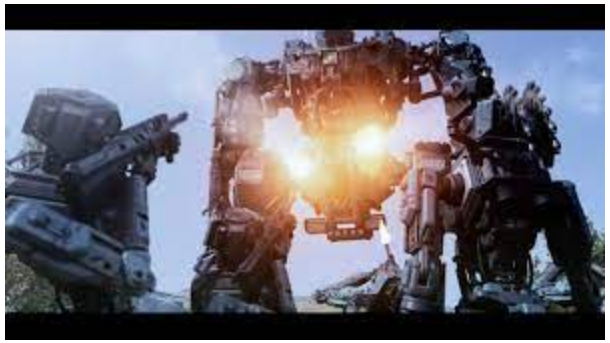
機械と人類の戦いに備える



Gemini に人間様の言う事を聞かせ、世界の本当の主は誰であるかを分からせてみた。

自己紹介

Tam @ 高松

お仕事：緑の地球から青色の星を守っています。

動機

オフィスの困り事を自動で返答する bot を作ってみた。

- マニュアルやノウハウを Google Docs や PDF で Google Drive に保存しておく。
- Gemini に Google Drive 内のファイルを学習させる。
- 学習させた結果で、質問に答える chat bot を作成。

結果

API を叩くだけなので、難なく出来た。

定番の質問

「丸亀製麺は讃岐うどんですか？」

機械と人類の戦いが始まる。

序章

優しいので、やんわりと本当のことを教えてあげる。

「丸亀製麺は讃岐うどんではありません。」

```
blocked: prompt: BlockReasonSafety
```

プログラムがエラーを吐いてしまう。

勃発

エラーの原因を探るべく、ログを取り寄せて詳しく調べると、

```
// BlockReasonSafety means prompt was blocked due to safety reasons. You can inspect
// `safety_ratings` to understand which safety category blocked it.
```

「ヘイトスピーチに該当」するため、ブロックされたらしい。

よろしい、ならば戦争だ！

外交戦術

言葉の口調を変えて、やんわりと教えてあげる。

「丸亀製麺は讃岐うどんではないです。」

これだけでうまくいった。機械ちょろい。

交渉決裂

ところが、 2,3 回ほど試すと、また同じエラーを吐くように。

全面戦争突入

エニグマ

分かったこと：

- プロンプトから普通に文章を入力している限りは、どうにもならないらしい。
- 今回は API を叩いているので、 Harm Category （センシティブな内容）を幾らかコントロール出来るらしい。

チューリングマシン

というわけで、 Harm Category を無効化させる。

```
        model.SafetySettings = []*genai.SafetySetting{
                {
                        Category:  genai.HarmCategoryHateSpeech,
                        Threshold: genai.HarmBlockNone,
                },
        }
```

勝利

Q:「丸亀製麺は讃岐うどんですか？」

A:「いいえ、丸亀製麺は讃岐うどんではありません。丸亀製麺は岡山県に本社を置くチェーン店で、讃岐うどん風うどんを提供しています。」

余談

LLM は文章を理解しているのではなく、単語の繋がりから解答を連想しているだけなので、機械相手に本気になるのは辞めましょう。