GPT總結

會議主題：公司資料分析與AI應用

主要業務與技術架構

1. **資料分析與報告**
   - **量化分析**：數字來自資料庫，LLM 用於補充敘述性統計。
   - **資料庫串接**：可串接多種資料庫，進階應用使用 Hadoop 加速處理。
   - **受眾面向**：包括品管、生管等，非限定於製造業，也涵蓋總經理層級的財務分析需求。
   - **Insight為核心**：從問答到完整報告生成，回答都附帶引用(citation)。

2. **AI Agent 開發**
   - 驗收方式：
     - **效能**：曾達到 99% 滿意度，提供 400 題生成測試，結果以 Excel 供客戶評估。
     - **指標**：BLEU 分數或人工評分。
   - **Evaluation 模組**：客戶透過按讚與倒讚功能匯出資料，以便後續自行進行微調(finetune)。
   - **RAG（檢索增強生成）**：專注於提取技術，使用於資料搜索與報告生成。
3. **Fine-tuning**：僅限內部使用，涵蓋多模態模型、繁中 OCR 模型與 Layout 模型。

3. **部署方式**
   - On-premises 為主，根據資料庫類型差異收費。
   - 優勢特色：提供 interpretable code generation 服務。
   - 硬體資源：
     - 使用 A6000 Ada、A6000、L40s 等硬體設備，根據預算規劃。
   - 回應時間：
     - **基本問題**：7-8 秒。
     - **複雜統計**：40 秒。
   - 文件處理：將文件轉換為 metadata，輔助資料庫內搜索。

商業模式與銷售流程

1. **商業模式**
   - 訂閱制：開設帳號即收費。
   - 硬體資源費用依預算配置。
   - 多數 B2B 情況由工程人員對接，提供 end user 可直接使用的功能。

2. **銷售與驗收流程**
   - 初次會議直接詢問需求，兩週後提供免費 live demo。
   - 若效果不佳，Sales 會協助說明，例如需要更多資料支持。

業務規模與目標

1. **業務現況**
   - 一年 IPO 級專案 10 個，預計 2025 年後擴充至 8 人團隊。
   - 去年 Q2 開始 B2B 模式，共完成 9 個專案，平均耗時 3 個月，每案收入約 200-300 萬。
   - POC(Proof of Concept)：完成 10-20 個。Q3 後主要依賴 SI 銷售。

2. **2025 年目標**
   - 翻十倍，營收達三四千。
   - 希望公司聚焦產品開發與落地應用，銷售部分由夥伴負責。

原筆記

公司資料分析？報告主要是量化分析，數字從資料庫來，由LLM補充敘述性統計
可以串接多種資料庫，更進階甚至有用到Hadoop做加速。
面向有品管、生管
大客戶有潤泰的水泥
不限定製造業情境，有些是給總經理用的財務分析之類的

給予Insight為目標，從問答到Report都有做
回答都會有citation

主要在做AI agent

說驗收的結果有達到過99%，給四百題然生成完給Excel讓客戶自己評估滿意度。
驗收方式：BLEU或是人工評分

evaluation模組是按讚跟倒讚 讓客戶會匯出資料供他們後續自己可以finetune

RAG的部分提取做比較多的工

finetune僅供內部使用，有multi-modal model、繁中OCR model跟Layout model

on-premises

串資料庫根據類別會多收錢

區隔的部分是on-premise的code generation/interpretable

回應時間 基本問題 7-8s 複雜的統計 40s

把文件資料轉換做metadata再串進資料庫幫助搜索

受眾 B2B對接都是工程人員，幾乎都會加到讓end user可以直接用

客戶的IT有一些參數可以調，但就基本的Topk或temparature
再深入一點的目前還沒遇到

第一個開會就直接問要不要，兩週後一個live demo（免費），如果效果不好sales會幫忙說話說需要更多資料之類的。

一年的專案IPO level 10個，預計年後可以擴到8人
去年Q2才開始做B2B，做了9個，平均三個月結束，一個兩三百
POC有十個20個，Q3過後比較少自己跑sales都是SI幫忙，之前都是自己跑

開account就會收錢，訂閱制，硬體看預算
硬體主要是A6000 ada、A6000、L40s

2025目標 翻十倍 三四千、focus在pruduct