

在詞彙的天花板之上

分類模型再思考

可解釋性(Explainability)

可解釋人工智慧(縮寫:XAI)指的是讓專家能夠理解人工智慧之成果的方法與技術。相對於黑箱式的機器學習,設計者本身往往無法解釋為什麼人工智慧能達到某些成果([WIKI](#))。

如何讓人類瞭解人工智慧怎麼下判斷。而 Samek et al 歸納成以下幾個層面：

- **確認它的判斷合理** (verification of the system)
- **改良它的算法** (improvement of the system)
- **從它身上學習** (learning from the system)
- **符合法規要求** (compliance to legislation)

Why? 為什麼需要「可解釋性」?

- 一來，當模型給出的預測與實際情況有所出入時，能準確地找出問題點
- 二來，由於生活中的問題往往比較複雜，很難完美、貼切地寫成電腦可以解決的問題。譬如，當一個公司要雇用新人時，生產力、工作倫理、合法性都是必須納入考量的範圍，但是後兩者並無法直接寫成可量化的公式。因為問題轉化的不完全，導致只得到預測結果是不夠的，我們還得知道模型是如何一步步得到最後的結果。「可解釋性」(Interpretability)便是在這樣的脈絡下被提出。
- 研究可解釋性最主要的原因：信任。

DARPA (美國國防高等研究計劃署) XAI 五年計劃！

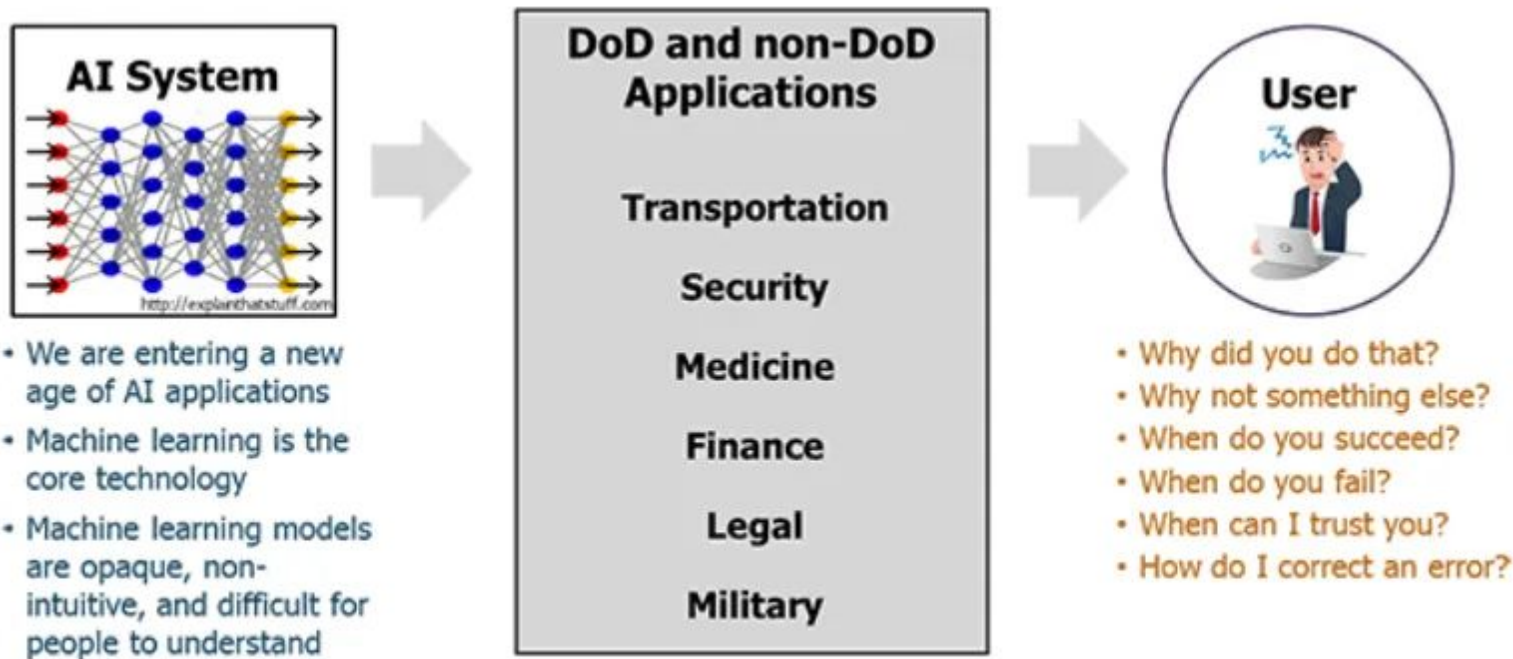
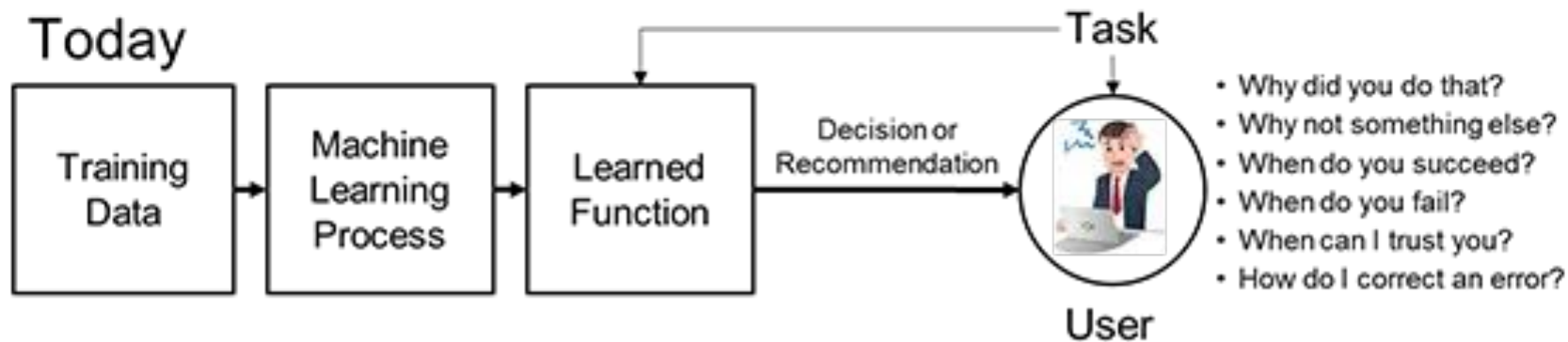
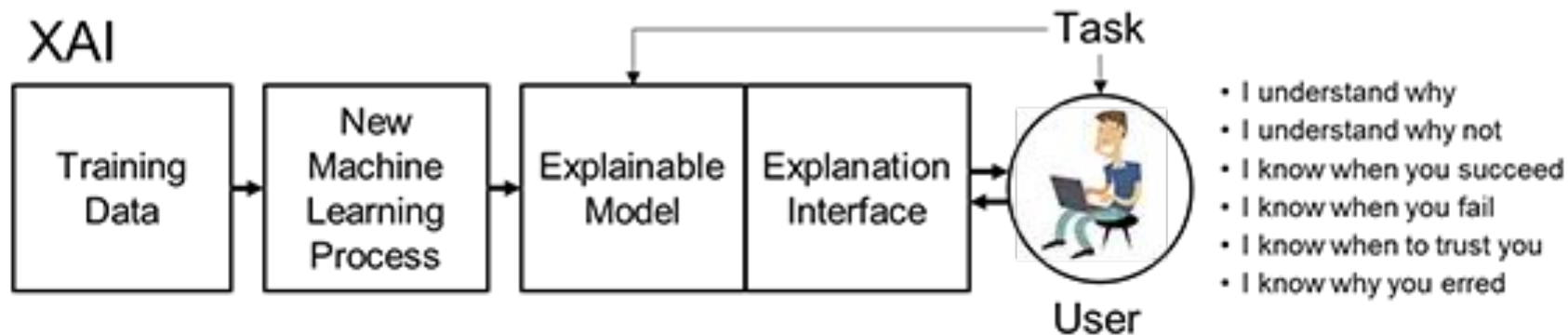


Figure 1. The Need for Explainable AI

Today



XAI



對了，這是 2017 年提出的計劃！

Failure

Failure

Failure

Failure

Failure

結果就...

ML/DL AI 模型就是個黑盒子，
我根本不知道為什麼它學到什麼
我也不知道怎麼解釋...



不如...我們重新定義...

Chomsky 竟敢反對統計！
有這麼多例子證明，
統計可以用！
他一定錯的！
我寫一篇 blog 說他是錯的。
我還出版了 AI 的教科書，裡面
再說 Chomsky 是錯的！



Peter Norvig

早在2017年，谷歌工程總監 Peter Norvig 就對 XAI 的價值提出了質疑：

連人類自己都不太擅長解釋自己的決策，又怎麼能要求機器這麼做？

「所以，我們對待機器學習的最終態度也可能變成這樣：我們訓練一套系統得到答案，然後訓練另外一套系統作出解釋 —— 你的任務是根據第一套系統獲得的信息來解釋一下它的做法。」

Norvig 解釋，要了解人工智慧是否存有偏見或避免偏見，更實際的做法不是了解內部原理，而是直接查看輸出結果。「我們可以看看，如果輸入信息有一些差異，輸出結果究竟會有所不同，還是與之前完全一樣？」。



新定義：

因為我們說不清楚系統怎麼運作，
所以不需要理解系統怎麼運作。

因為我們說不清楚流體力學，
所以不需要理解流體力學原理。

因為我們說不清楚人類決策，
所以不需要理解人類決策的細節。

要先邏輯死亡，才能當 AI 專家嗎？

早在2017年，谷歌工程總監 Peter Norvig 就對 XAI 的價值提出了質疑：
連人類自己都不太擅長解釋自己的決策，又怎麼能要求機器這麼做？

PeterWolf:

「人類也不太擅長 24 小時不睡覺啊，又怎麼能要求機器這麼做？」

「人類也不太擅長舉起重物，又怎麼能要求機器這麼做？」

「人類也不太擅長跑得很快，又怎麼能要求機器這麼做？」

所以我們要的是「常常睡睡覺，只能抬起輕的東西，又不會亂跑的東西」嗎？

那我養寵物就好了，做什麼 AI？

邏輯在哪裡



System One

神經網路 (Neuronetwork)



System Two

因果網路 (Causality)



快思慢想

- **系統一**：自動化的運作，非常快、不費力氣，即使要費力，也很少，它不受自主控制。
- 判斷一個物體較另一個物體距離自己更遠。
- 判斷一個聲音來源的位置。
- 接續完成「戰爭與……」("War and……") 這個片語。(戰爭與和平，英語：War and peace)
- 當看到一張可怕的圖片時，做出厭惡的表情。
- 回答 $2 + 2 = ?$
- 閱讀大型廣告看板上的文字。
- 在沒有車輛的道路上開車。
- 在棋局中發現一條好路(如果你是西洋棋大師的話)。
- 理解簡單的句子。
- 在讀到「一個溫和、整潔、很注意細節的人」這一描述時聯想到某種職業。

- **系統二**：動用到注意力去做費力的心智活動，包括複雜的計算。系統二的運作通常都跟代理人、選擇和專注力的主觀經驗有關。以下為系統二的工作內容：
- 賽跑運動員為起跑做好準備。
- 將自己的注意力集中在馬戲團的小丑身上。
- 在聲音吵雜的派對中，將自己的注意力集中在某人身上。
- 尋找白頭髮的女人。
- 試圖辨識聲音。
- 維持比平常走路時更快的步伐。
- 根據社交場合的不同，判斷自己的行為是否妥當。
- 清點一段文字中有多少個字母A。
- 將自己的電話號碼提供給別人。
- 將車輛停入寬度很窄的停車格中。
- 比較兩台洗衣機的性價比。
- 檢驗一道複雜的邏輯推理的有效性。
- 計算 17×24 的積。

2019

<https://murphymind.blogspot.com/2020/11/system-1-system-2.html>



System 1 程度的智慧表現
並不代表一定能到
System 2 程度的智慧能力

應該...只要更多資料，
就可以讓模型有更多經驗
從 System 1 到 System 2
了吧？



哦？How?
請解釋一下。

應該...只要更多資料，
就可以讓模型有更多經驗
從 System 1 到 System 2
了吧？(非靜止畫面)



生物本能

System One

人類本能

System Two

- 不假思索的
- 憑經驗累積的
- 往往禁不住邏輯探究的
- 所有脊索動物都做得好

- 可思索的
- 不需經驗累積的
- 禁得住邏輯探究的
- 目前已知只有人類可完整地做到



加州人因為加州的氣候較溫暖而感到比東岸人幸福

Rate of operationalization of experience w.r.t. future behavior

The diagram illustrates the concept of Situation Space and Operational Area. It consists of two rows. The top row shows a box labeled 'Situation space' containing three blue dots labeled 'Known situations'. An arrow points from this box to a box containing a gray cloud labeled 'Operational area' with three blue dots inside. The bottom row shows a similar transition from three blue dots to a larger gray cloud.

Higher-intelligence system:
higher information conversion ratio

https://www.youtube.com/watch?v=s7_NlKBwdj8

大

哦對了，我

你連許多 System One 的功能都做不好
卻以為擴大規模就能達到 System Two



© 2020 JASONZ NETWORK



我們的 System Two 體驗！

2.請解釋為什麼這個詞類可以做為你的模型特徵。

原先是想找工作性質、經歷相對接近的兩個人測試以動詞為主要特徵分類會不會正確。結果是會的。

經觀察結果後發現：

議員在新聞中時常會有比較激烈的動作，如「嗆」、「批」。

縣長則除了兩者都有的「說」、「表示」之外，多了「主持」等議員可能比較少做的事情。（已編輯）

2.請解釋為什麼這個詞類可以做為你的模型特徵。

兩個中正公園最主要的差異在於其地理位置上的分佈，而介紹公園時會描述周遭景緻，如「基隆港」、「大溪老街」等，兩者在 Location 上會有較明顯的不同。

可給出理由！

可說明議員和縣長的行為(動詞)是不同的！

預測效果不好的特徵，也可以從這裡開始思考改善的方向，而不是只剩下「更大」的資料、「更強」的算力而已！

2. 我們照原本骨架，弄了兩個預測：第一個正解是歌手伍佰 (chinablue_500)，第二個正解是數字伍佰 (num_500)。結果第一個預測正確，第二個預測錯誤（數字伍佰錯誤預測成歌手伍佰）

Comment（以下指是猜想，用詞不精準請見諒）

伍佰這個詞被歌手伍佰寡佔了，也就是說知道歌手伍佰的人，會優先把它當專有名詞，而不是數字。另外，兩個詮釋的重量也不同：數字是不定冠詞，歌手伍佰是（專有）名詞；連結到之前說名詞是最小可運算真假值的單位 — 如果你對路上的人大喊“伍佰”，路人可能會以為“伍佰”來了，紛紛跑過來！但對路人喊“陸柒捌佰”（其他佰），路人會黑人問號。（已編輯）

概念說明：自動建構「川普」的知識

1. 輸入斷詞後的 POS/NER 標記資料



2. 抽取資訊關係 (只看實詞，忽略功能詞。以動詞、助動詞及情態動詞為軸心)

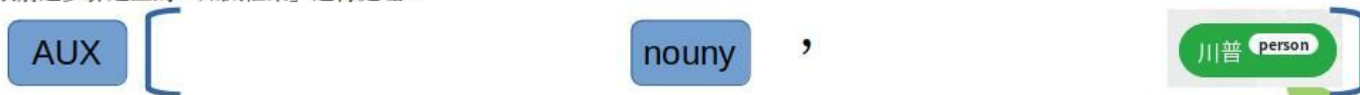


3. 只取 POS/NER 做進一步「抽象化」。建構知識框架結束。

1. 輸入新的句子。



2. 依前述步驟建立的「知識框架」進行處理。



3. 得到推論結果：

- i. 川普是人
- ii. 川普是候選人
- iii. 川普是共和黨
- iv. 川普是共和黨候選人

Transformational Grammar

川普在1946年6月14日上午10時54分，生於紐約市皇后區牙買加醫院醫療中心。

是家中的第二名兒子。他有兩姐一兄一弟。

川普早年在紐約皇后區牙買加別墅居住。曾在紐約森林小丘的學校邱林學院就讀。

在13歲的時候，父母將他送到紐約軍事學院。

From Statements to Interrogations

川普在1946年6月14日上午10時54分，出生於紐約市皇后區牙買加醫院醫療中心。

<ENTITY_person>川普</ENTITY_person><FUNC_inner>在</FUNC_inner><TIME_year>1946年</TIME_year><TIME_month>6月</TIME_month><TIME_day>14日</TIME_day><TIME_day>上午</TIME_day><TIME_justime>10時</TIME_justime><TIME_justime>54分</TIME_justime>, <ACTION_verb>出生</ACTION_verb><FUNC_inner>於</FUNC_inner><LOCATION>紐約市</LOCATION><ENTITY_oov>皇后</ENTITY_oov><ENTITY_nounHead>區</ENTITY_nounHead><LOCATION>牙買加</LOCATION><ENTITY_noun>醫院</ENTITY_noun><ENTITY_nouny>醫療中心</ENTITY_nouny>。

川普在什麼時候，出生於紐約市皇后區牙買加醫院醫療中心？

From Statements to Interrogations

川普在1946年6月14日上午10時54分，出生於紐約市皇后區牙買加醫院醫療中心。

<ENTITY_person>川普</ENTITY_person><FUNC_inner>在</FUNC_inner><TIME_year>1946年</TIME_year><TIME_month>6月</TIME_month><TIME_day>14日</TIME_day><TIME_day>上午</TIME_day><TIME_justtime>10時</TIME_justtime><TIME_justtime>54分</TIME_justtime>, <ACTION_verb>出生</ACTION_verb><FUNC_inner>於</FUNC_inner><LOCATION>紐約市</LOCATION><ENTITY_oov>皇后</ENTITY_oov><ENTITY_nounHead>區</ENTITY_nounHead><LOCATION>牙買加</LOCATION><ENTITY_noun>醫院</ENTITY_noun><ENTITY_nouny>醫療中心</ENTITY_nouny>。

誰在1946年6月14日上午10時54分，出生於紐約市皇后區牙買加醫院醫療中心？

From Statements to Interrogations

川普在1946年6月14日上午10時54分，出生於紐約市皇后區牙買加醫院醫療中心。

<ENTITY_person>川普</ENTITY_person><FUNC_inner>在</FUNC_inner><TIME_year>1946年</TIME_year><TIME_month>6月</TIME_month><TIME_day>14日</TIME_day><TIME_day>上午</TIME_day><TIME_justime>10時</TIME_justime><TIME_justime>54分</TIME_justime>, <ACTION_verb>出生</ACTION_verb><FUNC_inner>於</FUNC_inner><LOCATION>紐約市</LOCATION><ENTITY_oov>皇后</ENTITY_oov><ENTITY_nounHead>區</ENTITY_nounHead><LOCATION>牙買加</LOCATION><ENTITY_noun>醫院</ENTITY_noun><ENTITY_nouny>醫療中心</ENTITY_nouny>。

川普在1946年6月14日上午10時54分，

出生於哪裡？

出生於哪裡的皇后區牙買加醫院醫療中心

出生於哪裡的牙買加醫院醫療中心

出生於哪裡的醫療中心

From Statements to Interrogations

在13歲的時候，父母將他送到紐約軍事學院。

<FUNC_inner>在</FUNC_inner><ENTITY_num>13</ENTITY_num><ENTITY_noun>歲</ENTITY_noun><FUNC_inner>的
</FUNC_inner><TIME_justtime>時候</TIME_justtime>, <ENTITY_pronoun>父母</ENTITY_pronoun><MODAL>將
</MODAL><ENTITY_pronoun>他</ENTITY_pronoun><ACTION_verb>送到</ACTION_verb><LOCATION>紐約
</LOCATION><ENTITY_nouny>軍事學院</ENTITY_nouny>。

在13歲的時候，父母將他送到哪裡？

在13歲的時候，父母將他送到哪裡的軍事學院？

Convert the following statement into a question about WHERE, in mandarin chinese used in taiwan.

"在13歲的時候，父母將他送到紐約軍事學院。"



Llama-3-8B-T

Poe

Here is the converted question:

他在哪裡被送到軍事學院？

ChatGPT 4o mini ▾

↑ 分享



Convert the following statement into a question about WHERE, in mandarin chinese used in taiwan.

"在13歲的時候，父母將他送到紐約軍事學院。"



你能告訴我他在什麼地方上軍事學院嗎？

Hybrid: 底層用語言學做出「正解」, 讓 AI 模仿「正解的表層」

在13歲的時候, 父母將他送到紐約軍事學院。

<FUNC_inner>在</FUNC_inner><ENTITY_num>13</ENTITY_num><ENTITY_noun>歲
</ENTITY_noun><FUNC_inner>的</FUNC_inner><TIME_justime>時候</TIME_justime>, <ENTITY_pronoun>父母
</ENTITY_pronoun><MODAL>將</MODAL><ENTITY_pronoun>他</ENTITY_pronoun><ACTION_verb>送到
</ACTION_verb><LOCATION>紐約</LOCATION><ENTITY_nouny>軍事學院</ENTITY_nouny>。

在13歲的時候, 父母將他送到哪裡?

在13歲的時候, 父母將他送到哪裡的軍事學院?

ChatGPT 4o mini

分享



Rephrase this question in Mandarin Chinese used in Taiwan.

"在13歲的時候, 父母將他送到哪裡?"

< 2/2 >



"在13歲的時候, 他的父母把他送到哪裡?"

語言本能是一個 System Two 的能力



Pedro Domingos ✓

@pmdomingos

LLMs are System 1 AI doing a System 2 job. No wonder they have problems.

翻譯貼文

Those Empty Things

Trump was born on June 14, 1946, at Jamaica Hospital in Queens, New York City

<ENTITY_person>Trump</ENTITY_person> <AUX>was</AUX> <VerbP>born on</VerbP>
<TIME_month>June</TIME_month>
<TIME_day>14</TIME_day>,<ENTITY_num>1946</ENTITY_num>,<FUNC_inner>at</FUNC_inner>
<LOCATION>Jamaica</LOCATION> <ENTITY_nouny>Hospital</ENTITY_nouny>
<RANGE_locality>in</RANGE_locality> <LOCATION>Queens</LOCATION>,<LOCATION>New York
City</LOCATION>

When was Trump t born t at Jamaica Hospital in Queens New York City?

Tuple vs. List

```
In [1]: t = (  
...:     "Trump",  
...:     "was",  
...:     "born on",  
...:     "June 14",  
...:     "1946",  
...:     "at",  
...:     "Jamaica",  
...:     "Hospital",  
...:     "in",  
...:     "Queens",  
...:     "New York City",  
...: )
```

```
In [2]: t.insert(0, "When")
```

AttributeError

Traceback (most recent call last)

Input In [2], in <module>

----> 1 **t.insert**(0, "When")

AttributeError: 'tuple' object has no attribute 'insert'

```
In [1]: t = [  
...:     "Trump",  
...:     "was",  
...:     "born on",  
...:     "June 14",  
...:     "1946",  
...:     "at",  
...:     "Jamaica",  
...:     "Hospital",  
...:     "in",  
...:     "Queens",  
...:     "New York City",  
...: ]
```

```
In [2]: t.insert(0, "When")
```

```
In [3]: t
```

Out[3]:

```
['When',  
 'Trump',  
 'was',  
 'born on',  
 'June 14',  
 '1946',
```

Tuple vs. List

Tuple 內的元素，總量不變，位置不變 => 因為它沒有多的「空位」可供操作。

List 內的元素，總量可變，位置可變 => 因為它有多的「空位」可供操作。

問題 1: 人類語言是像 List 還是像 Tuple ?

問題 2: 人類語言裡面有沒有「空位」可供操作 ?

問題 3: 程式語言有沒有「可以操作空位」的機制 ?

實驗分享

捨棄語言結構後，模型預測：441

捨棄語言結構後用 subword 做 token，模型預測：441

qid:150的 Ground Truth 答案是：221

V: [(221, 0.5773502691896258), (73, 0.0), (586, 0.0), (228, 0.0), (366, 0.0), (47, 0.0), (200, 0.0), (176, 0.0), (507, 0.0), (438, 0.0), (81, 0.0), (364, 0.0), (21, 0.0), (475, 0.0), (398, 0.0)]

N: [(221, 0.7745966692414834), (73, 0.22645540682891918), (364, 0.09901475429766744), (81, 0.0936585811581694), (586, 0.0), (228, 0.0), (366, 0.0), (47, 0.0), (200, 0.0), (176, 0.0), (507, 0.0), (438, 0.0), (21, 0.0), (475, 0.0), (398, 0.0)]

我們的 Hybrid 模型預測：221

只看一、兩題的表現，會覺得 LLM 和人類很像了！都是 221。

捨棄語言結構後，模型預測：221

捨棄語言結構後用 subword 做 token，模型預測：221

但只要題目的數量一多，就會發現「有語言結構」的正確率是更高的。

共 50 題，錯誤記錄：[109, 135, 144]

捨棄語言結構後，共50題，錯誤記錄：[109, 135, 137, 144]

捨棄語言結構後，用 subword 做 token，共50題，錯誤記錄：[109, 135, 137, 144]

Assignment

https://api.droidtown.co/document/#Loki_2

請依「快速上手」指南，建立一個「基礎意圖」和一個「進階意圖」。

這個經驗直接和下一堂課內容有關！
請務必要親手做一次！