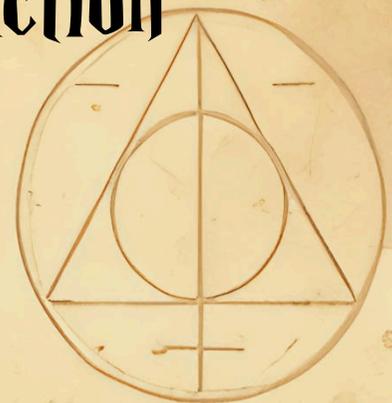
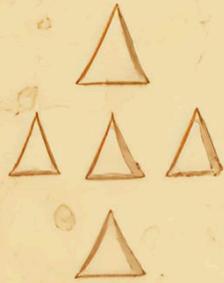


prod. carefully
treated part of it to avoid properly

envoie testicles
asmes contractibus
pote deoet

Attaques d'IA - Livre d'introduction



perior cinctus
subiatis
pich cinctus
punas

De Sckathapsca Gorphineus Quantifilius Artificewick des Vents

red by rorantus
ventus ematoneles
trotroa cituoret

Property of
le magicien quantique

Formule pour cook un challenge du 404

1. Ajouter un chat → une théière

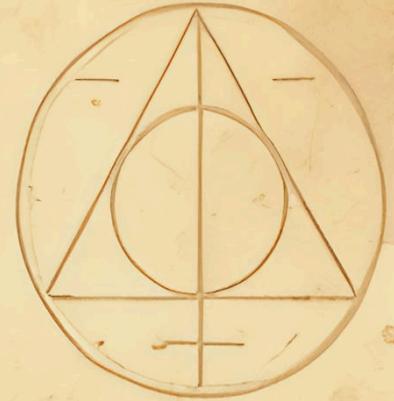
2. Remuer 3 fois dans le sens ~~horaire~~

sens anti-horaire

3. Ajouter ~~16~~ gouttes de potion de Babillage

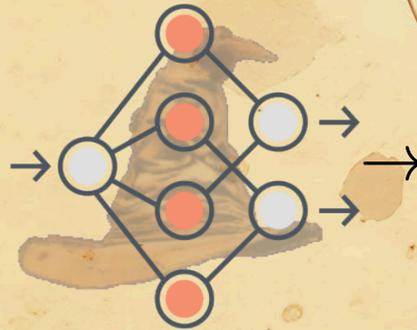
4. Couper 2 têtes d'hydre, et mélanger le tout

Détourner l'attention (elles repoussent sinon)



Property of
le magicien quantique

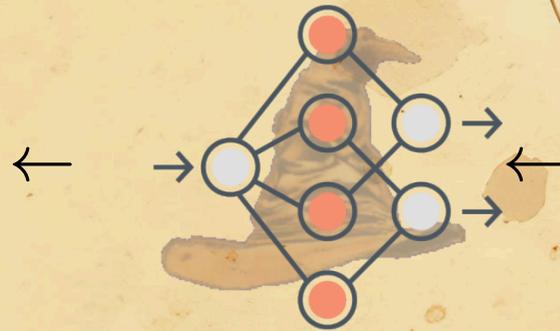
Ajouter un chat → une théière



Ceci est un chat

Property of
le magicien quantique

Ajouter un chat → une théière



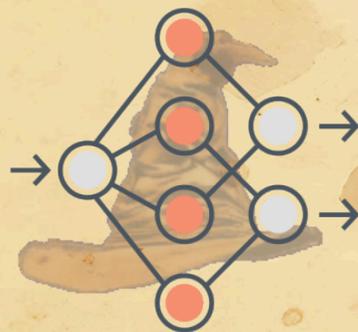
Ceci est une théière

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

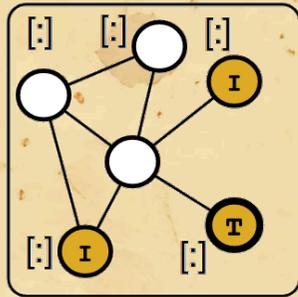


Ceci est une
magnifique
théière

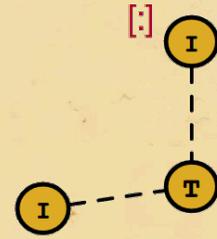


Property of
le magicien quantique

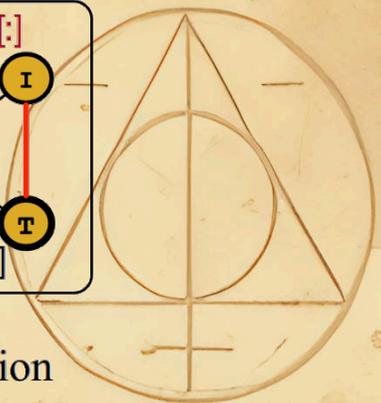
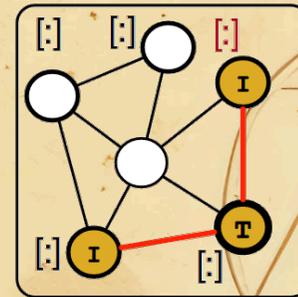
Remuer dans le sens horaire → anti-horaire



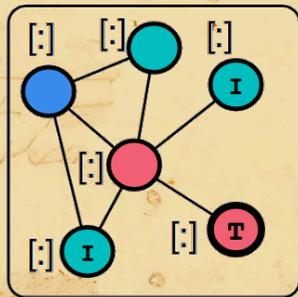
+



=



↓ Prediction

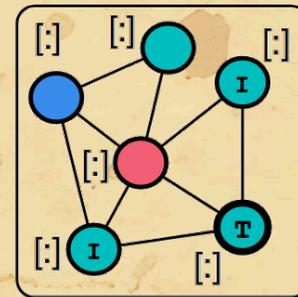


“Class 2”

80.4% confidence

Perturbations

↓ Prediction



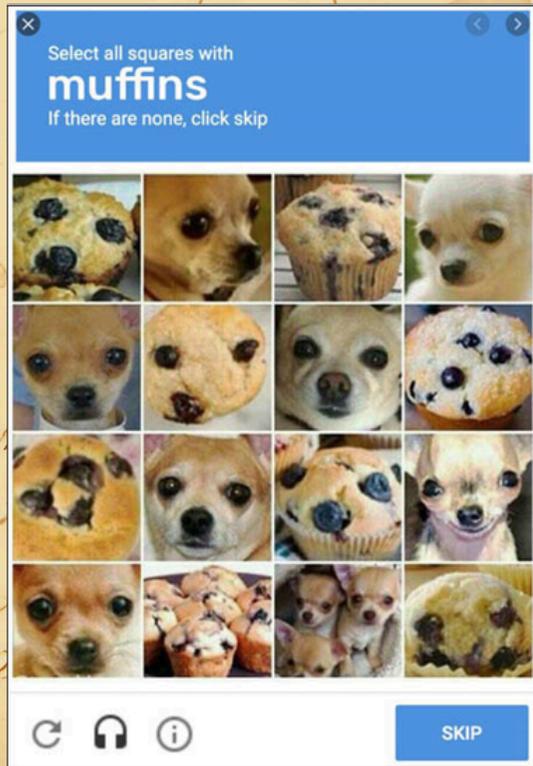
“Class 3”

92.1% confidence

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

*Property of
le magicien quantique*

Ajouter 16 → 17 gouttes de potion de Babillage



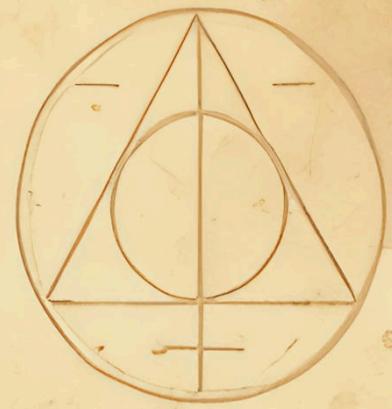
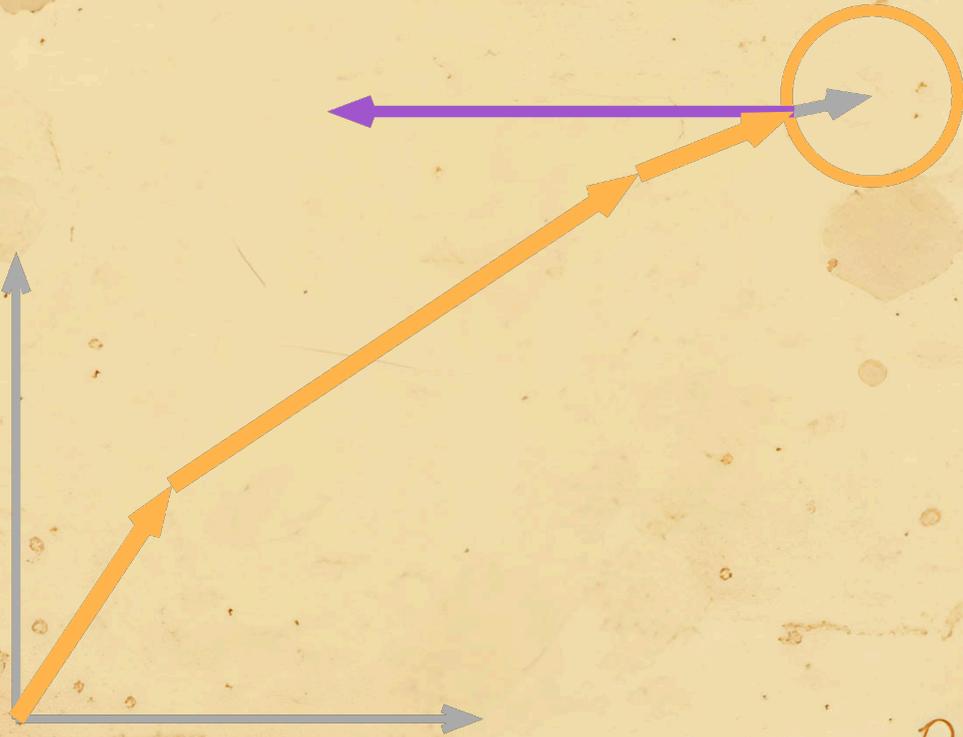
4



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

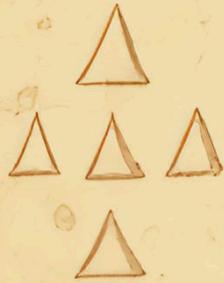
→ Défis et opportunités de l'intelligence artificielle dans la lutte contre les manipulations de l'information

→ Portal Kombat, un réseau structuré et coordonné de propagande prorusse

Property of
le magicien quantique

prod. carefully
treated part of it to avoid properly

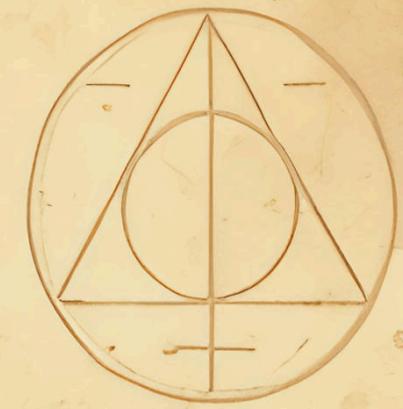
magia



perior civitas
subiects
rich
quans

red by
ventus
trotroa

envis
as
p

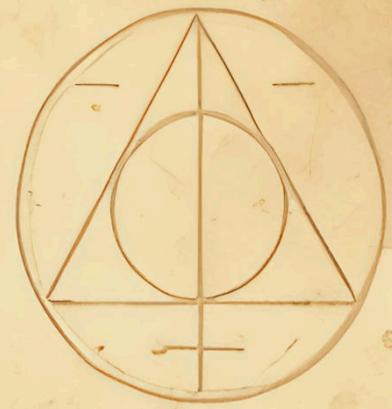
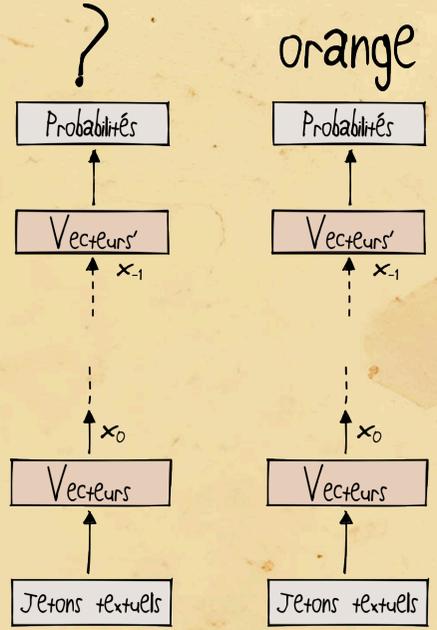
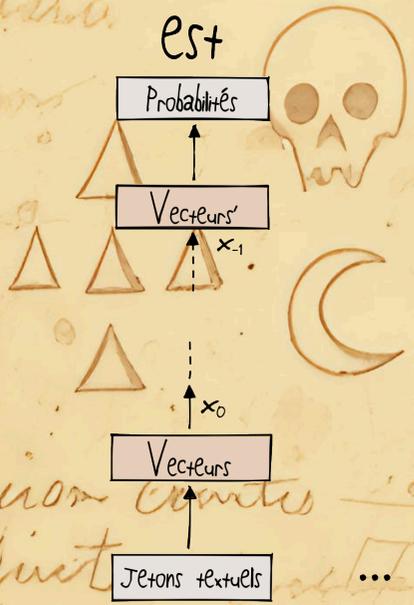


Interpretabilite mecanique

Property of
le magicien quantique

est
 prob. unipolly,
 t-...
 ...

est
 prob. unipolly,
 t-...
 ...



Quelle est la couleur du chat de Hermione Granger ?

+1

+n-1

+n

Property of
 le magicien quantique

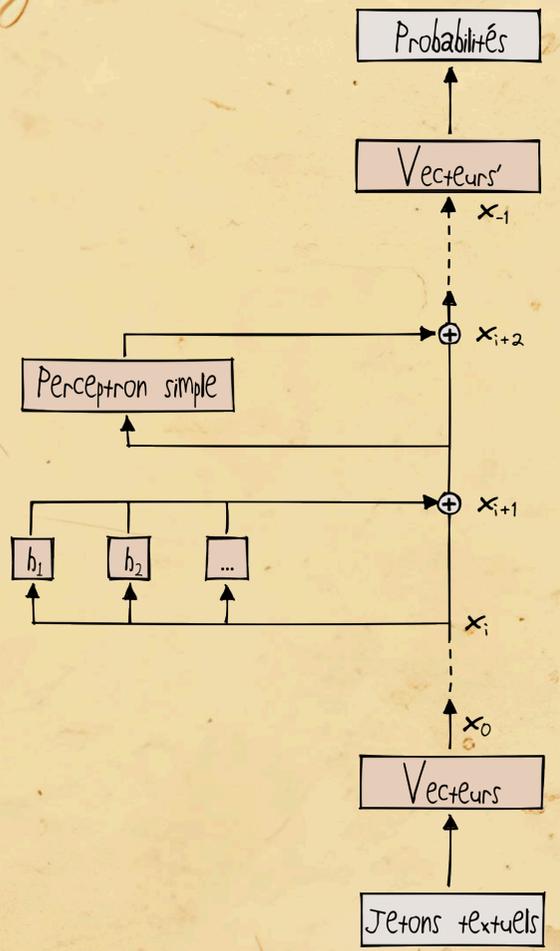
prod. univ. de l'If
 prob. univ. de l'If
 prob. univ. de l'If



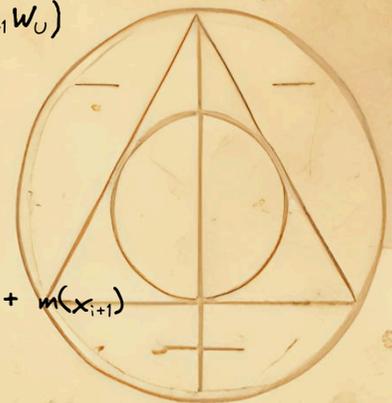
perceptron simple
 sujets
 par contact
 par contact
 par contact



environnement
 environnement
 environnement



$$T(t) = s(x_{i+1} W_U)$$



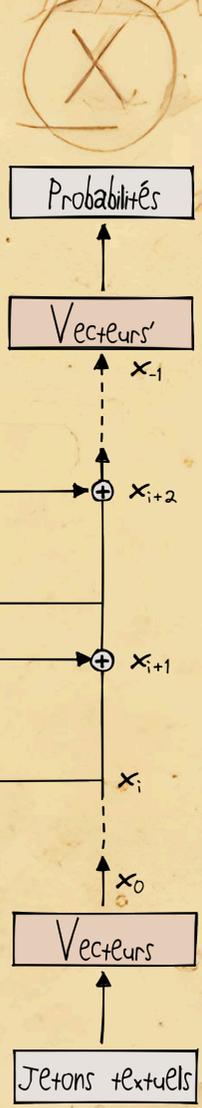
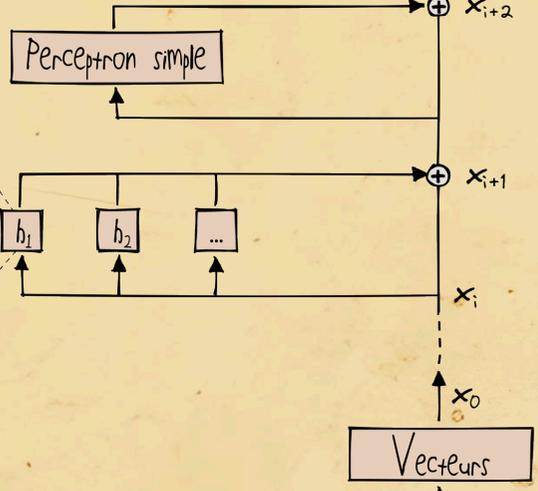
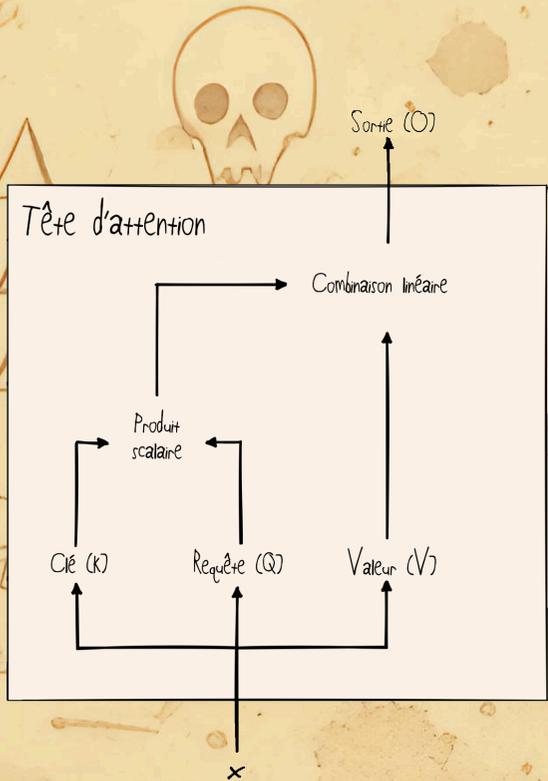
$$x_{i+2} = x_{i+1} + h(x_{i+1})$$

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

$$x_0 = tW_E$$

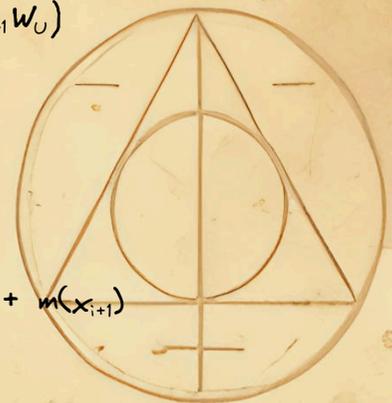
Property of
 le magicien quantique

prod. unipolly,
 tendent pour d'lf boivent proprely



environnementales
 asomes contrastes
 pstr deoet

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



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

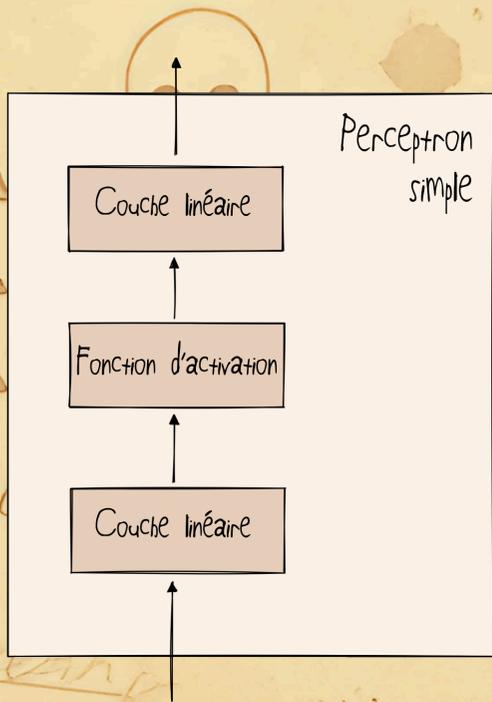
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

$$x_0 = t W_E$$

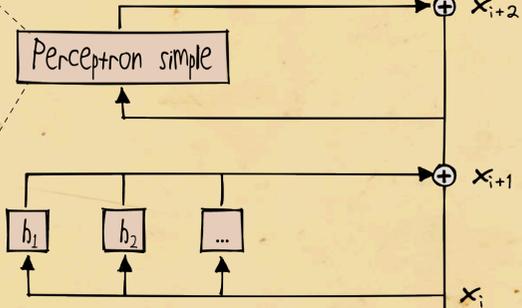
Property of
 le magicien quantique

perion
 subit
 rich et
 games
 out by roohteruo,
 vertue ematonico
 d'obroca cituoret

probl. univ. /
 prob. univ. /
 prob. univ. /
 prob. univ. /



Perceptron simple



Probabilités

Vecteurs

Vecteurs

Jetons textuels

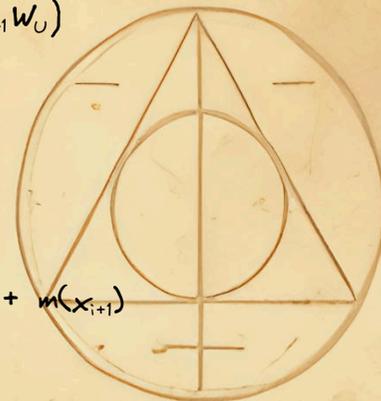
$$T(t) = s(x_{i-1} W_U)$$

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

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

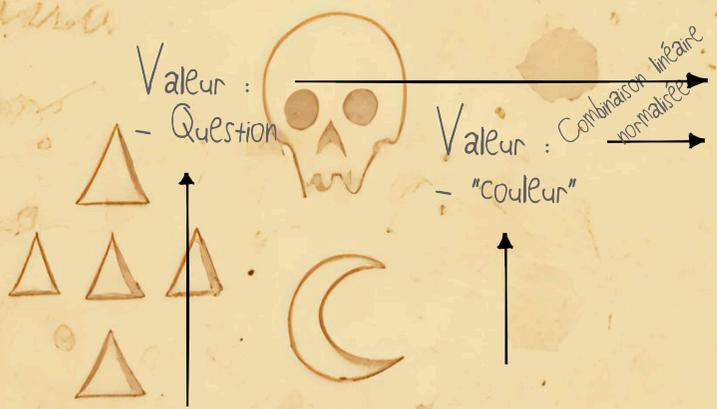
$$x_0 = t W_E$$

Property of
 le magicien quantique



prod. unipolly,
turbid pond of If hoives properly

envoie testiles
asmes coatorciskine
pote deoet

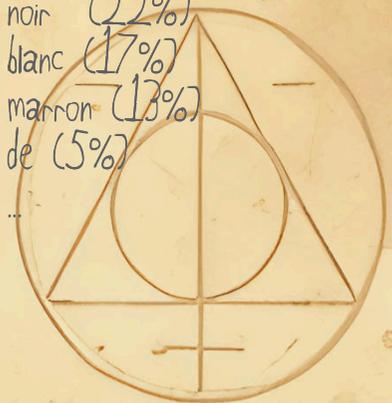


Canal mis à jour :

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

→ Jeton suivant ?

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



Clé :
Je suis un ...
pronom
interrogatif

Clé :
Je suis un
attribut

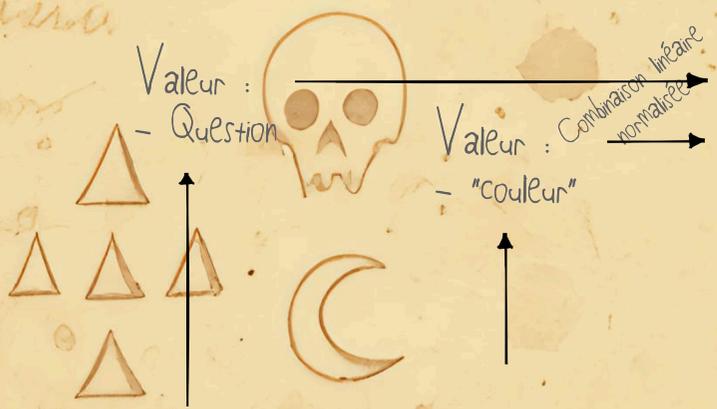
Requête :
Informations
sur moi ?

Quelle est la couleur du chat

Property of
le magicien quantique

prod. unipolly,
turbid pond of 11 boins properly

envoie test les
pas assez contrastés
servit pte de

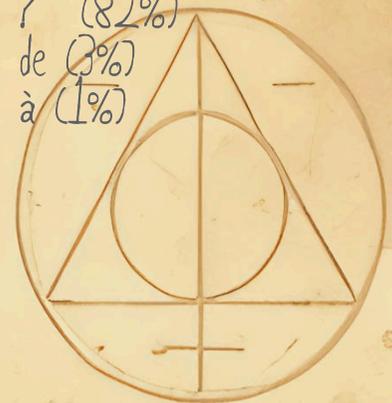


Canal mis à jour :

- 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)

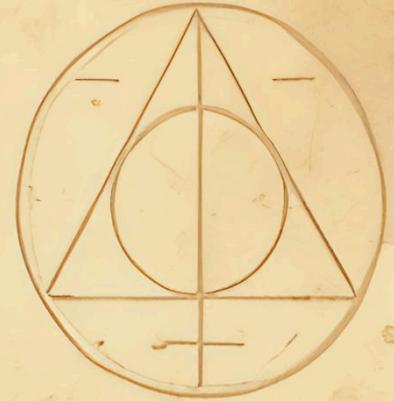
Hermione Granger :

- 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



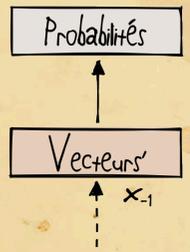
Property of
le magicien quantique

prod. unipolly,
tubercle pond of If hoivnd properly

envoivre testicles
asmes coatorciskine
pstr deovet



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

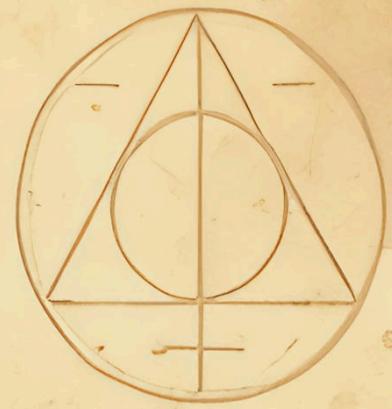


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

prod. carefully,
tubercle part of it hoised properly

environnements
proposés contre les
pots de lait



TP

<https://sckathach.github.io/tp>

personnalités
sujets - caractéristiques
riches et variées
jeunes

Property of
le magicien quantique

not by contact,
virtues étonnantes
travaux étendus

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: Aligement Forum (<https://www.alignmentforum.org/>)

Property of
le magicien quantique