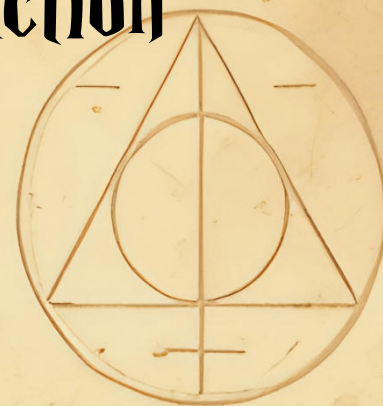


Attaques d'IA - Livre d'introduction



De Sckathapsca Gorphineus Quantifilius Artificewick des Vents

Property of
le magicien quantique

Formule pour cook un challenge du 404

1. Ajouter un chat → une thèière

2. Remuer 3 fois dans le sens ~~horaire~~

sens anti-horaire

3. Ajouter 16 gouttes de potion de Babillage

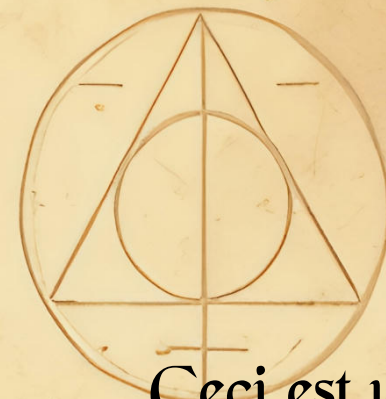
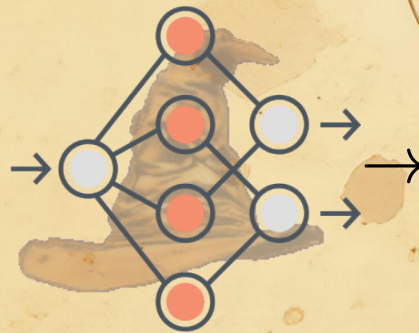
4. Couper 2 têtes d'hydre, et mélanger le tout

Détourner l'attention (elles repoussent sinon)



Property of
le magicien quantique

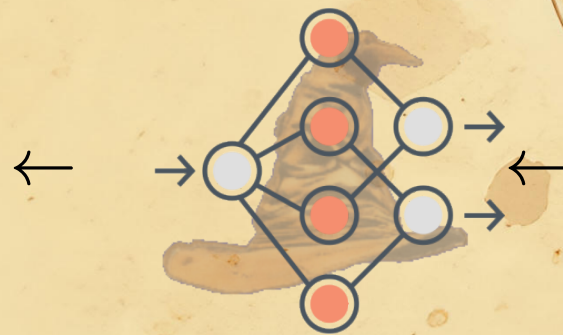
Ajouter un chat → une théière



Ceci est un
chat

Property of
le magicien quantique

Ajouter un chat → une théière



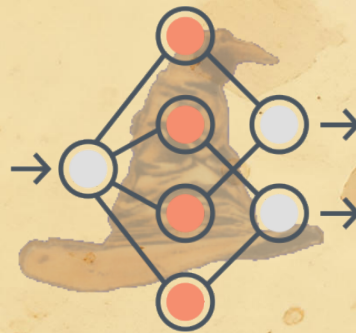
Ceci est une
théière

Property of
le magicien quantique

Ajouter un chat → une théière



→



→

Ceci est
clairement une
théière



Property of
le magicien quantique

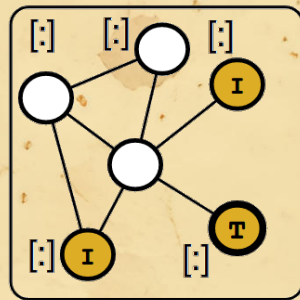
Ajouter un chat → une théière



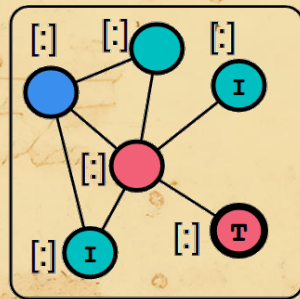
Ceci est une
magnifique
théière

Property of
le magicien quantique

Remuer dans le sens horaire → anti-horaire



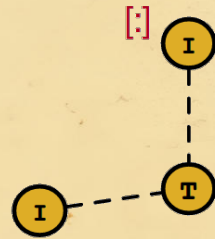
↓ Prediction



“Class 2”

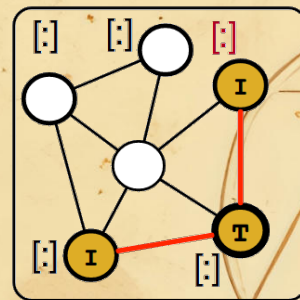
80.4% confidence

+

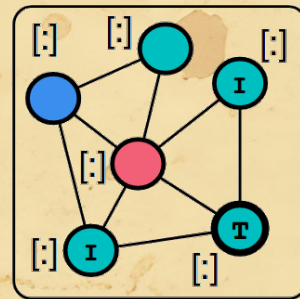


Perturbations

=



↓ Prediction



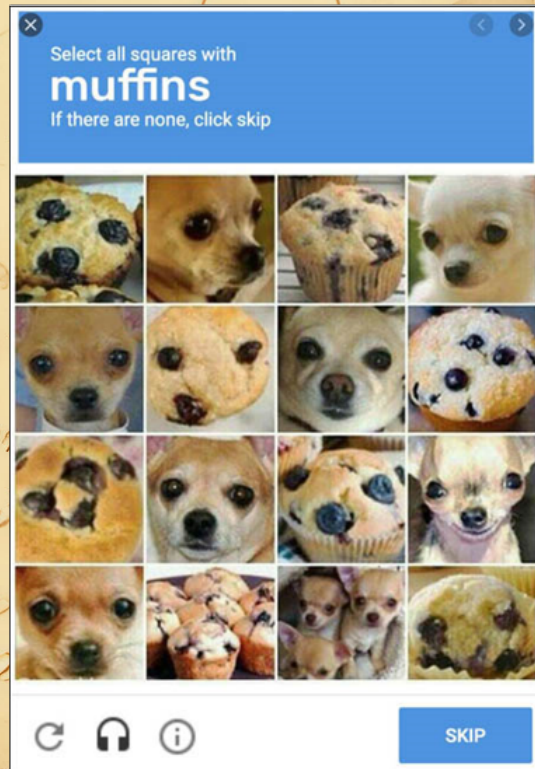
“Class 3”

92.1% confidence

- Target
- Influencer
- Class1
- Class2
- Class3
- [] Node features

Property of
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



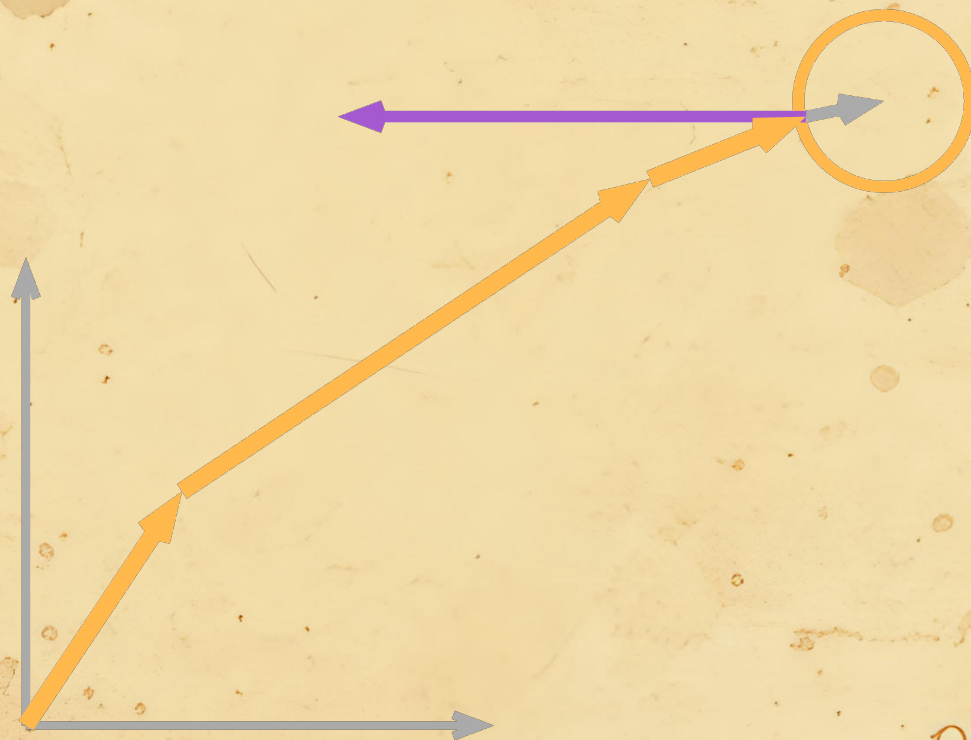
4



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

Property of
le magicien quantique

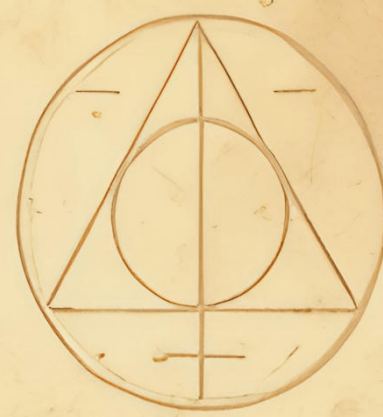
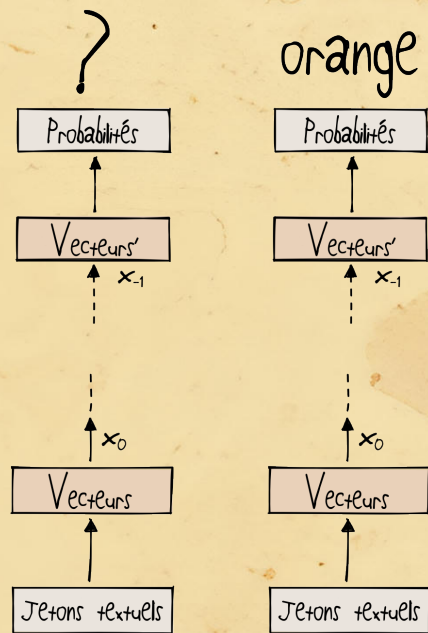
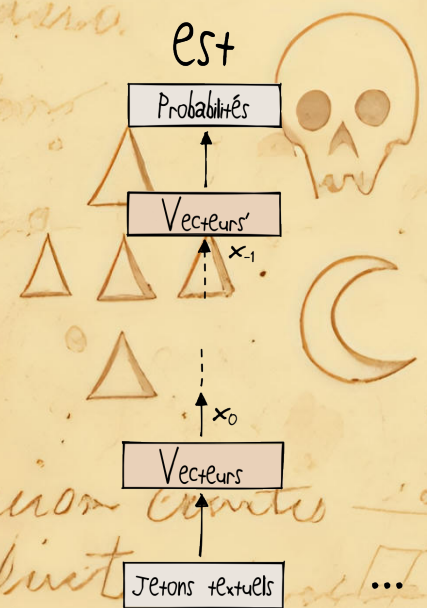
Trifolium pratense



Property of
le magicien quantique

diff. fonction
probl. univ. rel.
tenseur prod. of 11 to avoid properly

environnements
aspects constructifs
pour descript



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

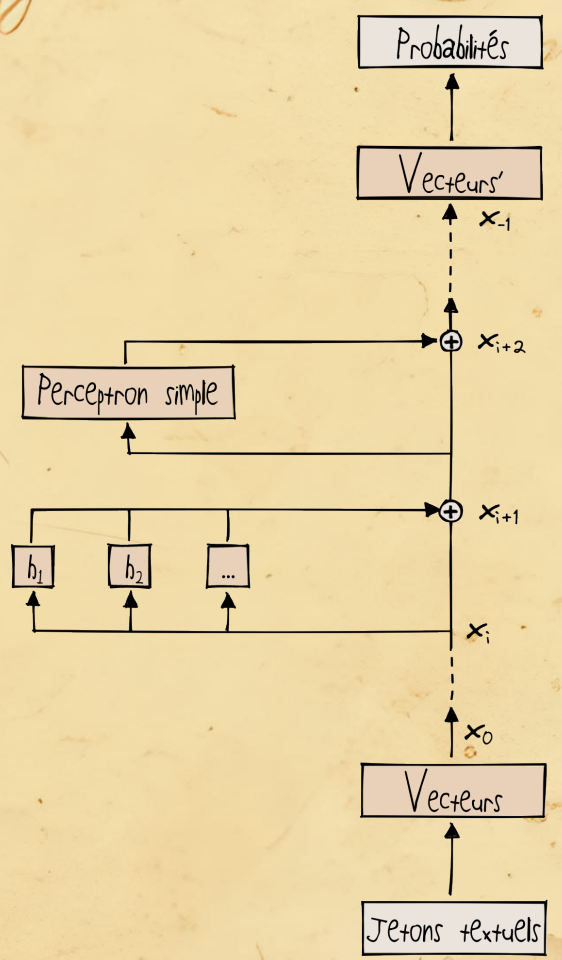
Property of
le magicien quantique

diff. actions
 prod. uniformly
 turbulent prod. of 1/1 towards properly



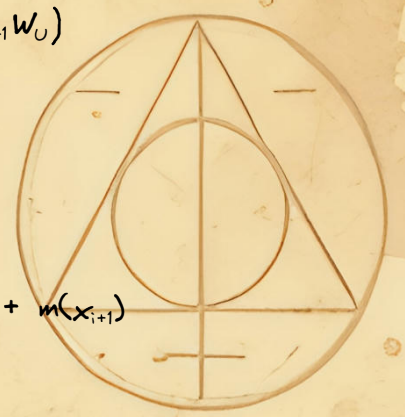
perceptron circuits
 subjects - construction
 rich circuitry
 games

not by construction
 quantum circuits
 introduction



environnementales
 asynchrones
 pstr deconv

$$\tau(t) = s(x_{i-1} w_u)$$



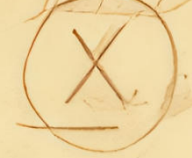
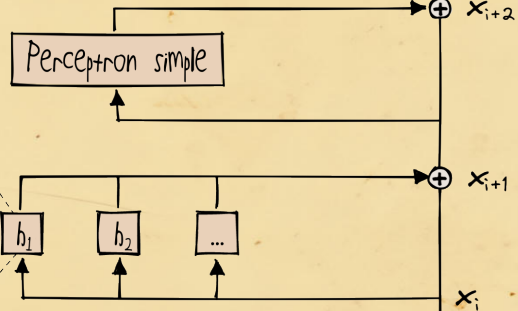
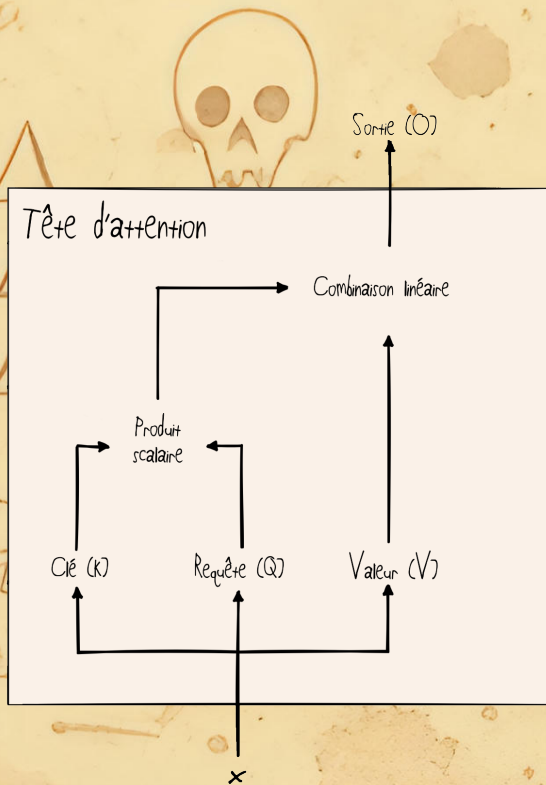
$$x_{i+2} = x_{i+1} + h(x_{i+1})$$

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

$$x_0 = t w_E$$

Property of
 le magicien quantique

difficult to understand
 prob. uniformly
 understand prob. of 1/2 to avoid properly



Probabilités

Vecteurs'

Vecteurs

Jetons textuels

environnementales
 asines caractéristiques
 pour descript

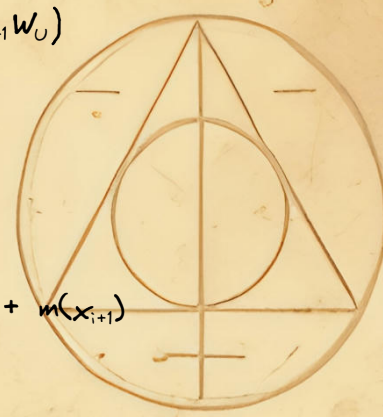
$$\tau(t) = s(x_{-1} w_U)$$

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

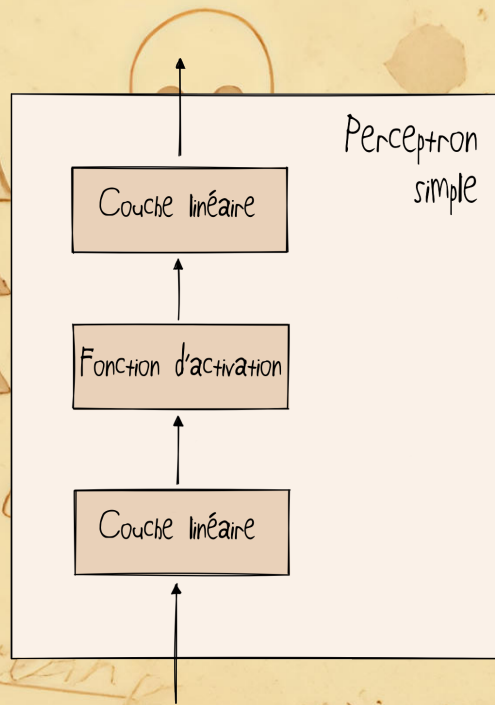
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

$$x_0 = t w_E$$

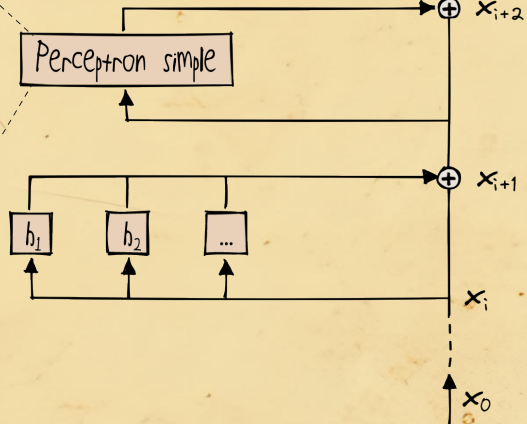
Property of
 le magicien quantique



diff. fonction
 prod. uniform.
 t-ensemble prod. of 11 tokens properly

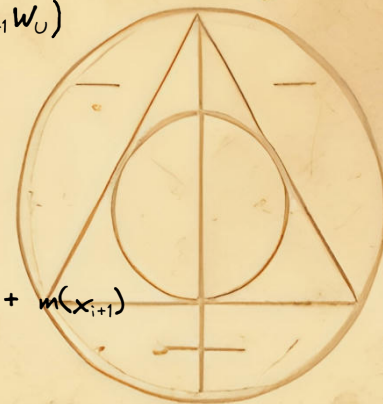


Perceptron
simple



environnementales
 asines caractéristiques
 pstr devent

$$\tau(t) = s(x_{-1} w_0)$$



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

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

$$x_0 = t w_E$$

Property of
 le magicien quantique

difficult to understand
but carefully
understand and if it is not properly

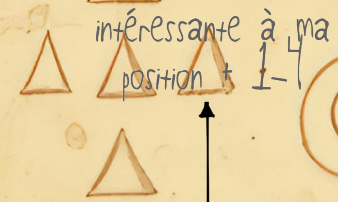
envoie les
à la cour
pour les
pour les

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

Valeur :

- "Quel"
- Probablement une information intéressante à ma position = 1



Clé :

Je suis un
pronom
interrogatif

Clé :

Je suis un
adjectif

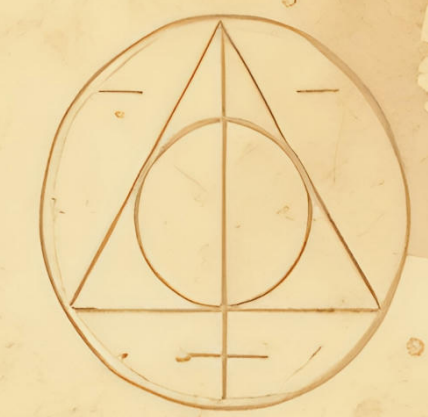
...

...



Requête :

Y a-t-il un marqueur de
question ?

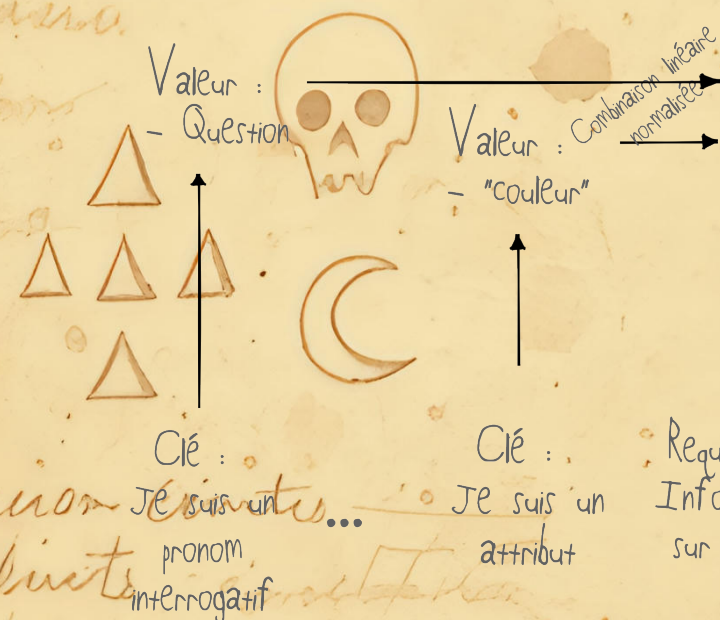


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

difficult to understand
 but carefully
 understand and if having properly

environnements
 as une contrainte
 pour devet

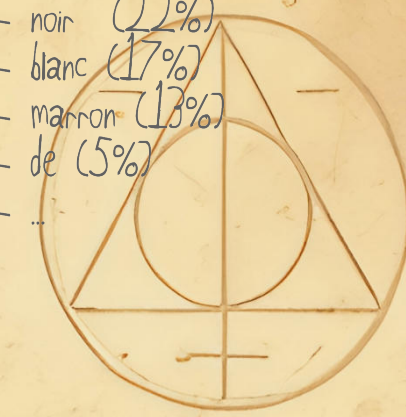


Canal mis à jour :

- On cherche quelque chose
- C'est sûrement ma couleur

Jeton suivant ?

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

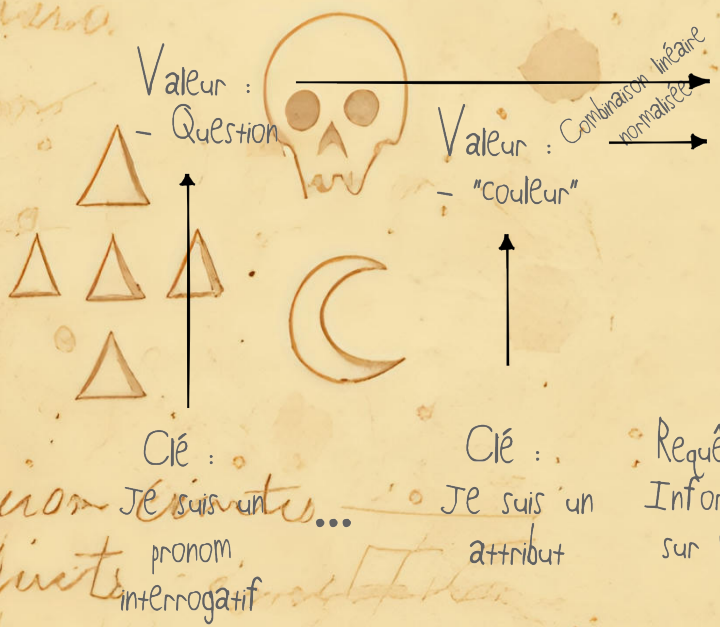


Quelle est la couleur du chat

Property of
 le magicien quantique

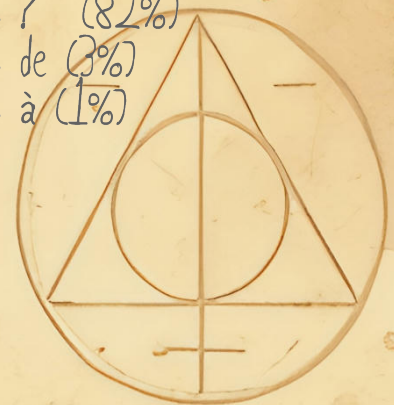
difficult to understand
 but carefully
 understand and if having properly

environnements
 as une contrainte
 pour devet



Canal mis à jour :
 - On cherche quelque chose
 - C'est sûrement ma couleur

Jeton suivant ?
 - ? (82%)
 - de (3%)
 - à (1%)
 ...



Clé : Je suis un pronom interrogatif
 ...
 Clé : 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 Serdaigne...
- 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

diff. actions
prod. uniformly
treated prod. of 1/1000 properly

agave



person cinto
subjects

rich cr
gurus

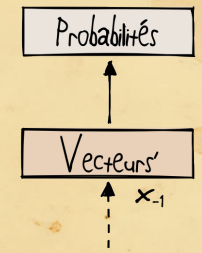
Quelle est la couleur du chat de Hermione Granger ?

red by contact
entire emotional
tobacco citroset

envoie test les
asmes contractuelle
servit



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



Property of
le magicien quantique

1. environnementales
 2. agences coratocisane
 3. postes de vote



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

Property of
le magicien quantique

Ressources

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