Requirements for Training Data: Scam Call Detection

- **Use Available in Standard Malay, English, and Manglish only. (dialects are not allowed).**
  - For commercialization, dialects can be covered as more training data is gathered.
- **Each conversation(in a call) must have at least 20 turns and a total length of over 5 minutes.**
  - Definition of 'Turn' : **the period in a conversation** involving multiple participants **during which a particular speaker begins and finishes speaking**.
    A. When the other person starts speaking after the first speaker finishes, it is counted as a new turn.
    B. Even if there is a brief silence and the original speaker resumes speaking, this is also counted as a new turn.
  - Length estimation : 5 minutes is about **1,500~1,750 syllables,** since Malaysian speaks about 300~350 syllables a minute.
- **It doesn't matter whether the person who initiates the conversation is the victim or the scammer.**
- **Keywords commonly used in voice phishing in Malaysia should appear across different conversations for effective learning of AI model. However, it is not necessary to include all such keywords intentionally; rather, they should be distributed in proportion to their actual frequency in real incidents.**
- **Local proper nouns** (such as names of administrative, bank, financial company, police office in Malaysia) **should be included without alteration.**
- **Data must be real-world scenarios representative of today's crime reality. Snippets must be available for ≥80% of the data; only then can it be augmented and used.**
- **We do not require voice data.**
- **Data file formats, size, situations**
  - The total amount of data we are requesting is 50,000, consisting of 30,000 General, 10,000 Specific Domains, and 10,000 Scam cases.
    Please refer to the details below for further information.

Types	Description	Format	Each File	Total
General	1. (10,000 files) Similar to local scam scenarios in Malaysia, but in the form of everyday conversations (such as police calls, investment consultations, delivery inquiries, etc.), not actual crime situations 2. (20,000 files) instances of everyday conversations that may occur in call situations	json (*.txt)	Min. 20 turns & Min. 5 minutes. (1,500~1,750 syllables)	30,000 files
Specific Domains	Call center data related to finance, telecommunications, and shopping payments, Etc			10,000 files
Scam	Criminal cases arising in the different scenarios outlined on the next page			10,000 files

- **Preferred Types of Scam Data**

  - We have researched the current state of scam-related crimes in Malaysia as outlined below.
  - We would like the scam data to be structured in proportions simillar to each category.
  - It would be helpful if the scenarios included clear distinguishing points between criminal situations and real-life situations, so that the model can be trained effectively.
  - Additionally, the categories defined here should be indicated with the 'category' meta tag in the JSON file.

Scam Status	Proportions	Categorize	Meta tags	Example
Macau Scam	-	**impersonation of authorities or public institutions**	"Macau Scam"	A victim living in Kuala Lumpur received a phone call from someone impersonating a 'Chinese embassy official'. The scammer deceived the victim by claiming that their passport had been used for illegal activities and demanded a deposit to resolve the matter."
E-commerce Fraud	33.2%	Delivery Scam	"Delivery Scam"	A scammer impersonating a delivery company told the victim that a package had been ordered under the spouse's name and demanded cash on delivery. The victim paid for an item they never ordered, but the package was never delivered.
		E-commerce Fraud	"E-commerce Fraud"	Fraudsters promoted non-existent products at cheap prices via social media or advertisements, luring victims into making payments.
Phone Message Scam	30.0%	Voice Scam	"Voice Scam"	Using AI to clone or mimic the voice of a boss or acquaintance, scammers deceived victims by claiming an urgent matter had arisen.
		Phishing SMS Scam	"SMS Scam"	Impersonating banks, government agencies, or courier companies, scammers sent text messages containing payment links. When victims clicked the links, they were redirected to fake websites where their financial information was stolen.
		Social Media Giveaway Scam	"Giveaway Scam"	Victims paid participation fees for social media giveaway events, then received WhatsApp calls claiming they had won a prize. To receive the winnings, they were asked to make additional transfers.
Investment Fraud	15.6%	Investment Scam	"Investment Scam"	Through unknown platforms, scammers approached victims with promises of high returns. After clicking links, victims made multiple transfers to various accounts and were urged to send even more for "extra profits," but never received any money back.
Loan Fraud	12.3%	Loan Fraud	"Loan Fraud"	Fraudsters posted "easy loan" advertisements and contacted victims, claiming that processing fees had to be paid before loan approval. They collected these fees but never provided the promised loans.

- **Tagging Rules of General Data**
  - For General data, the meta tags should broadly cover categories such as banking consultation, insurance consultation, telecom consultation, delivery, e-commerce, and general conversation. Alternatively, recommended category tags would also be helpful.

Sample Data: ① Scam Call Detection #1

- **Sample** : Input and target data combined into single JSON format

```
[
  {
    "region": "Malaysia",
    "category": "Investment scam",  type of call Category
    "is_vp": 1,   indicator of scam: 1=Yes(It is scam)
    "conversation_id": 2,
    "full_content": [
      {
        "sent_id": 1,
        "RX/TX": "TX",   speakers: TX=caller
        "stt_text": "Hai, perkhidmatan manual TTC Education. Perlukan soalan peperiksaan? Isi
pautan ke simpanan."
                                                └ PROPER NOUN of a specific
      },                                          Entity in Malaysia
      {
        "sent_id": 2,
        "RX/TX": "RX",
        "stt_text": "Oh, betul -betul? Itu akan membantu! Di mana pautannya?"
      }
    ]
  }
]
```