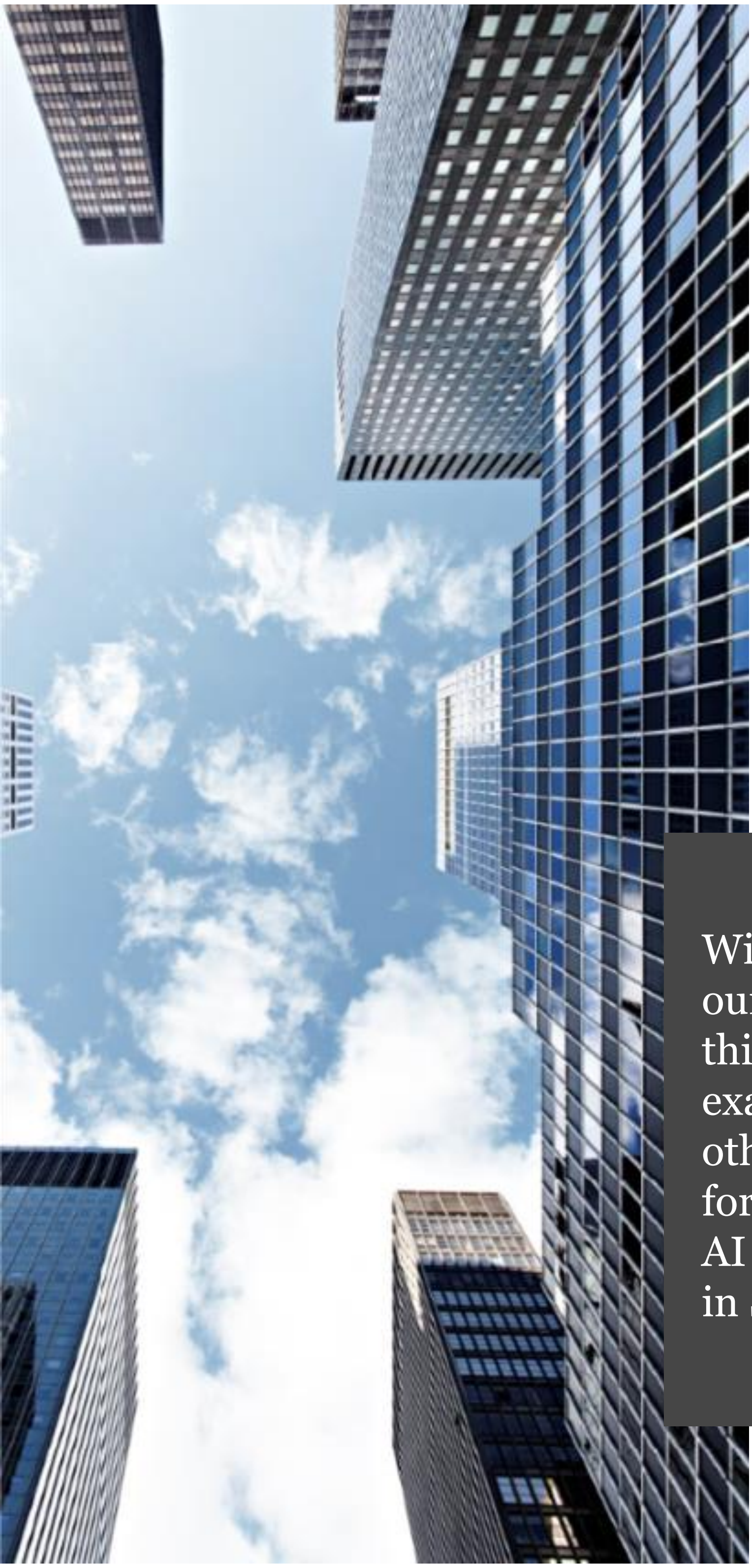
Eventual benefits that may be realised	How our PoC demonstrated some aspects of this benefit
Data can be autonomously mined and retrieved from multiple data sources and platforms to auto-populate reports. This reduces the manual efforts in report production.	Our NLP model is able to automatically interpret what the reporting request is asking for and thereafter trawl through the database to retrieve the right data needed to answer the requests, with high accuracy. Additionally, as reports are essentially several reporting requests, our NLP model can auto-populate reports, as long as the report structure is given in advance to teach the NLP model where to populate the result of each reporting request.
Consistent industry returns as standard definitions and requirements are set within systems to organise and classify data. The idea is to improve the results comparability for deeper analysis by investors and regulators	We were able to train the NLP model in a consistent manner that would replicate what happens when standard definitions and requirements are set and applied in systems. This was achieved through our use of (i)FIBO ontology, (ii) an agreed set of business terms and definitions, (iii) one common interpretation of each reporting request, and (iv) the expected result for these requests.

Looking ahead, we aim to further develop this solution and make it production-ready - one that can scale-up for the industry. We believe that more updates can be made such as:

- expanding the scope of this initiative to include other reporting processes;
- increasing the number of training data/domains (i.e., beyond loan data);
- reviewing the effectiveness of chatbots;
- rephrasing reporting requests in a machine-readable format; and
- focusing on standardisation.

These refinements can help bring to fruition the full benefits that we envisioned for the future of digital reporting.These additional future benefits include: automation of the end-to-end reporting process, ease of implementing new regulations and/or changes to regulations where needed, and overall improved quality of submitted reports.

7. Conclusion



Our PoC displayed strong potential. Business users can utilise the NLP model to streamline the reporting process. This will enable the banking industry to reap benefits such as time-savings or reduced manual efforts for reporting.

Overall, the PoC demonstrated AI's ability to ingest English requests, with only minor rephrasing or fragmentation where human requests are 'unclear' (in a technical sense - for the NLP model), interpret requests into SQL code and produce correct results from the data model.

The PoC has also demonstrated the ability to produce a technology-agnostic representation of the request (pseudocode), which ensures the scalability for usage across the banking industry.

Furthermore, relevant solutions and workarounds were applied successfully to address and resolve the technical challenges that arose in the course of developing a NLP model fit for reporting purposes.

With promising results from our PoC, our next step is to expand and enhance this solution to a wider scope, with more examples and additional inputs from other industry stakeholders. We look forward to realising our vision of using AI to meet reporting needs for all banks in Singapore, and beyond.

8. Acknowledgments

Pricewaterhousecoopers Consulting (Singapore) Pte. Ltd.was awarded the Monetary Authority of Singapore's Financial Sector Technology and Innovation (FSTI) Proof of Concept (PoC) grant on 26-May-2020. The FSTI PoC grant provides funding support for experimentation, development and dissemination of nascent innovative technologies in the financial services sector.

Industry advisers

We are grateful to the following organisations for being a part of our journey. The insights shared by the individuals have proved to be a valuable sounding board. It helped us in shaping the scope of the research/ proof of concept, which is based on our understanding of existing AI use cases and lessons learned from banks who have tried adopting AI technologies.

#	Organisation	Name
1	Bank of China	Feng Zhigang
2	Bank of China	Ray Cui Wei
3	Hongkong Shanghai Banking Corporation	Sundar Iyer
4	Hongkong Shanghai Banking Corporation	Kevin Lim
5	Hongkong Shanghai Banking Corporation	Wong Seow Hong
6	Oversea-China Banking Corporation	Noel D'Cruz
7	Oversea-China Banking Corporation	Andrew Tan
8	Oversea-China Banking Corporation	Tan May Ling
9	Sumitomo Mitsui Banking Corporation	Lum Chong Keen
10	Sumitomo Mitsui Banking Corporation	Chin Chih Ching
11	Sumitomo Mitsui Banking Corporation	Lim Wei Han

PwC Project Team

We also want to express our gratitude to the project team members for being part of this journey.

#	Name	Project Role
1	Irene Liu	Project Leader
2	David Yakowitz	Solution Architect
3	Andy Goldin	Lead AI Designer
4	Catherine Lee	Project Director
5	Shierly Mondianti	Project Management Office
6	Melissa Lin	Project Management Office
7	Coen Hennipman	Senior Consultant
8	Disha Grover	Senior Consultant
9	Cindy Sutanto	Consultant
10	Kenneth Guo	AI Specialists
11	Jonathan Chen	AI Specialists

Contact

Irene Liu
Partner
Risk & Regulatory
PwC South East Asia Consulting
irene.l.liu@pwc.com

Catherine Lee
Director
Risk and Regulatory
PwC South East Asia Consulting
catherine.hs.lee@pwc.com