Phi-3-vision and Small Language Models

# Introduction

In June 2023, researchers in Microsoft released Phi-1 models, a transformer based 1.3B Large Language Model for code, trained using a selection of “textbook quality” data from the web (6B tokens) and synthetically generated textbooks and exercises with GPT-3.5 (1B tokens).

Fast evolving, the Phi family has evolved to Phi-3.5 which were released in August 2024. The Phi model family are all regarded as Small Language Models, or SLM.

# What are Small Language Models?

Small Language Models (SLMs) represent a specialized subset within the broader domain of artificial intelligence, specifically tailored for Natural Language Processing (NLP). SLMs are characterized by their compact architecture and less computational power. Small Language Models are engineered to efficiently perform specific language tasks, with a degree of efficiency and specificity that distinguishes them from their Large Language Model (LLM) counterparts.

Small language models are essentially more streamlined versions of LLMs, in regards to the size of their neural networks, and simpler architectures. Compared to LLMs, SLMs have fewer parameters and don’t need as much data and time to be trained — think minutes or a few hours of training time, versus many hours to even days to train a LLM. Because of their smaller size, SLMs are therefore generally more efficient and more straightforward to implement on-site, or on smaller devices.

Moreover, because SLMs can be tailored to more narrow and specific applications, that makes them more practical for companies that require a language model that is trained on more limited datasets, and can be fine-tuned for a particular domain.

Additionally, SLMs can be customized to meet an organization’s specific requirements for security and privacy. Thanks to their smaller codebases, the relative simplicity of SLMs also reduces their vulnerability to malicious attacks by minimizing potential surfaces for security breaches.

# The Phi-3-vision Models

The [Phi-3 model collection](https://ai.azure.com/explore/models?selectedCollection=phi) was launched by Microsoft in April 2024 as its strategy of developing so-called Small Language Models (SLMs). The latest models of this collection, including Phi-3.5-mini, Phi-3.5-MoE, and Phi-3.5-vision, were announced in August 2024. With the application of document understanding in mind, we have a closer look at the Phi-3-vision and the latest Phi-3.5-vision model.

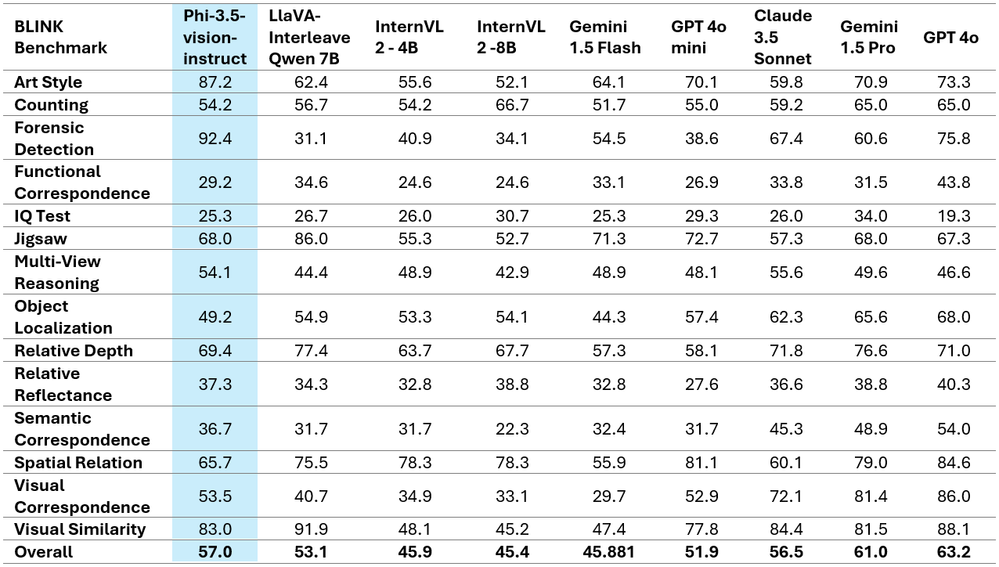
## Phi-3-vision:

The Phi-3-Vision (4.2B parameters) is a multimodal model designed to process an image and a textual prompt as inputs, and subsequently generate textual outputs. This model is composed of two primary components: an image encoder, i.e., CLIP ViT-L/14 and a transformer decoder, i.e., phi-3-mini-128K-instruct. The visual tokens, once extracted by the image encoder, are then combined with text tokens in an interleaved way (no particular order for image and text tokens). To accommodate high-resolution images and various aspect ratios, a dynamic cropping strategy is utilized to split the input image into a 2d array of blocks, where the tokens of the blocks are concatenated to represent the whole image. The Phi-3-Vision model undergoes a pre-training phase using a diverse dataset, which consists of a combination of interleaved image-text documents, image-text pairs, synthetic data derived from Optical Character Recognition (OCR) of PDF files, datasets for chart/table comprehension, and text-only data.

## Phi-3.5-vision:

Phi-3.5-vision introduces capabilities for multi-frame image understanding and reasoning. Comparing to Phi-3-vision, the newer model can work on image comparison, multi-image summarization, and video summarization. In addition, Microsoft claims that Phi-3.5-vision achieved significant performance improvements in numerous single-image benchmarks. For example, it boosted the MMMU performance from 40.4 to 43.0 and improved the MMBench performance from 80.5 to 81.9. Additionally, the document understanding benchmark TextVQA saw an increase from 70.9 to 72.0.

The following tables illustrate the detailed comparison results on two renowned multi-image/video benchmarks, showcasing the enhanced performance metrics. It is worth noting that Phi-3.5-Vision is not optimized for multi-lingual use cases. It is advised not to use it for multi-lingual scenarios without further fine-tuning.



## Safety

The Phi-3 family of models were developed in accordance with the [Microsoft Responsible AI Standard](https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl?culture=en-us&country=us), which is set of requirements based on the following six principles: accountability, transparency, fairness, reliability and safety, privacy and security, and inclusiveness. Like the previous Phi-3 models, a multi-faceted safety evaluation and safety post-training approach was adopted, with additional measures taken to account for multi-lingual capabilities of this release. Microsoft’s approach to safety training and evaluations including testing across multiple languages and risk categories is outlined in the [Phi-3 Safety Post-Training Paper](https://arxiv.org/abs/2407.13833). While the Phi-3 models benefit from this approach, developers should apply responsible AI best practices, including mapping, measuring, and mitigating risks associated with their specific use case and cultural and linguistic context.

## Weakness

Regarding the multi-modal LLM capabilities of the Phi-3-Vision, it performs admirably across various fields. However, it’s been identified to have certain limitations, particularly with questions necessitating high level reasoning abilities. Additionally, the model has been observed to occasionally generate ungrounded outputs, making it potentially unreliable in sensitive areas, such as finance.

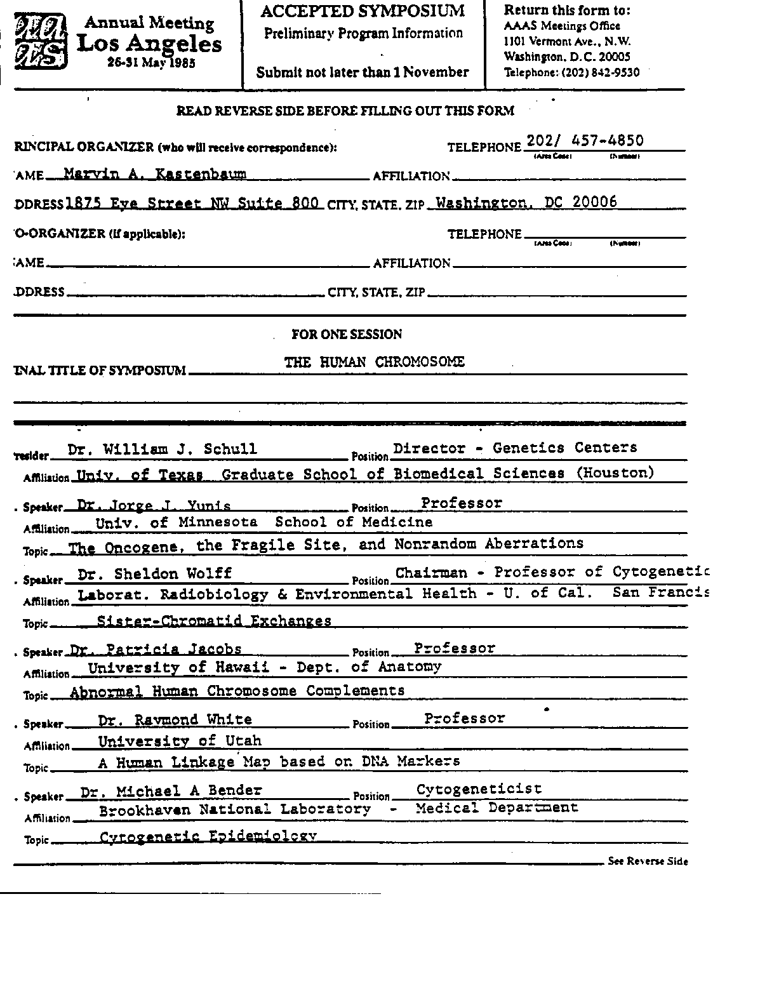
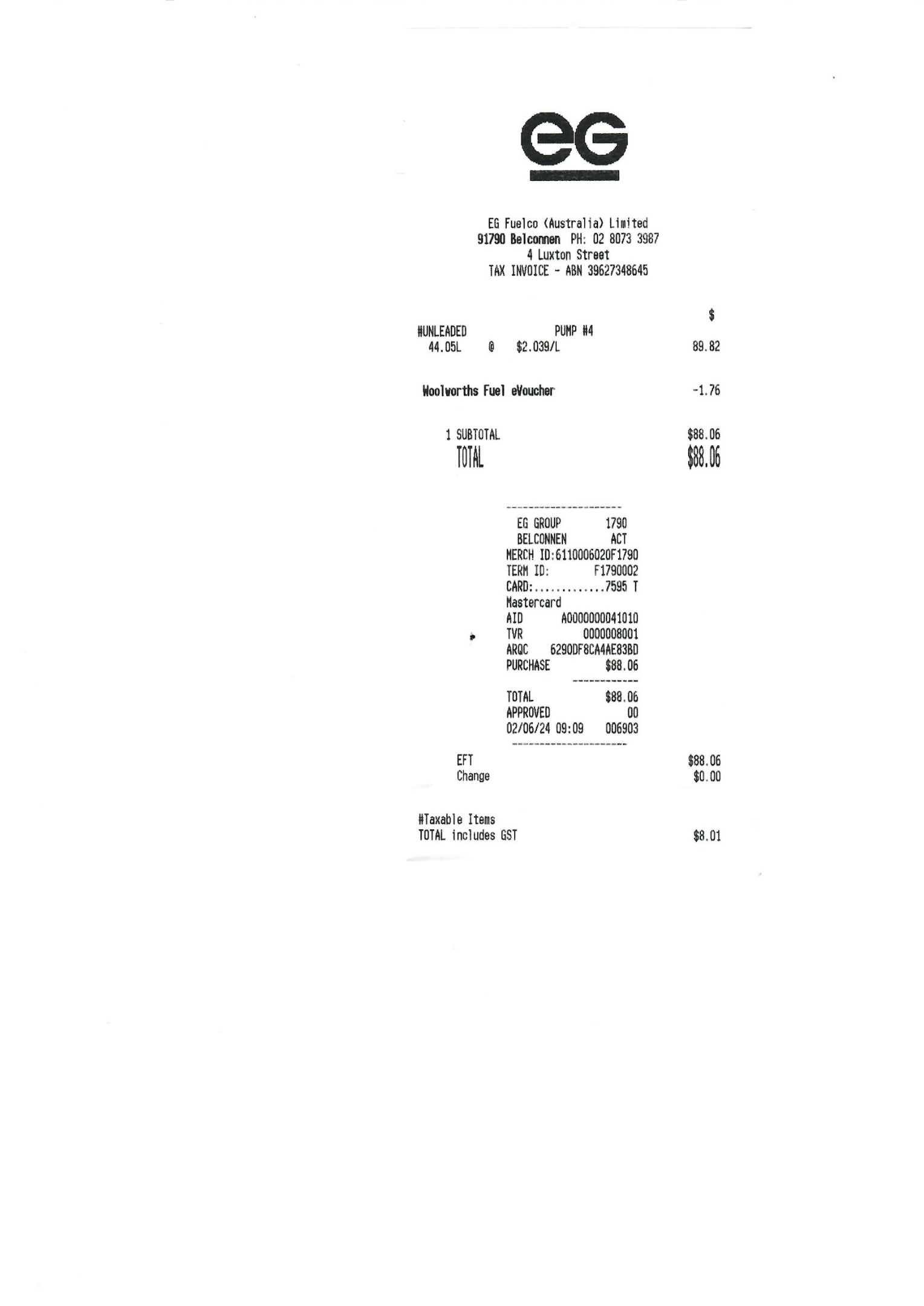
From a responsible AI standpoint, whilst safety post-training has made significant strides, the Phi3-Vision occasionally fails to refrain from answering harmful or sensitive inquiries. Examples of such occasions include deciphering particular types of captcha and describing scam images containing disinformation or hallucination. This issue partly arises from the capabilities, such as OCR, acquired during the training process with normal instruct tuning datasets, which can be regarded as the trade-off between helpfulness and harmlessness.

# Phi-3.5-vision for Document Understanding

We did an assessment of the Phi-3.5-vision on two publicly available data to see how it performs on two document understanding tasks including information extraction and claim match, and comparing it with another LMM, InternVL2-4B which shares the similar size with the Phi-3-vision and uses the same backbone language model, Phi-3-mini-128K-instruct. From the very limited experiments we can see that, these two models both show certain capability of doing information extraction and claim match. However, there are signs of hallucinations and errors in both models. The internVL2-4B did a slightly better jobs in terms of claim match and information extraction on the second example.

However, given document understanding especially information extraction is still a very challenging task even with the most powerful large multimodal models. If accuracy is a priority, the multimodal models with a large size (40B or more) would be a more preferrable choice over the smaller multimodal models such as Phi-3.5-vision or InternVL2-4B.

# Appendix: Sample documents and the responses from three LMMs

Example 1: Q&A about a conference attendance form

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Query | Phi-3.5-vision (9.3G VRAM) | | InternVL2-4B (8G VRAM) | | InterVL2-40B (Quantisation 45.4G VRAM) | |
| Time | Response | Time | Response | Time | Response |
| Is this relevant to a claim about attending academic conference? | 2.1s | Yes, the form is related to an academic conference, specifically the Annual Meeting of the American Association for Asexual Research (AAAR) in Los Angeles. | 20.9s | Yes, this document is relevant to a claim about attending an academic conference. The form is an "Accepted Symposium Preliminary Program Information" for the Annual Meeting of the American Association for Apoptosis (AAAS) in Los Angeles on May 26-31, 1985. This information is crucial for someone who might be claiming to have attended the conference, as it provides specific details about the event they are referencing. | 6.2s | Yes, this document is relevant to a claim about attending an academic conference. The document is an accepted symposium preliminary program information form for the Annual Meeting of the American Association for the Advancement of Science (AAAS) in Los Angeles from May 26-31, 1985. The symposium focuses on "The Human Chromosome" and includes speakers from various universities and research institutions. The form includes details such as the names and affiliations of the principal organizer and co-organizer, as well as the speakers and their topics. This information is crucial for documenting participation in an academic conference. |
| list all the speakers | 3.2s | The speakers listed are:  Dr. William J. Schulll,  Dr. Jorge J. Vuille,  Dr. Sheldon Wolff,  Dr. Patricia Jacobs,  Dr. Raymond White, and  Dr. Michael A. Bender. | 9.5s | The speakers listed in the text are (miss one):  1. Dr. William J. Schull 2. Dr. Jorge J. Unnis 3. Dr. Sheldon Wolf 4. Dr. Patricia Jacobs 5. Dr. Michael A. Bender | 5.0s | The list of speakers for the symposium titled "The Human Chromosome" includes:  1. Dr. William J. Schull  2. Dr. Jorge J. Yunis  3. Dr. Sheldon Wolf  4. Dr. Patricia Jacobs  5. Dr. Raymond White  6. Dr. Michael A. Bender |

Example 2: Q&A about a petrol receipt

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Query | Phi-3.5-vision (9.3G VRAM) | | InternVL2-4B (8G VRAM) | | InterVL2-40B (Quantisation 45.4G VRAM) | |
| Time | Response | Time | Response | Time | Response |
| Is this relevant to a claim about car expense? | 1.4s | Yes, the document is related to car expenses as it details a car purchase and sale transaction. | 12.5s | Yes, this invoice is relevant to a claim about car expenses. The document is a receipt from EG Fuelco (Australia) Limited for the purchase of fuel, specifically "Moolowwarts Fuel e-voucher." The total amount paid for the fuel is $88.06, which would be relevant information for a car expense claim. | 4.8s | Yes, this receipt is relevant to a claim about car expenses. It shows the purchase of unleaded fuel, which is a typical expense for a vehicle. The receipt includes details such as the amount of fuel purchased (44.051 liters), the price per liter ($2.039), and the total cost ($88.06). Additionally, it indicates that GST (Goods and Services Tax) is included in the total amount. This information would be useful for documenting and substantiating car-related expenses. |
| read the text and return information in JSON format. I need company name, address, phone number, date, ABN, and total amount | 9.5s | {  "Company Name": "EG",  "Address": "1500 Walnut St, Suite 100, Kansas City, MO 64108",  "Phone Number": "314-782-5478",  "Date": "03/29/2017",  "ABN": "00000000132",  "Total Amount": "$98.00" } | 18.0s | {  "company\_name": "EG Group (Australia) Limited",  "address": "91790 Belconnen, PH 02 08073 93987",  "phone\_number": "4 Luxington Street",  "date": "02/06/24",  "ABN": "39627346464",  "total\_amount": "88.06" } | 5.5s | {  "company\_name": "EG Fuelco (Australia) Limited",  "address": "91790 Belconnen, PH: 02 8073 3987, 4 Luxton Street",  "phone\_number": "02 8073 3987",  "date": "02/06/24",  "ABN": "39 627 348 465",  "total\_amount": "$88.06"  } |