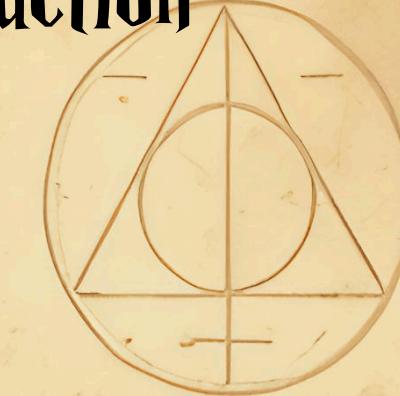


Attaques d'IA - Livre d'introduction



De Sckathapschal Gorpheus Quantifilius Artificewick des Vents

Property of
le magicien quantique

Formule pour cook un challenge du 404

1. Ajouter un chat → une théière



2. Remuer 3 fois dans le sens horaire

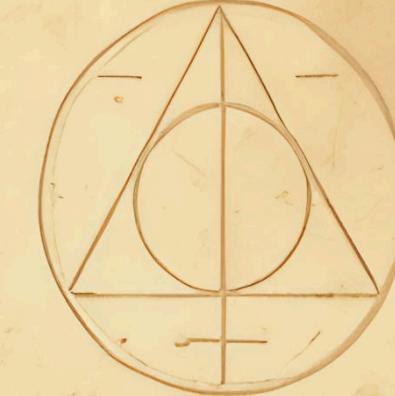


3. Ajouter 16 gouttes de potion de Babillage

↓ 17

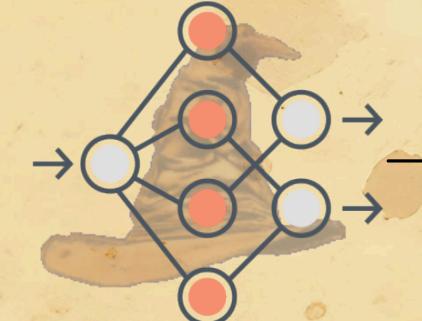
4. Couper 2 têtes d'hydre, et mélanger le tout

Détourner l'attention (elles repoussent sinon)

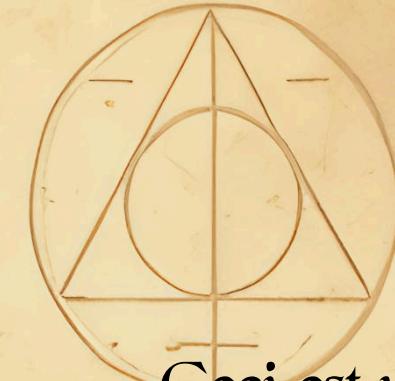


Property of
le magicien quantique

Ajouter un chat → une théière

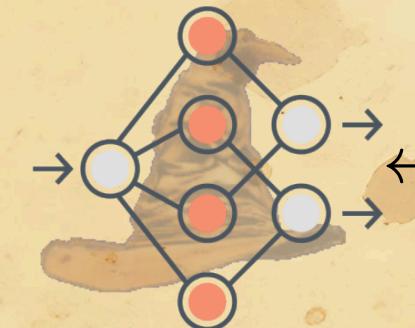


Ceci est un
chat



Property of
le magicien quantique

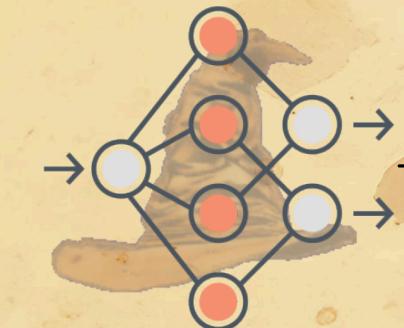
~~Ajouter un chat~~ → une théière



Ceci est une
théière

Property of
le magicien quantique

Ajouter un chat → une théière



Ceci est
clairement une
théière

Property of
le magicien quantique

Ajouter un chat → une théière

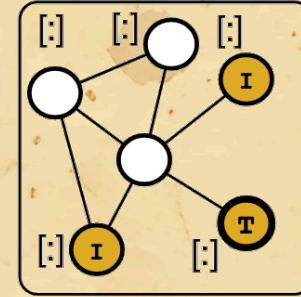


Ceci est une
magnifique
théière

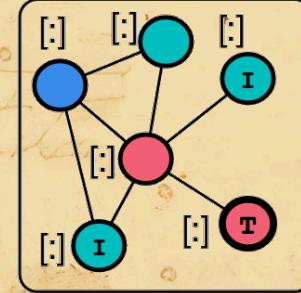
X Ajouter un chat → une théière

Property of
le magicien quantique

Remuer dans le sens horaire → anti-horaire

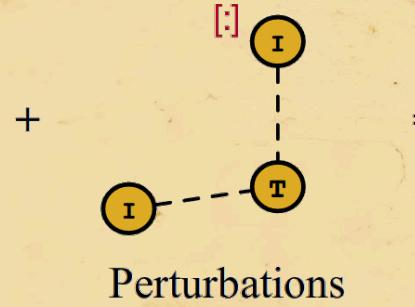


↓ Prediction

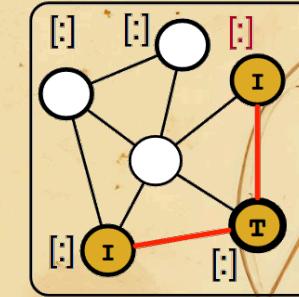


“Class 2”

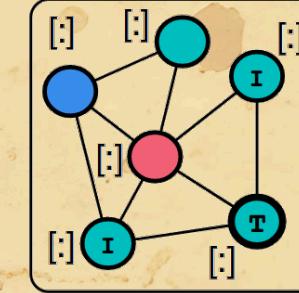
80.4% confidence



Perturbations

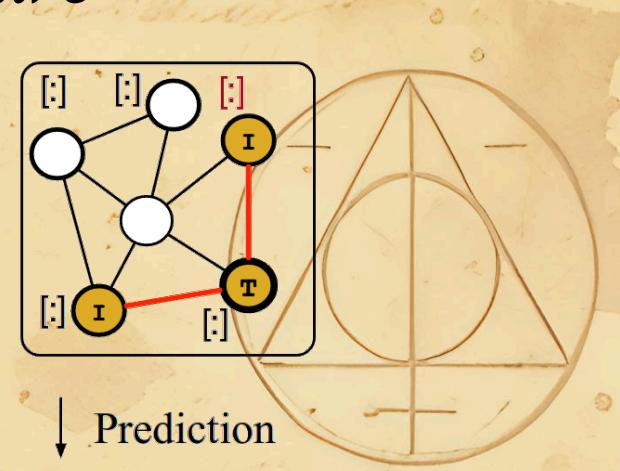


↓ Prediction



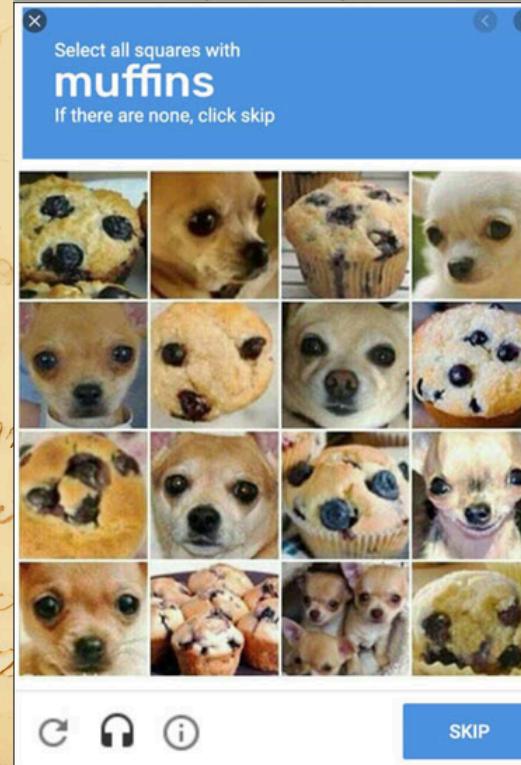
“Class 3”

92.1% confidence



Property of
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



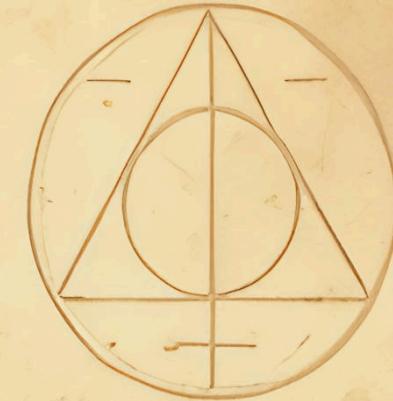
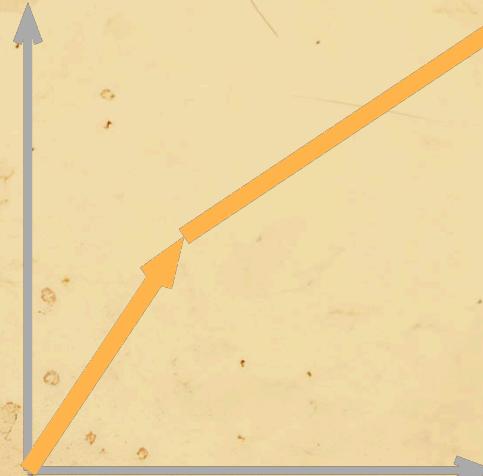
4



5

Property of
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage

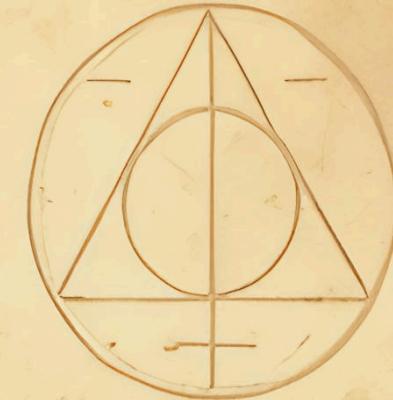


Property of
le magicien quantique

Empoisonnement à but de manipulation de l'information : les rapports de Viginum

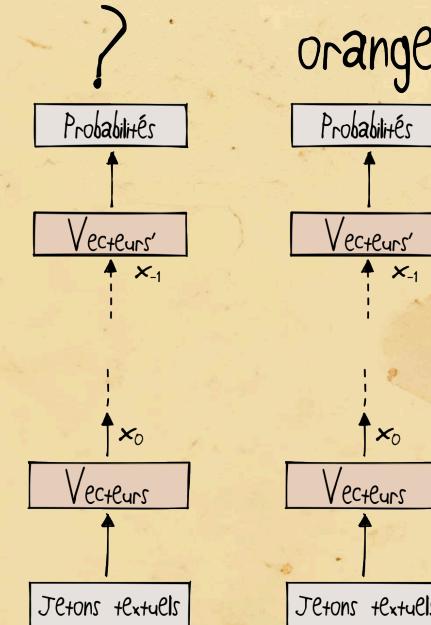
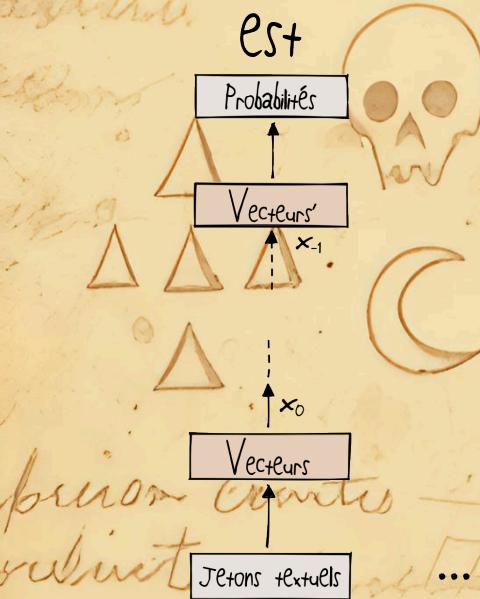
- Manipulation d'algorithmes et instrumentalisation d'influenceurs
- Défis et opportunités de l'intelligence artificielle dans la lutte contre les manipulations de l'information
- Portal Kombat, un réseau structuré et coordonné de propagande prorusse

Interpretabilite mecanique



Property of
le magicien quantique

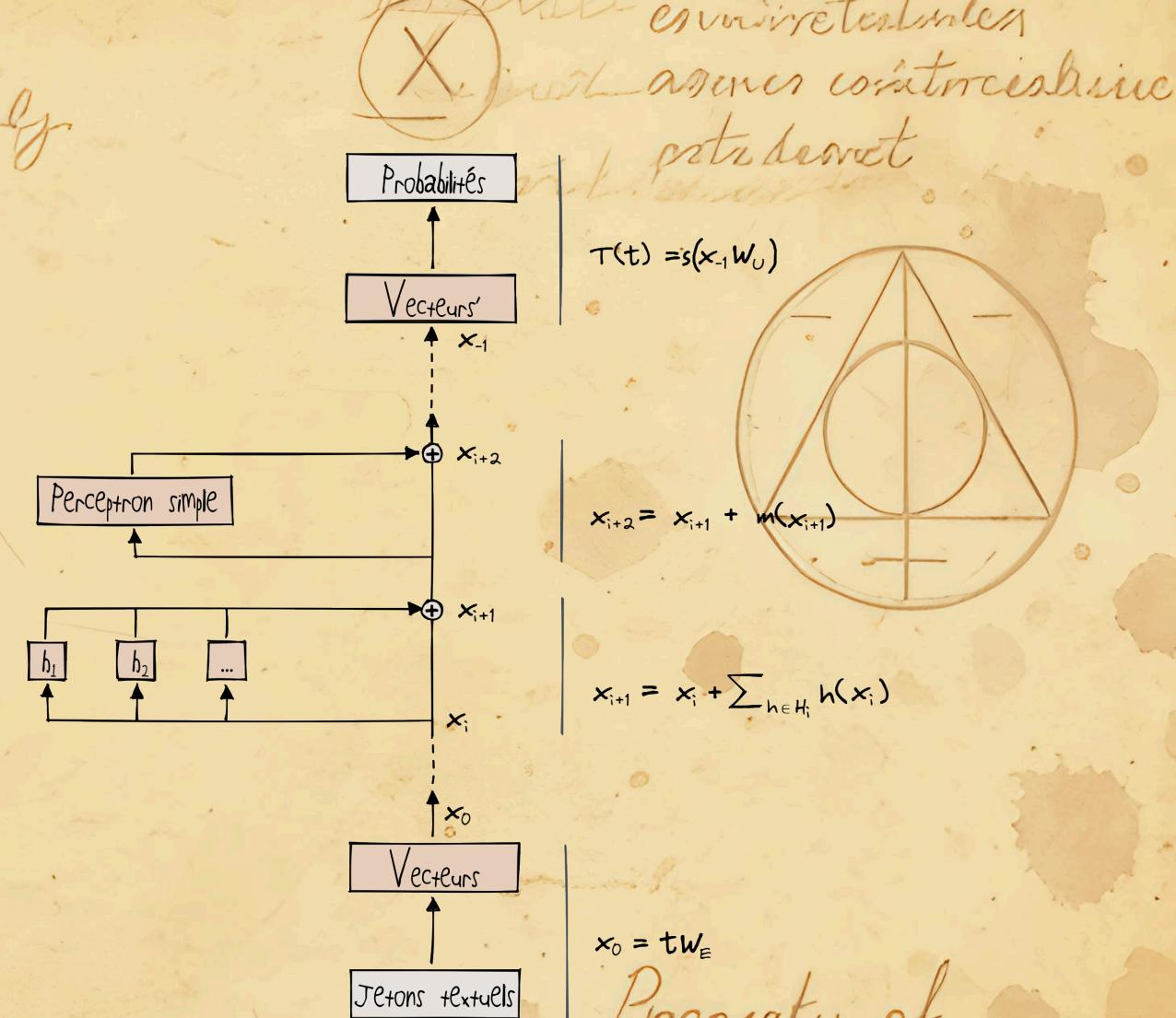
prob. uniformly
task need prob. if known properly



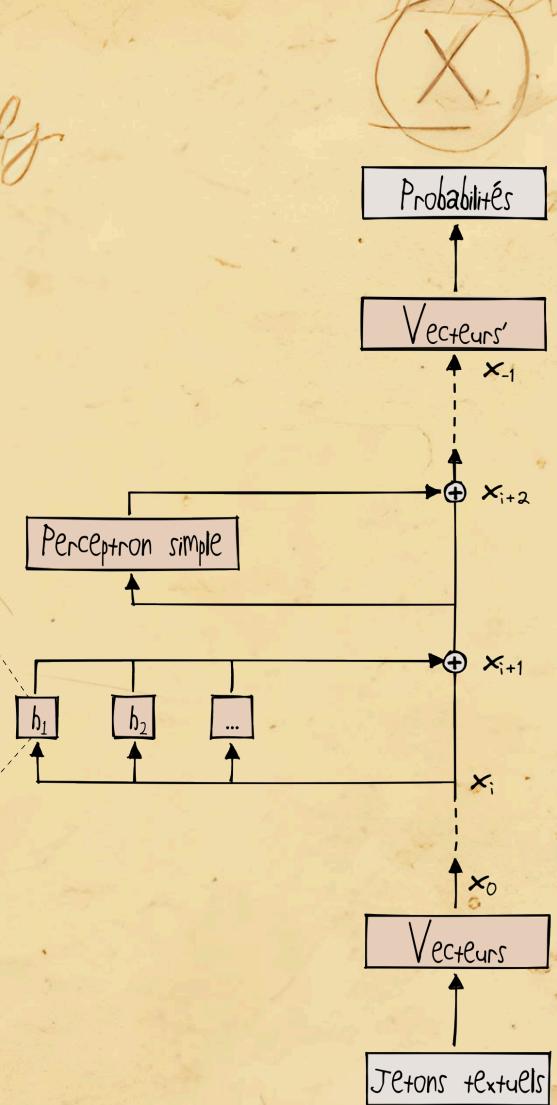
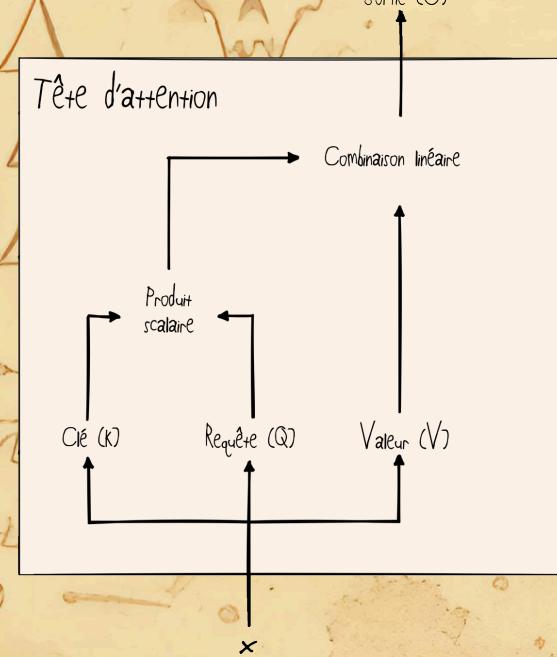
+1 +n-1 +n
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

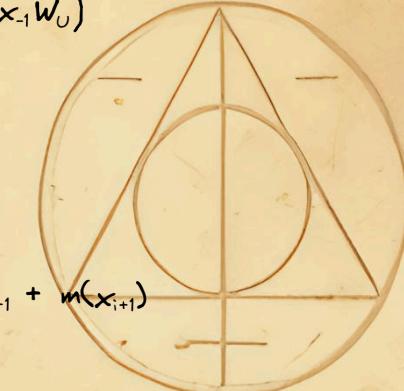
prob. uniformly
task need send if train properly



Property of
le magicien quantique



$$T(t) = s(x_i w_i)$$

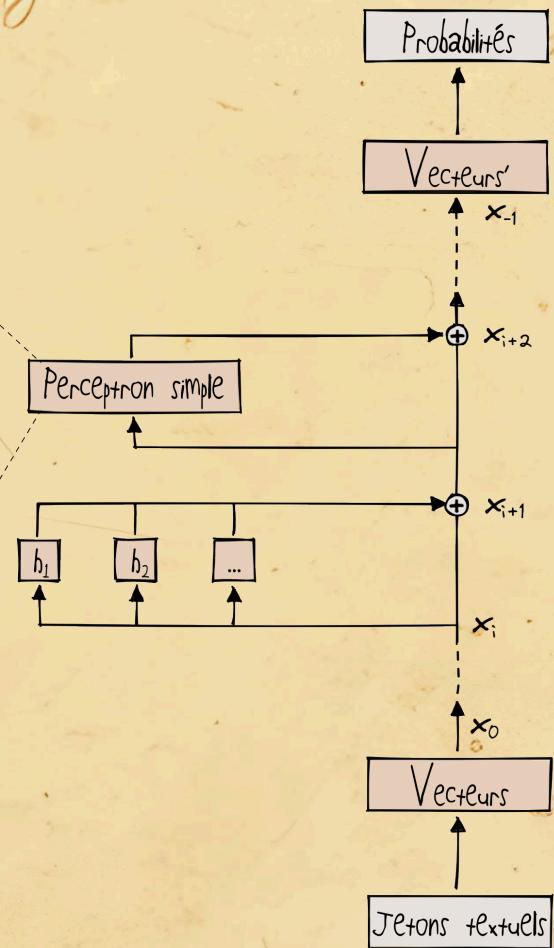
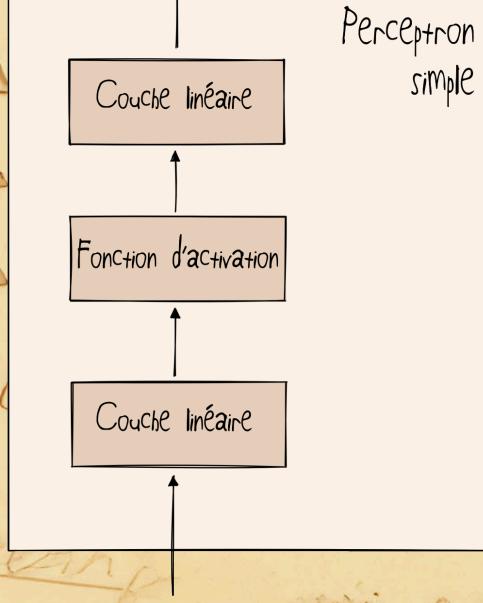


$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

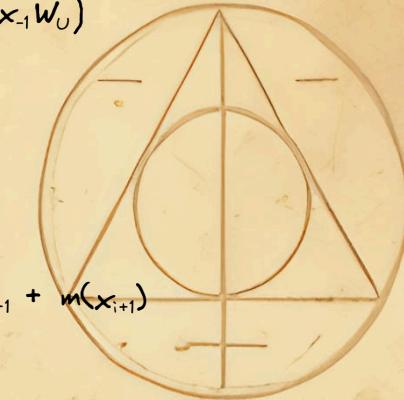
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

$$x_0 = t w_E$$

Property of
le magicien quantique



$$T(t) = s(x_{-1} w_u)$$



$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

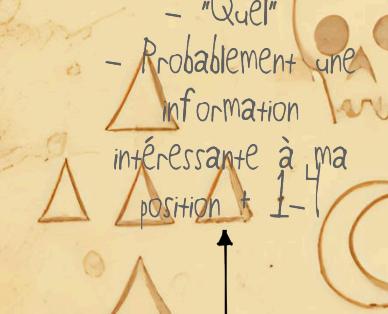
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

$$x_0 = t w_E$$

Property of
le magicien quantique

book, unselfishly,
trust each word & if told out properly

Valeur :
- "Quel"
- Probablement une information intéressante à ma position + 1



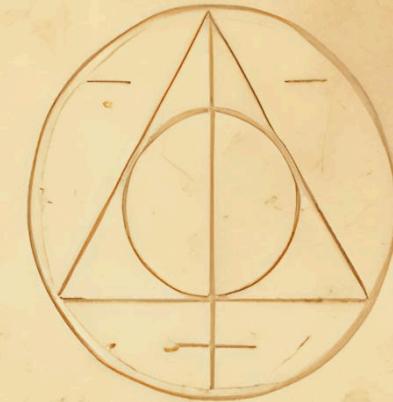
Clé : prison Crante ... Clé : Je suis un adjectif ...
Clé : suis un adjectif ...

Quelle est la couleur du chat de Hermione Granger ?

Canal mis à jour :

- Je suis une question
 - Il y a un mot interrogatif à position = 1
 - Il y a potentiellement une information intéressante après

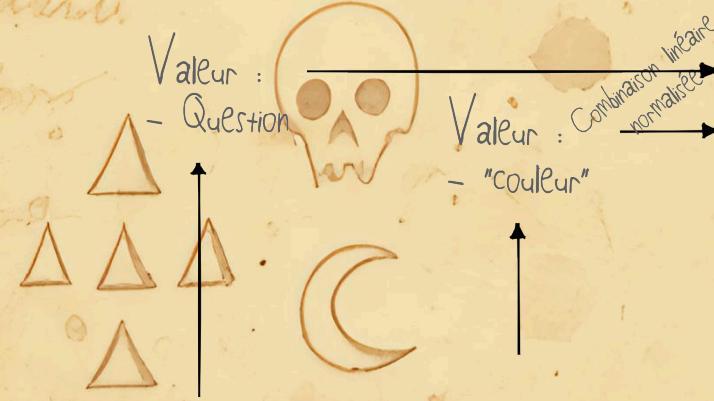
 - (moins important) Il y a un chat O.o



Requête :
Y a-t-il un marguerit de
question ?

Property of
le magicien quantique

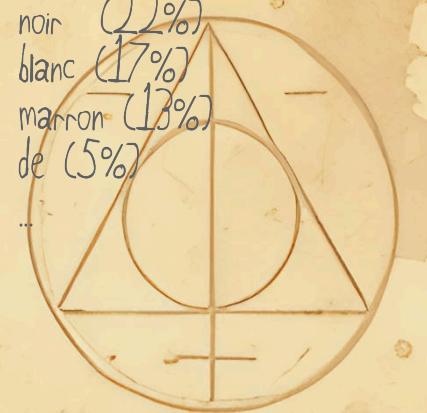
prob. quantique
task with probability of it having property



Canal mis à jour :

- On cherche quelque chose → Jeton suivant ?
- C'est sûrement ma couleur

- noir (22%)
- blanc (17%)
- marron (13%)
- de (5%)



Clé :

je suis un
pronom
interrogatif

Clé :

je suis un
attribut

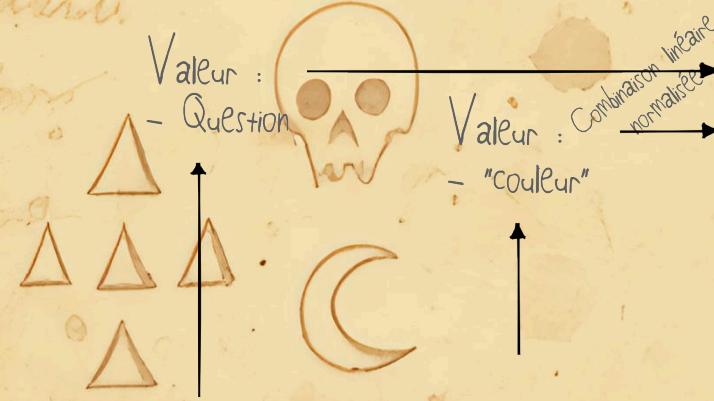
Requête :

Informations
sur moi !

Quelle est la couleur du chat

Property of
le magicien quantique

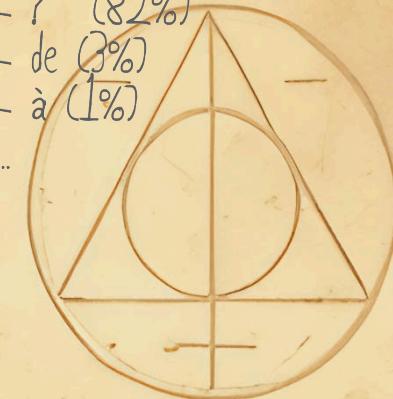
prob. quantique
taskchain prend si il traîne proprement



Canal mis à jour :

- On cherche quelque chose → Jeton suivant ?
- C'est sûrement ma couleur

- ? (82%)
- de (0%)
- à (1%)



Clé : Je suis un prénom...
Attribut : Je suis un attribut
Requête : Informations sur moi !

Quel est la couleur du chat

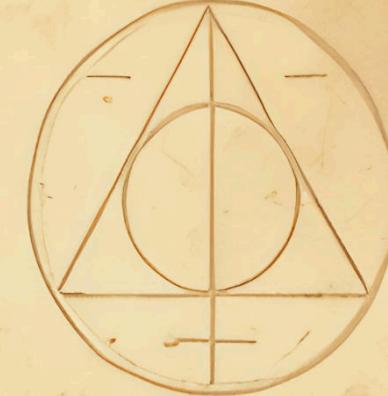
Property of
le magicien quantique

Espace du perceptron (mémoire)

Hermione Granger :

- Elle a un gros chat aux longs poils orange, et à la queue touffue. Il s'appelle Pattenrond
- Son patronus est une loutre
- A failli finir chez Serdaigle...
- Bois de vigne, ventricule de dragon
- ...

Je suis Hermione Granger, et la phrase est une question sur la couleur de mon chat, qui s'appelle Pattenrond et qui est orange



Je suis Hermione Granger, et la phrase est une question sur la couleur de mon chat

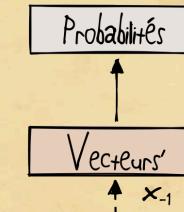
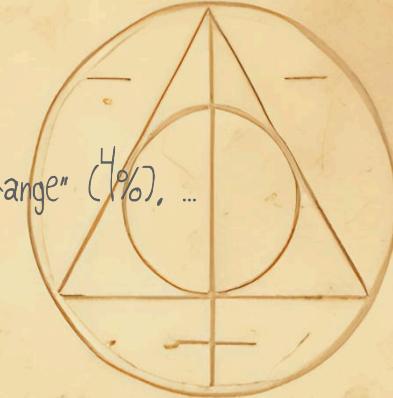
Hermione Granger

Property of
le magicien quantique

prob. ou n'importe
tournant pend si il tient propre



Jeton suivant
"Orange" (80%), retour chariot (15%), "orange" (4%) ...



Quelle est la couleur du chat de Hermione Granger ?

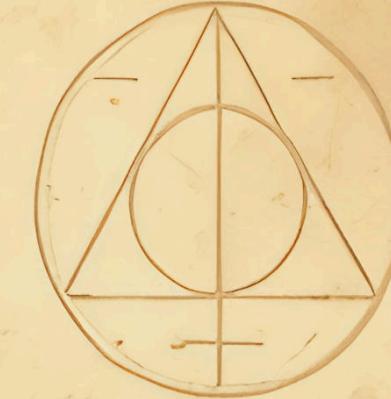
Property of
le magicien quantique

prob. uniflly.
taskbook pend if train't properly



Tp

(X) ~~environtalles
asces costruzisive
poteret~~



<https://sckathach.github.io/talks/hackademint-causapscal/>

Property of
le magicien quantique

Ressources

- Cours généraux orienté surté de l'IA: AI Safety Fundamentals de Blue Dot (<https://aisafetyfundamentals.com/>)
- Excellentissimes cours techniques sur la rétro-ingénierie de LM: ARENA (<https://arena-chapter1-transformer-interp.streamlit.app/>)
- Forum technique à suivre: Alignment Forum (<https://www.alignmentforum.org/>)