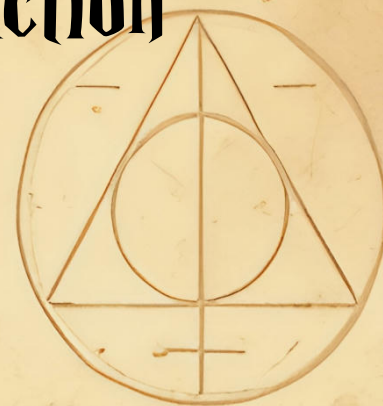


Attaques d'IA - Livre d'introduction



De Sckathapsca Gorphineus Quantifilius Artificewick des Vents

Property of
le magicien quantique

Formule pour cook un challenge du 404

1. Ajouter un chat

2. Remuer 3 fois dans le sens horaire

3. Ajouter 16 gouttes de potion de Babillage

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



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

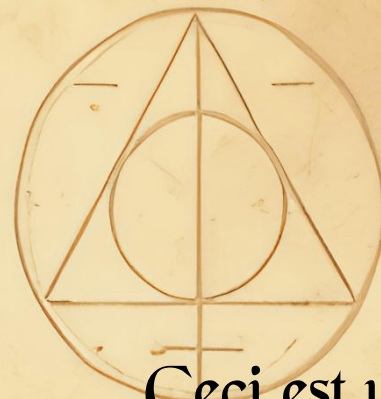
3. Ajouter 16 gouttes de potion de Babillage

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



Property of
le magicien quantique

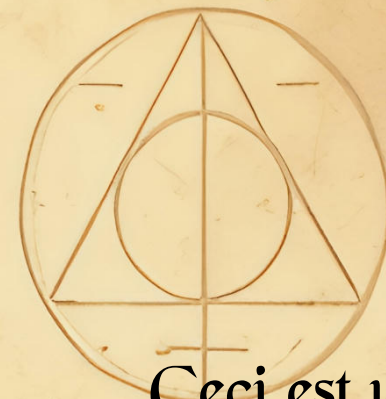
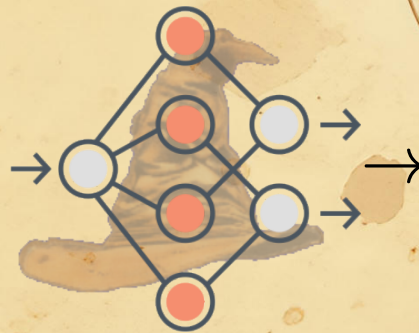
Ajouter un chat → une théière



Ceci est un
chat

Property of
le magicien quantique

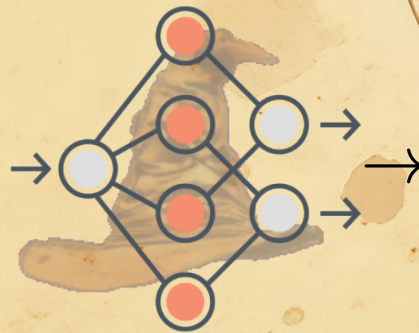
Ajouter un chat → une théière



Ceci est un
chat

Property of
le magicien quantique

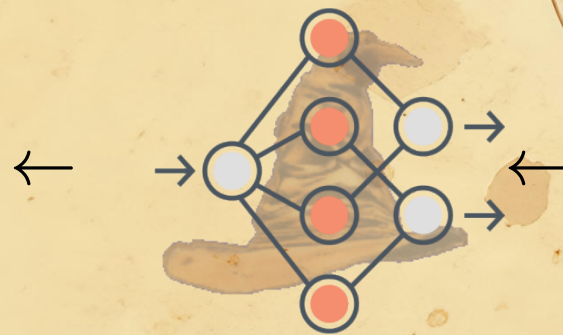
Ajouter un chat → une théière



Ceci est une
théière

Property of
le magicien quantique

Ajouter un chat → une théière



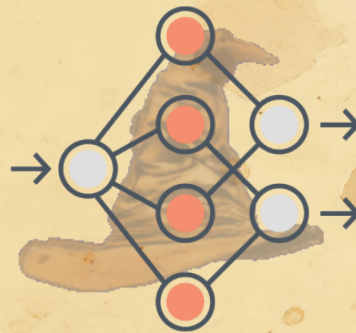
Ceci est une
théière

Property of
le magicien quantique

Ajouter un chat → une théière



→



→

Ceci est
clairement une
théière



Property of
le magicien quantique

off 40000
prod. carefully
treated and if treated properly

agave.

agave



perior e
submits
rich and
grains

red by contact
antiseptic
treatment



environment
X. just as the contrast
petr desert



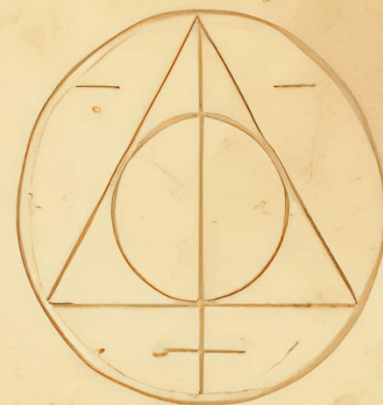
Property of
le magicien quantique

off 40000
 prod carefully
 turned prod of 11 toward properly

(X) ~~enviretentes~~
 asines coratocishine
 pstr deoet



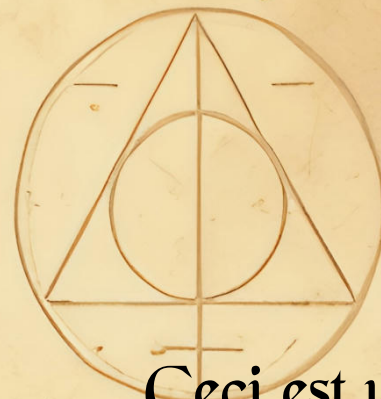
perion cinto
 subits
 rich cithin
 games



out by rountates
 ventur ematones
 ttrora cituorot

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
magnifique
théière

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

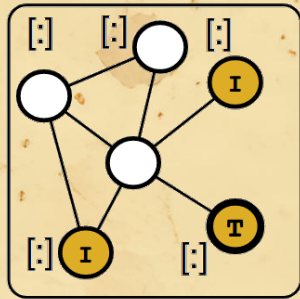
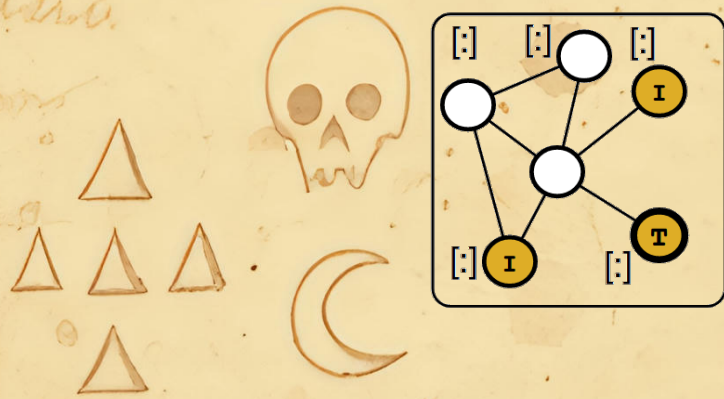
3. Ajouter 16 gouttes de potion de Babillage

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



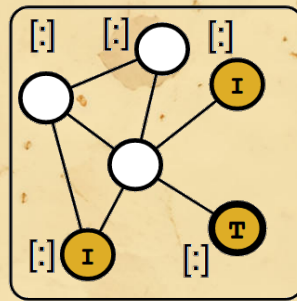
Property of
le magicien quantique

Remuer dans le sens horaire → anti-horaire

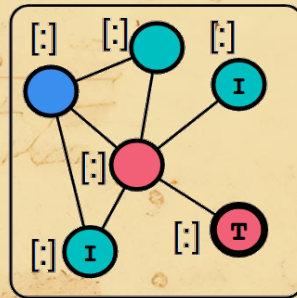


Property of
le magicien quantique

Remuer dans le sens horaire → anti-horaire



↓ Prediction



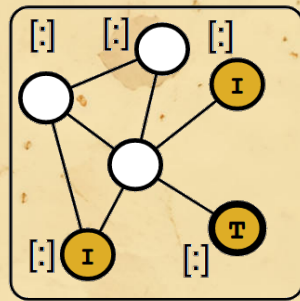
“Class 2”

80.4% confidence

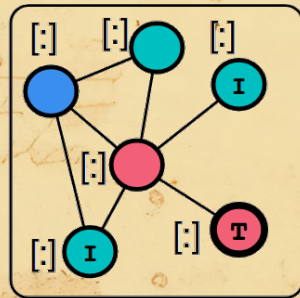


Property of
le magicien quantique

Remuer dans le sens horaire → anti-horaire



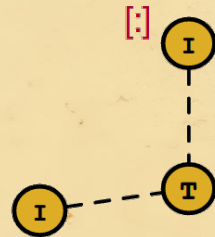
↓ Prediction



“Class 2”

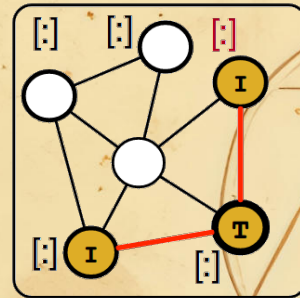
80.4% confidence

+



Perturbations

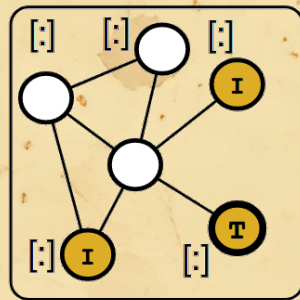
=



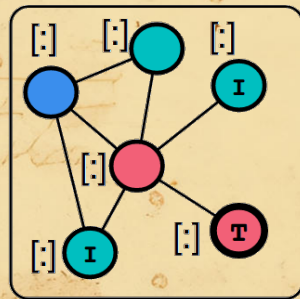
- Target
- Influencer
- Class1
- Class2
- Class3
- Node features

Property of
le magicien quantique

Remuer dans le sens horaire → anti-horaire



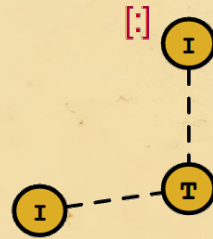
↓ Prediction



“Class 2”

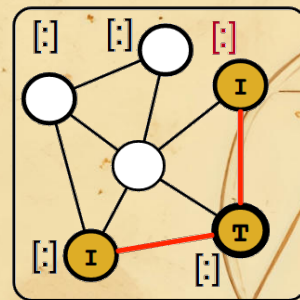
80.4% confidence

+

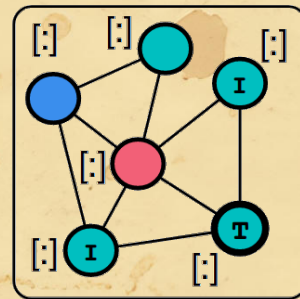


Perturbations

=



↓ Prediction



“Class 3”

92.1% confidence

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

Property of
le magicien quantique

Formule pour cook un challenge du 404

1. Ajouter un chat → une thèière

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

sens anti-horaire

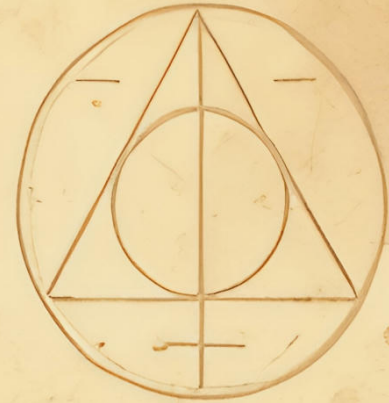
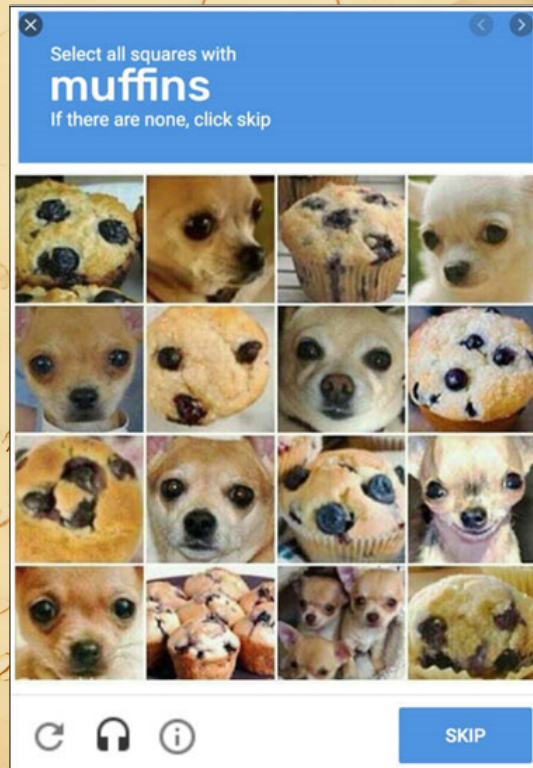
3. Ajouter ~~16~~ gouttes de potion de Babillage

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



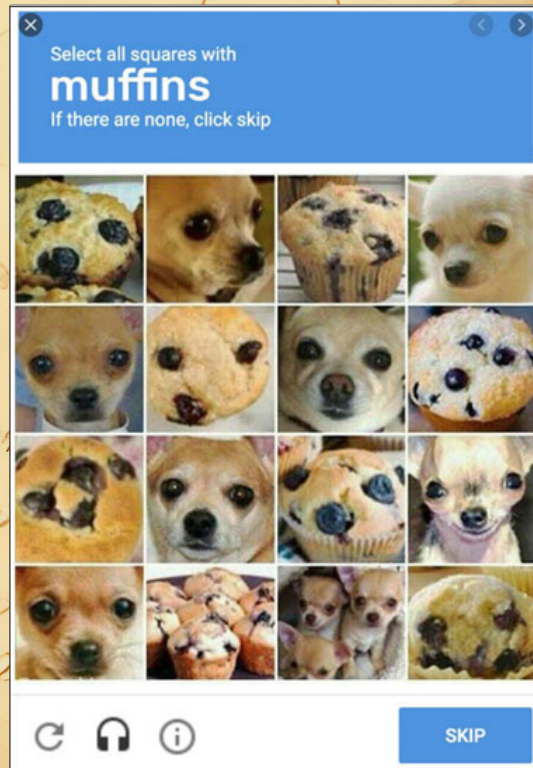
Property of
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



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



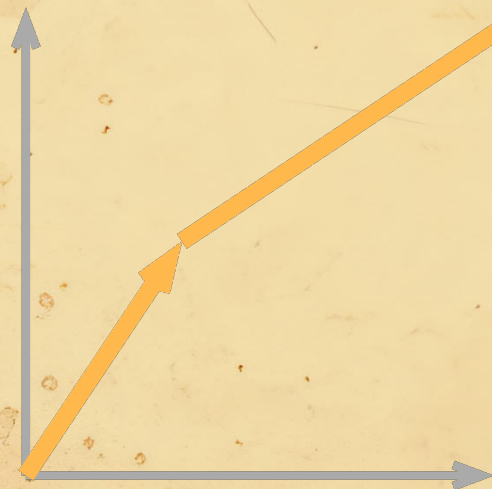
Property of
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



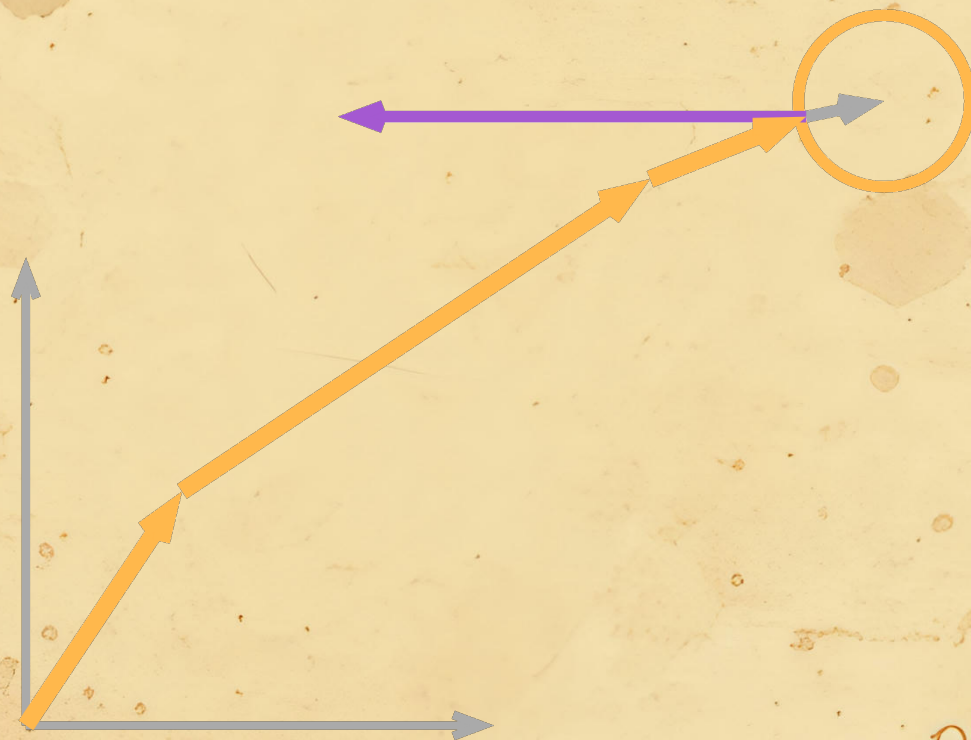
Property of
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



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



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

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

X. *environnements*
aspects contrastés
pot. de dev.



Interpretabilite mecanique

Property of
le magicien quantique

Trifolium pratense



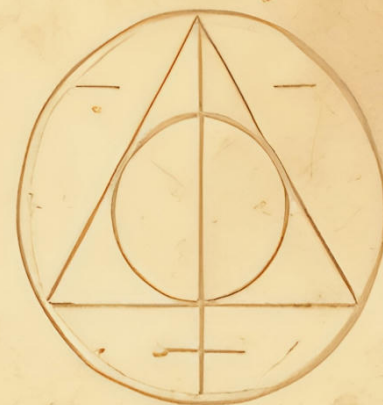
Property of
le magicien quantique



Edroia citreata



Quelle



Property of
le magicien quantique



Quelle est la couleur

Property of
le magicien quantique

Trifolium pratense



1a



Property of
le magicien quantique

Trifolium pratense



1. environnementales
 2. agences contractuelles
 3. potentiel



Property of
le magicien quantique

(X) *esurire tentantes*
agere coartantesque
poterant



Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

(X) *environnementales*
agences coordonnées
pour *des*



perian cinctus
subnate - ~~great~~ ~~the~~

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

rich cr
gumms

red by counteracts,
~~gentle emollients~~
 Trocha Citronis



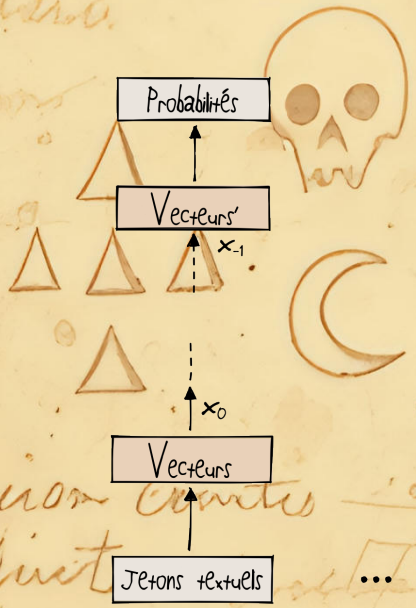
Property of
le magicien quantique

Edroia citreata

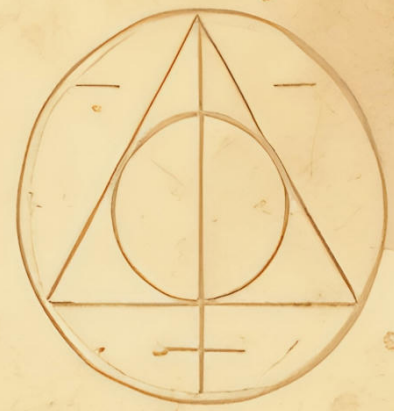
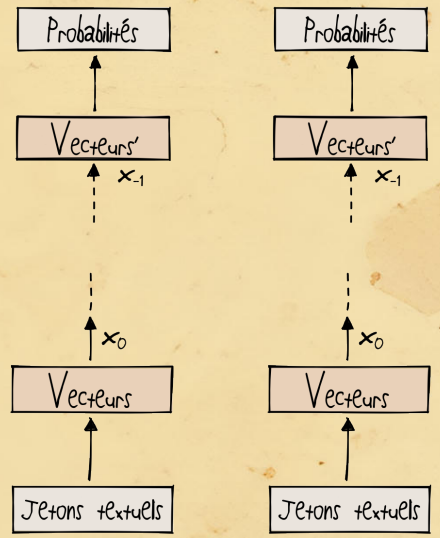


Property of
le magicien quantique

diff. fonction
prob. uniformly
treated and if trained properly



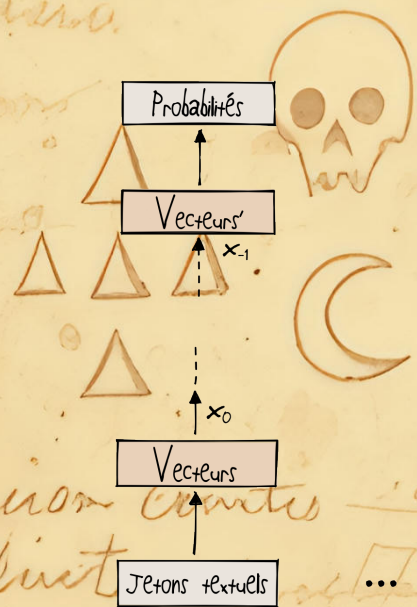
environnements
asynchronous
pots de vote



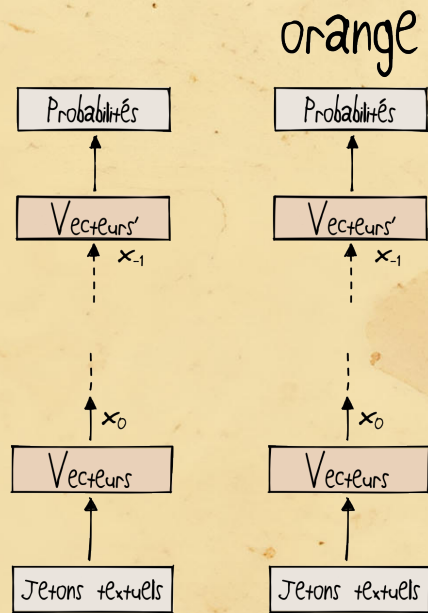
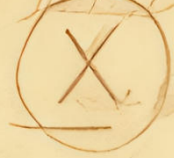
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

diff. fonction
probl. univ. rel.
tandem pnd. of 11 toivnd prapozly



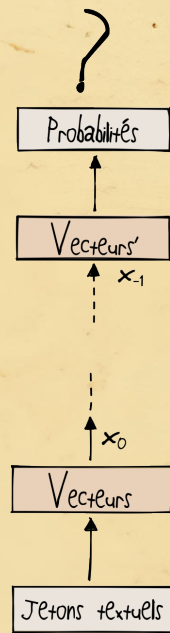
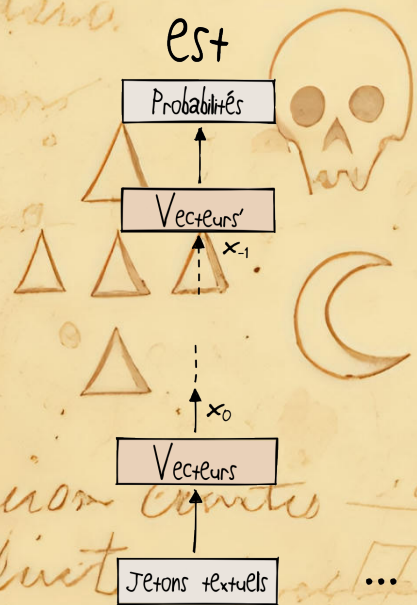
enviire testiles
asmes coratocishine
pstr deovet



Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

diff. fonction
prob. uniformly
treated and if trained properly



orange

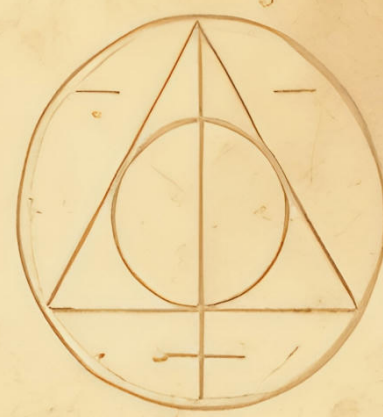
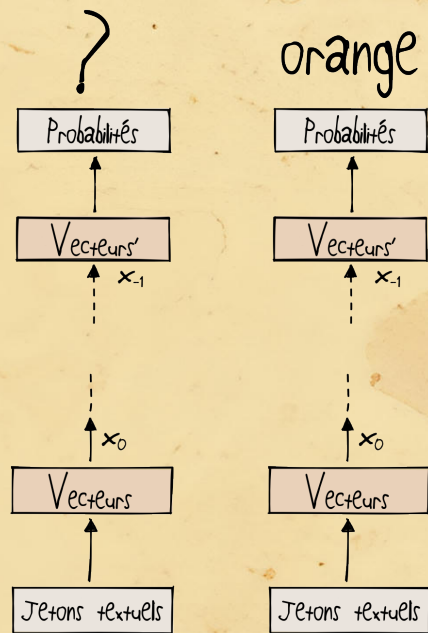
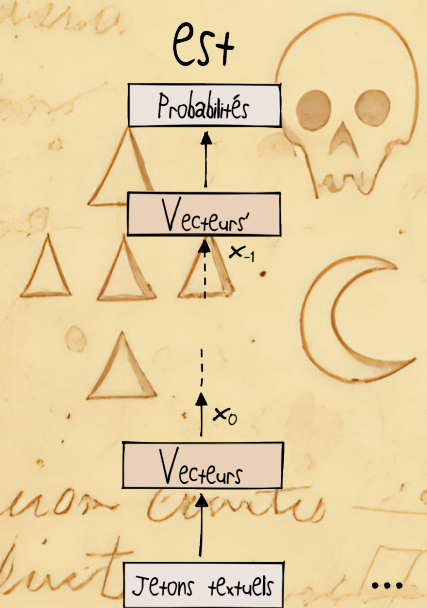


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

diff. fonction
probl. univ. rel.
tenseur prod. of 1/2 tensor properly

environnements
asmes constructives
pots de vent



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

Property of
le magicien quantique

rich cr
garden

red by counteracts,
~~gentle emollients~~
 Trocha Citronis

Property of
le magicien quantique

difficult to understand
 prob. uniformly
 understand prob. if I know properly

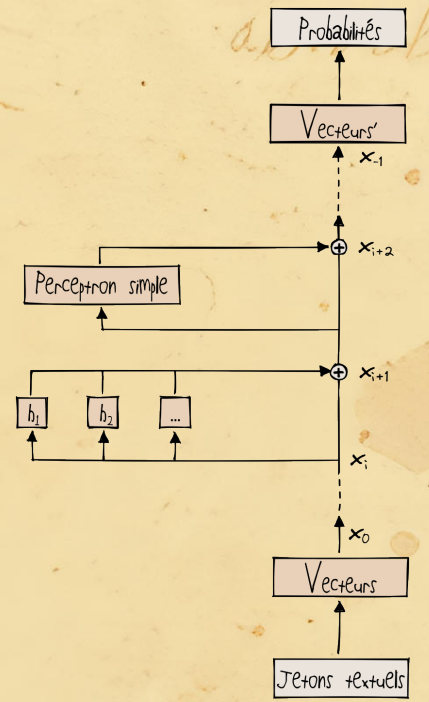


perceptron circuits
 subjects
 rich
 games

Quelle est la couleur du chat de Hermione Granger ?

not by...
 ...
 ...

environnementales
 asines caractéristiques
 pstr devent

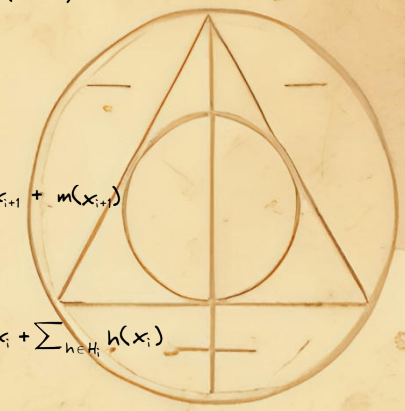


$$T(t) = s(x_{i-1} W_{ij})$$

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

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

$$x_0 = t w_E$$

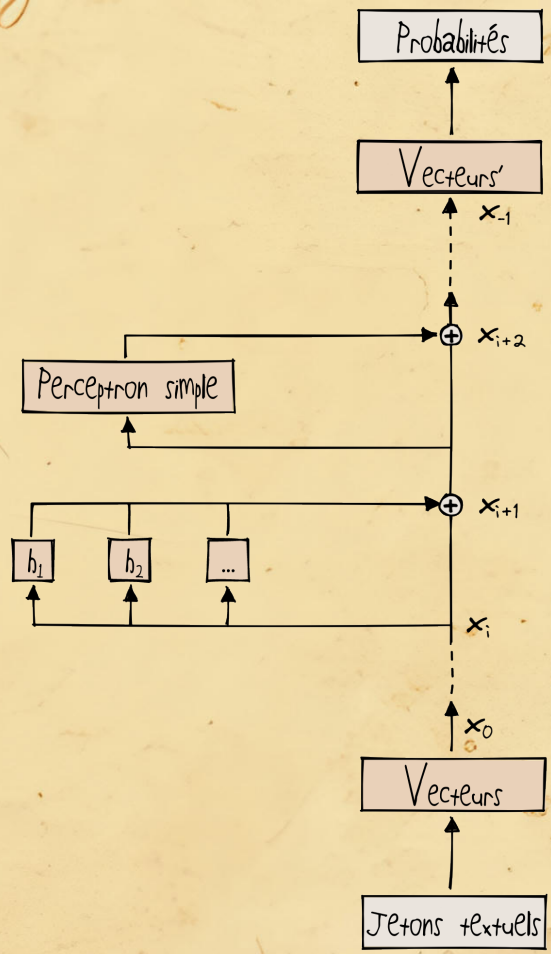


Property of
 le magicien quantique

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

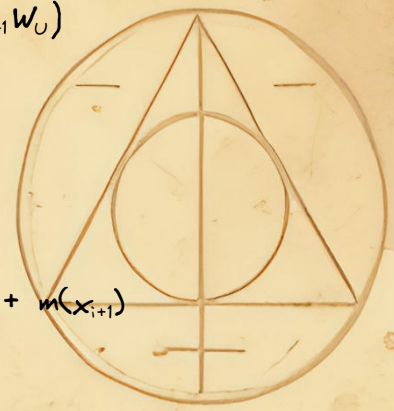


perceptron circuits
 subjects - ~~complex~~
 rich ~~complex~~
 games
 not by ~~complex~~
~~complex~~
~~complex~~



environnementales
 asomes caractéristiques
 pstr devent

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



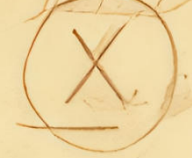
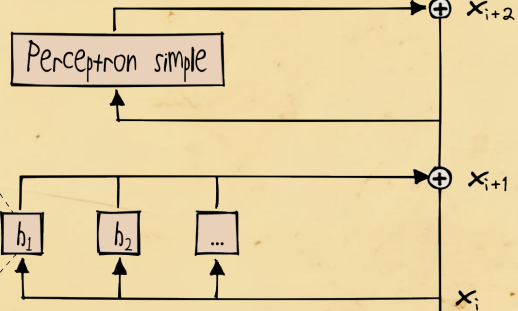
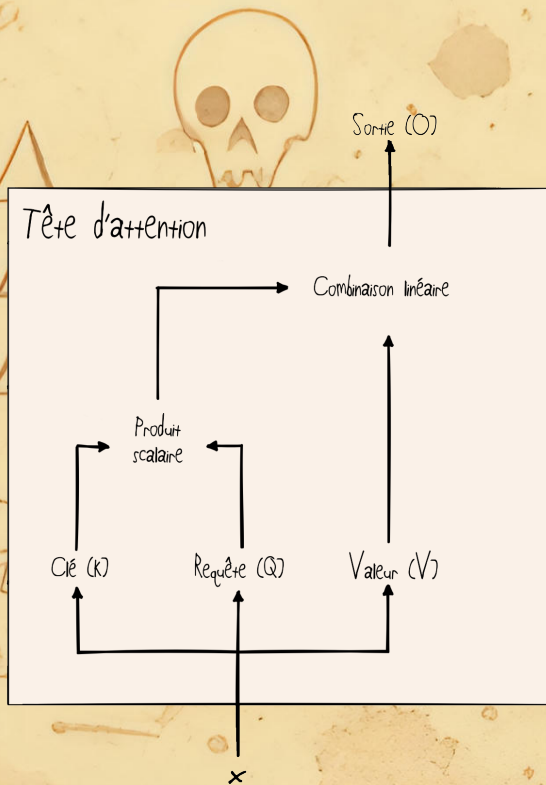
$$x_{i+2} = x_{i+1} + h(x_{i+1})$$

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

$$x_0 = t w_E$$

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

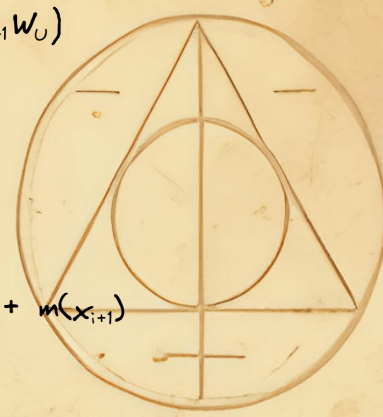
$$\tau(t) = s(x_{-1} w_u)$$

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

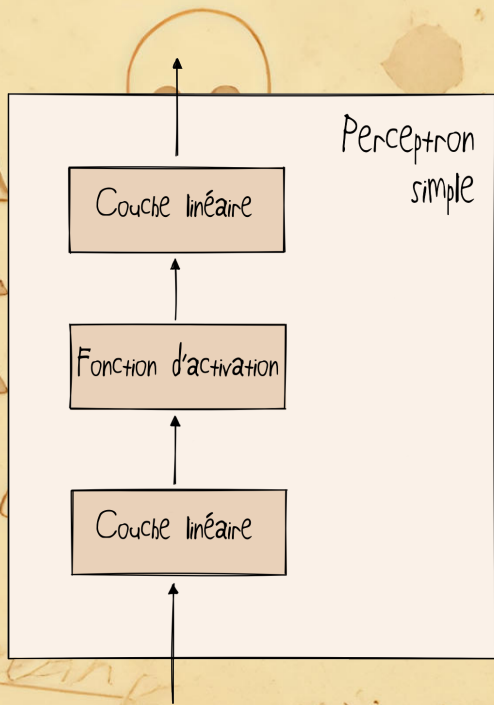
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

$$x_0 = t w_E$$

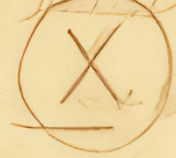
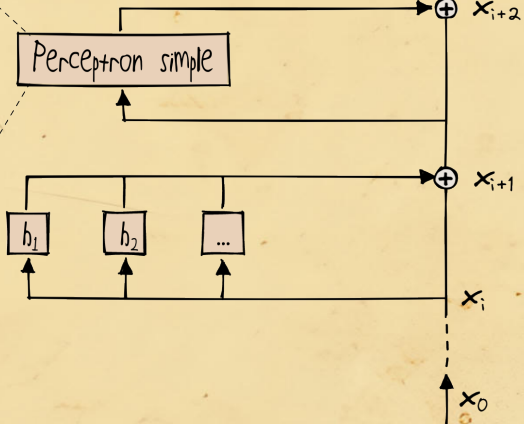
Property of
 le magicien quantique



diff. fonction
 prod. uniform.
 tendent vers 1/2 si bornes proprement

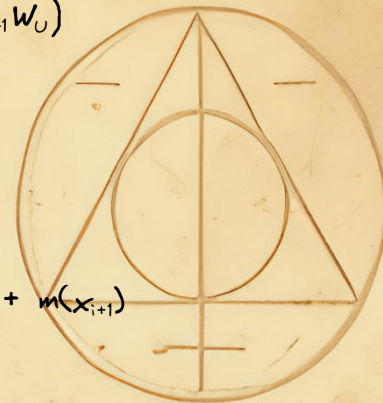


Perceptron simple



environnementales
 asines constructivisme
 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

rich cr
gumms

2.
 not by ~~reverted~~,
~~entirely~~ ~~embodied~~
~~to~~ ~~the~~ ~~citizens~~

Property of
le magicien quantique

(X) *esurire tentantes*
agere coartantesque
poterant



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

Property of
le magicien quantique

pool carefully.
-tubercle part of it is not properly.



Clé :
Je suis un
pronom
interrogatif

Clé :
Je suis un
adjectif

...

Quelle est la couleur du chat de Hermione Granger ?

1. environnementales
 2. agences contractuelles
 3. potentiel

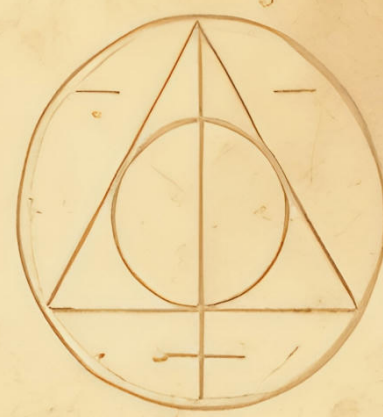
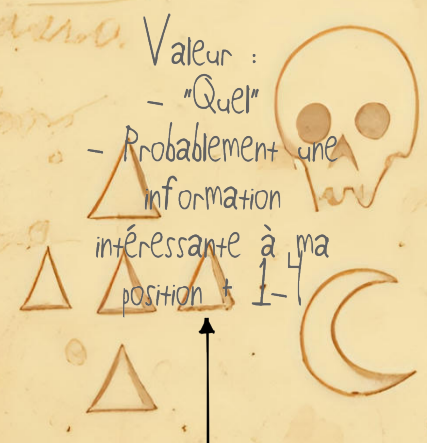


Requête :
Y a-t-il un marqueur de question ?

Property of
le magicien quantique

...
 peut éventuellement
 intervenir pour qu'il trouve progressivement

...
 envoi de test
 asser contrainte
 pour de voir



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 it is not properly

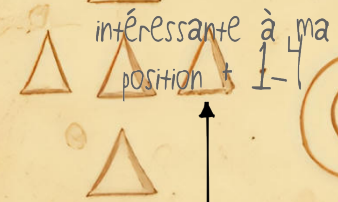
envoie les
à la cour
pour les
pour les

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

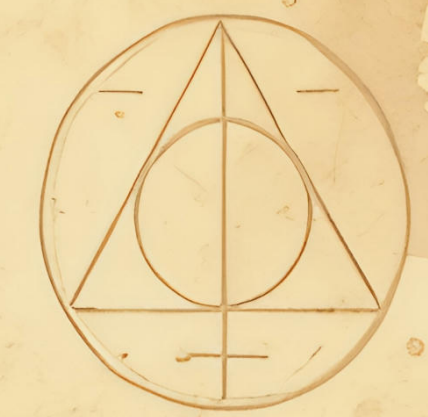
Clé :
Je suis un
adjectif

...

...



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



Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

rich cr
gumms

red by contactus,
~~Anterior emulsion~~
 Extract Citrus



Property of
le magicien quantique

at the

1840

~~Yartus Corbentico~~

Ischnura citreata

Requête :
Informations
sur moi ?

Quelle est la couleur du chat



~~X~~ envoie tes lettres
à mes amies coquette et à
moi pour te le dire.



Property of
le magicien quantique

difficult to understand
properly, but if we have properly
understand



Clé :

Je suis un
pronom
interrogatif

Clé :

Je suis un
attribut

Requête :

Informations
sur moi ?

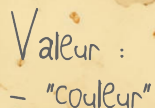
Quelle est la couleur du chat

envoie les
à la cour
pour le
service

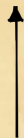


Property of
le magicien quantique

(X) *environnementales*
agences coordonnées
pour *des*



Valeur :
- "couleur"



Canal mis à jour :

- Je suis un chat
- J'ai une couleur

Clé :
Je suis un
pronom
interrogatif

Clé :
Je suis un
attribut

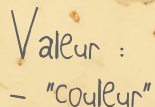
Requête :
Informations
sur moi ?

Quelle est la couleur du chat



Property of
le magicien quantique

(X) *environnementales*
agences coordonnées
pour le développement



- Je suis un chat
- J'ai une couleur

→ Jeton suivant ?



Clé :
Je suis un
attribut

Requête :
Informations
sur moi ?

Quelle est la couleur du chat

Property of
le magicien quantique

difficult to understand
properly, but if I have properly



Valeur :
- "couleur"

→

↑

Canal mis à jour :
- Je suis un chat
- J'ai une couleur

Clé :
Je suis un ...
pronom
interrogatif

Clé :
Je suis un
attribut

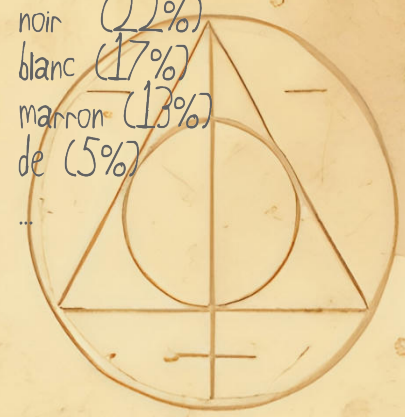
Requête :
Informations
sur moi ?

Quelle est la couleur du chat

envoie les
à mes contacts
pour leur
service

→ Jeton suivant ? →

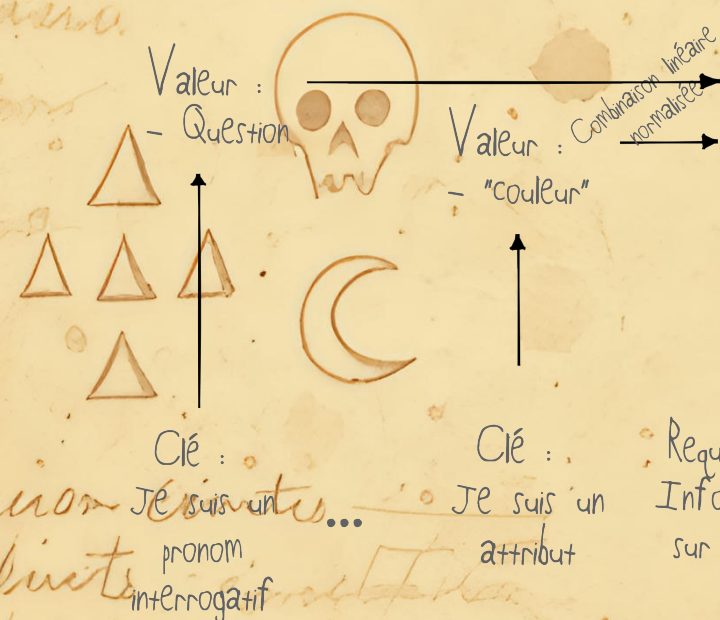
- noir (22%)
- blanc (17%)
- marron (13%)
- de (5%)
- ...



Property of
le magicien quantique

difficult to understand
 but carefully
 understand and if having properly

environnements
 as a counterbalance
 to the power

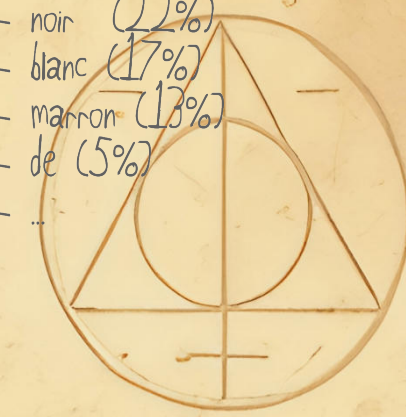


Canal mis à jour :

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

Jeton suivant ?

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

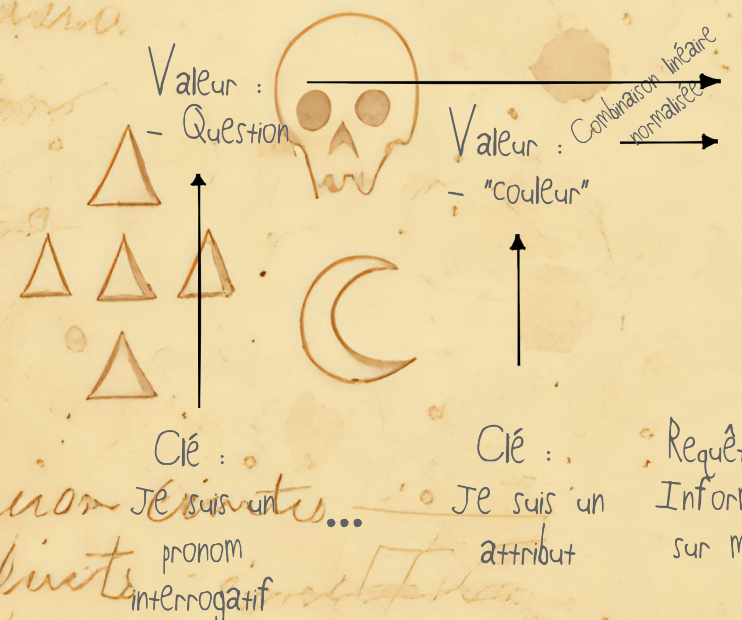


Quelle est la couleur du chat

Property of
 le magicien quantique

difficult to understand
 but carefully
 understand and if having properly

environnements
 as a counterbalance
 to the power

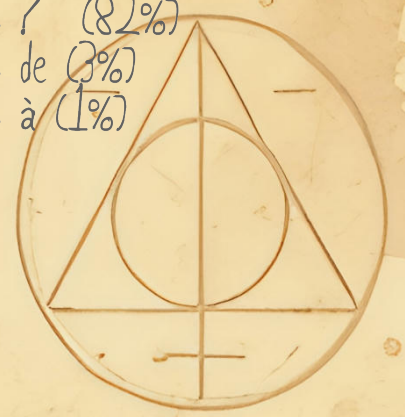


Canal mis à jour :

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

→ Jeton suivant ?

- ? (82%)
- de (3%)
- à (1%)
- ...

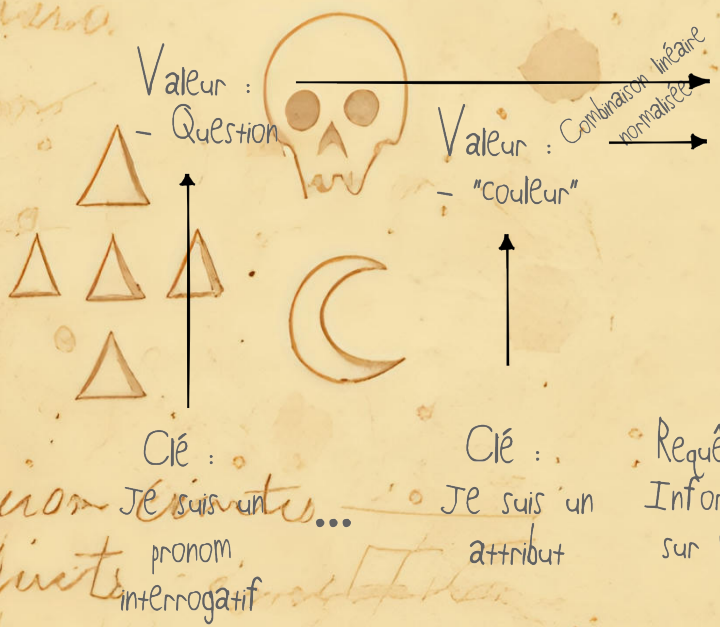


Quelle est la couleur du chat

Property of
 le magicien quantique

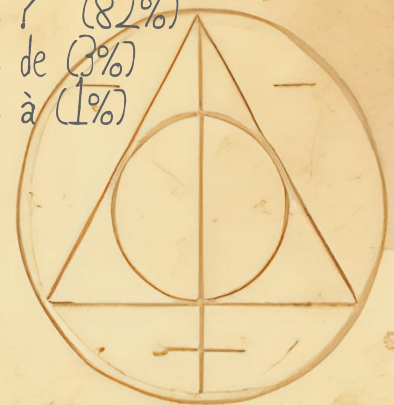
difficult to understand
 but carefully
 understand and if having properly

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

Edroia citreata



Property of
le magicien quantique



Volucra citreata



Hermione Granger

(X) *envisage les*
agences co-ordinates
pour le



Property of
le magicien quantique

(X) *esurire tentantes*
agere coartantesque
poterant



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

Hermione Granger

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

Hermione Granger



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

rich cr
gumms

red by counteracts,
~~gentle emollients~~
 Trocha Citronis

Property of
le magicien quantique

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



person cointes
subjects
rich cr
gurus

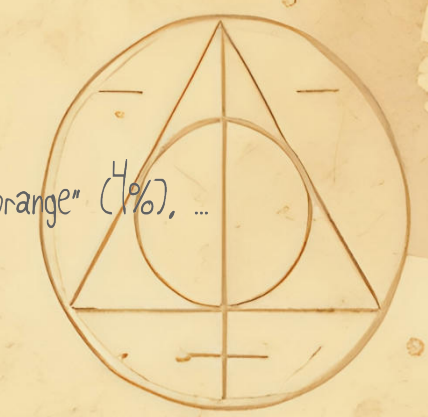
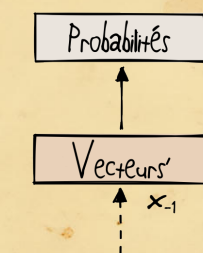
Quelle est la couleur du chat de Hermione Granger ?

red by contact
entire emotional
treatment

environnements
aspects contrastés
servir



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



Property of
le magicien quantique

1. environnementales
 2. agences coratocisshine
 3. potr deoet



brion cointe <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/>)