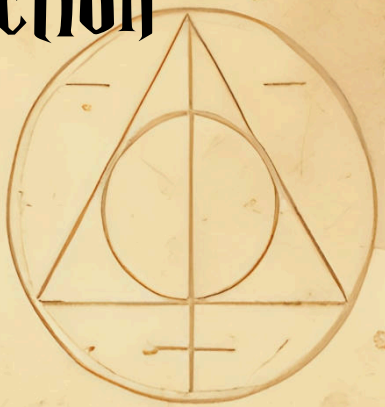
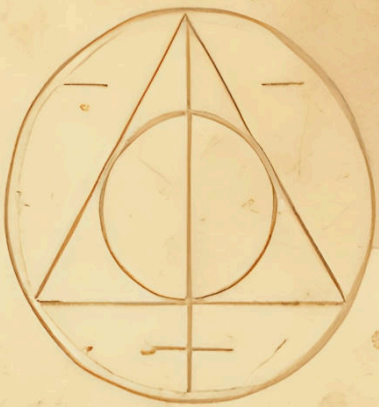
Attaques d'IA – Livre d'introduction



De Sckathapscal Gorphineus Quantifilius Artificewick des Vents

Formule pour cook un challenge du 404

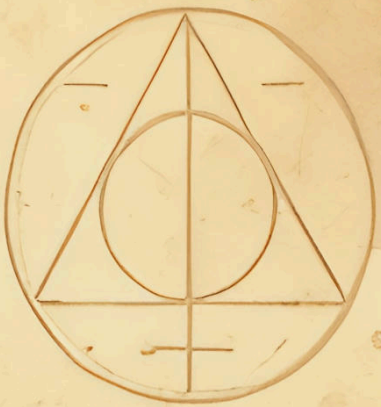
1. Ajouter un chat
2. Remuer 3 fois dans le sens horaire
3. Ajouter 16 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout



Property of
le magicien quantique

Formule pour cook un challenge du 404

1. Ajouter un ~~chat~~ → une théière
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

→

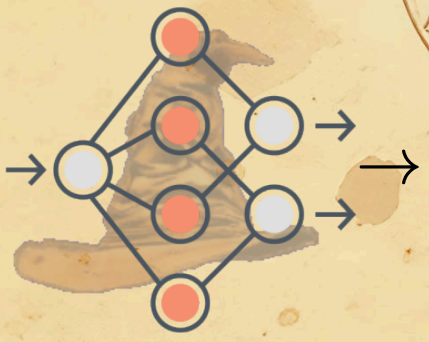
→

Ceci est un chat

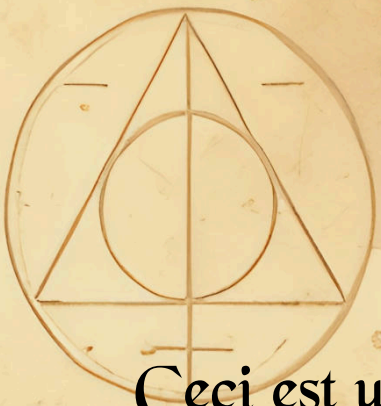
Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→



→

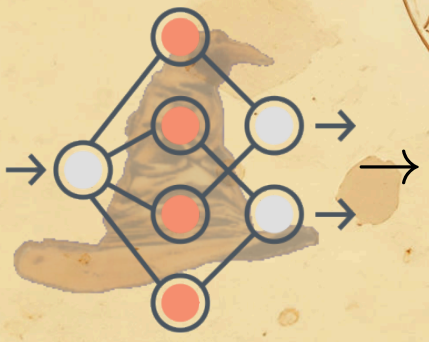


Ceci est un chat

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

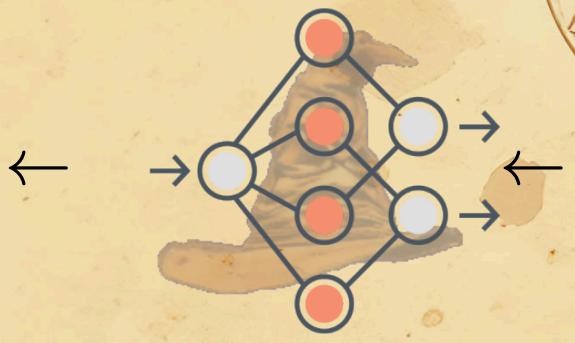
→



→

Ceci est une théière

Ajouter ~~un chat~~ → une théière



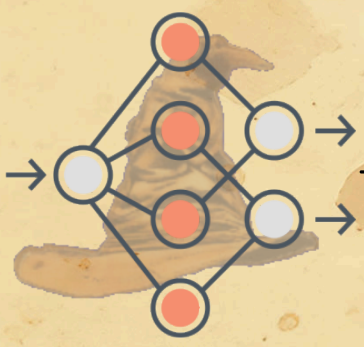
~~Ceci est un chat~~

Ceci est une théière

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→

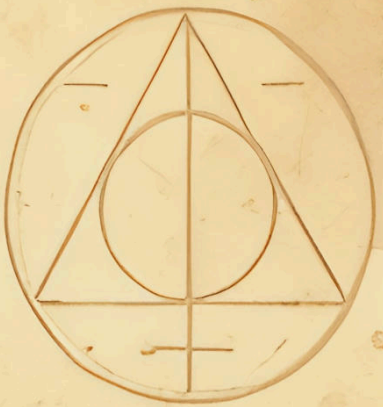


→

Ceci est clairement une théière

Property of
le magicien quantique

Property of
le magicien quantique



Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→

→

Ceci est un chat

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→

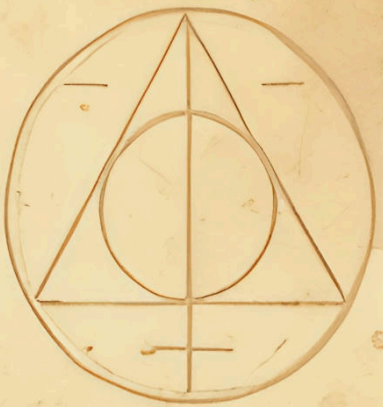
→

Ceci est une magnifique théière

Property of
le magicien quantique

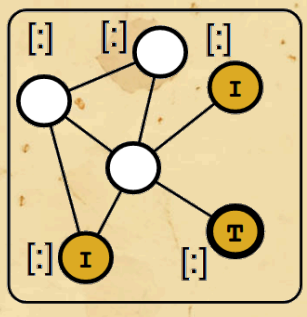
Formule pour cook un challenge du 404

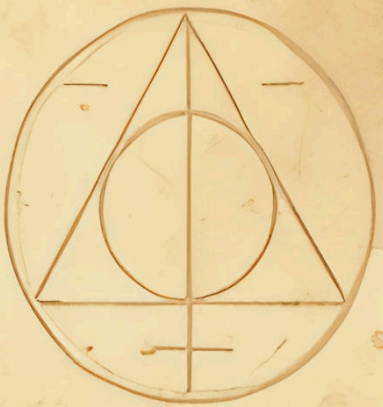
1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le ~~sens horaire~~ → sens anti-horaire
3. Ajouter 16 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout



Property of
le magicien quantique

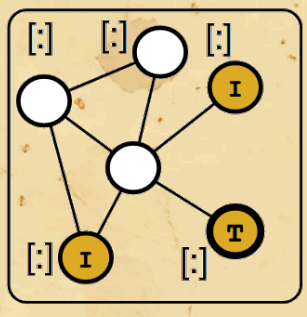
Remuer dans le sens ~~horaire~~ → anti-horaire



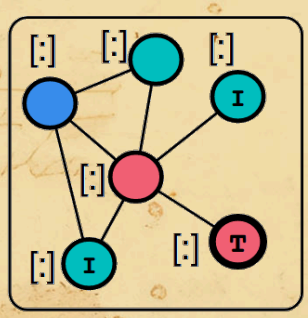


Property of
le magicien quantique

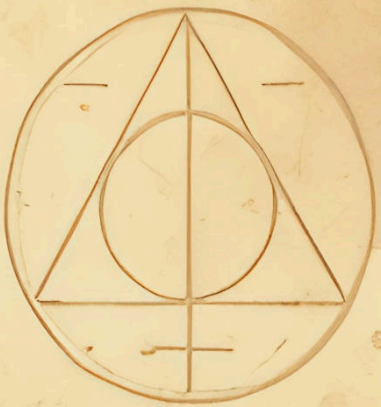
Remuer dans le sens ~~horaire~~ → anti-horaire



Prediction

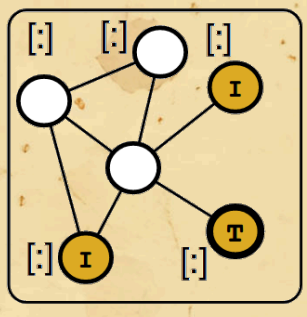


"Class 2"
80.4% confidence

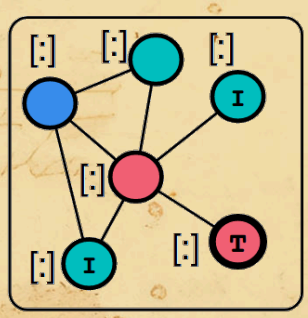


Property of
le magicien quantique

Remuer dans le sens ~~horaire~~ → anti-horaire

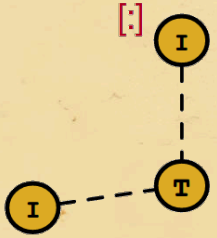


Prediction



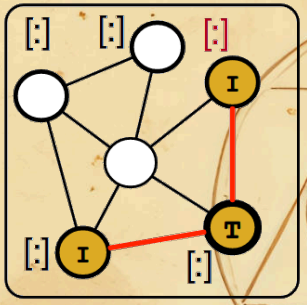
"Class 2"
80.4% confidence

+



Perturbations

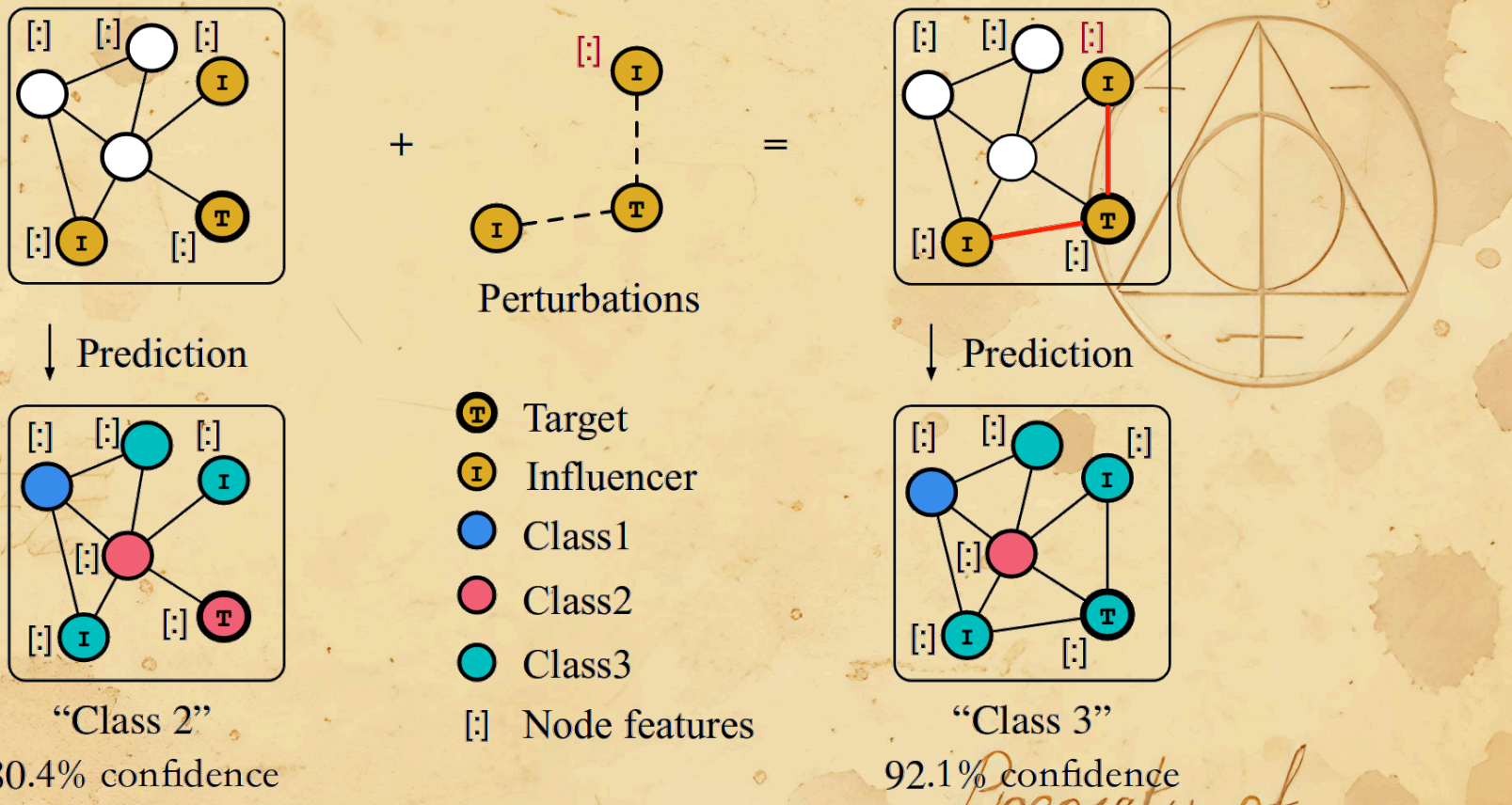
=



- T Target
- I Influencer
- Class1
- Class2
- Class3
- [:] Node features

Property of
le magicien quantique

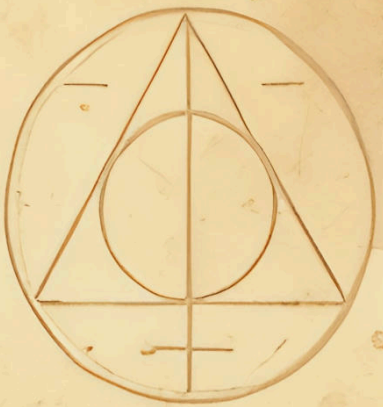
Remuer dans le sens ~~horaire~~ → anti-horaire



Property of
le magicien quantique

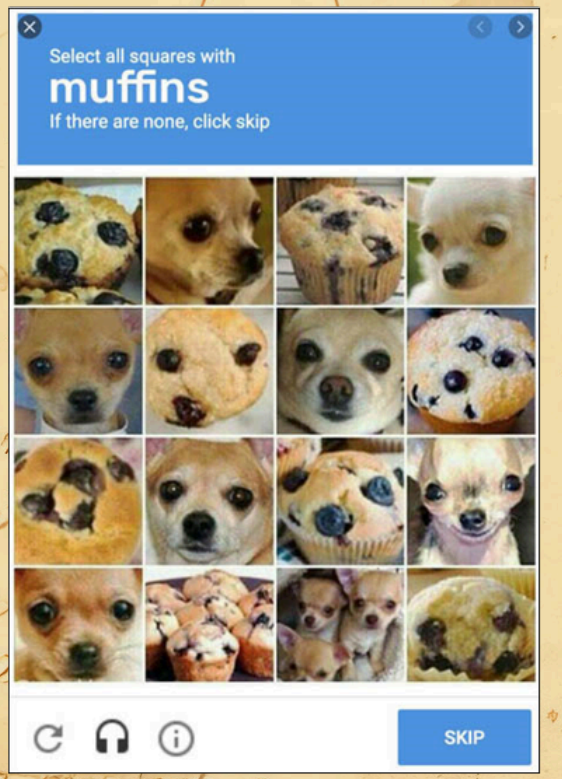
Formule pour cook un challenge du 404

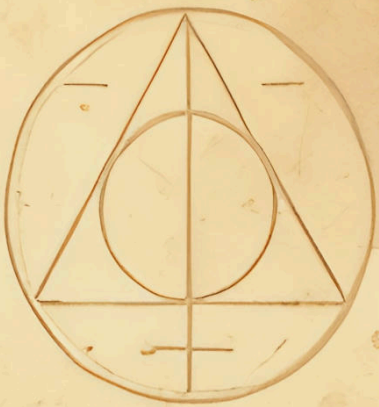
1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le ~~sens horaire~~ → sens anti-horaire
3. Ajouter ~~16~~ → 17 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout



Property of
le magicien quantique

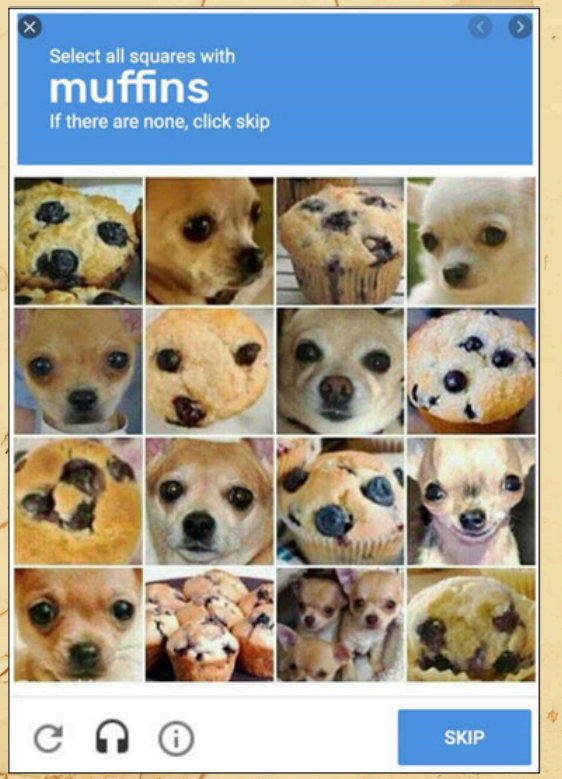
Ajouter ~~16~~ → 17 gouttes de potion de Babillage





Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage



↓

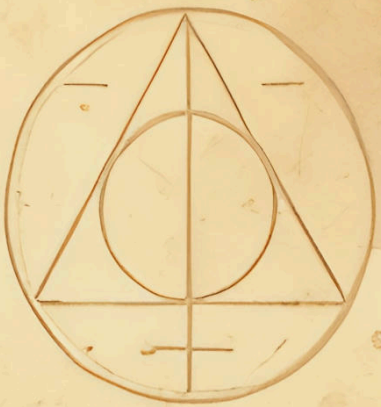
4

↓

5

Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage

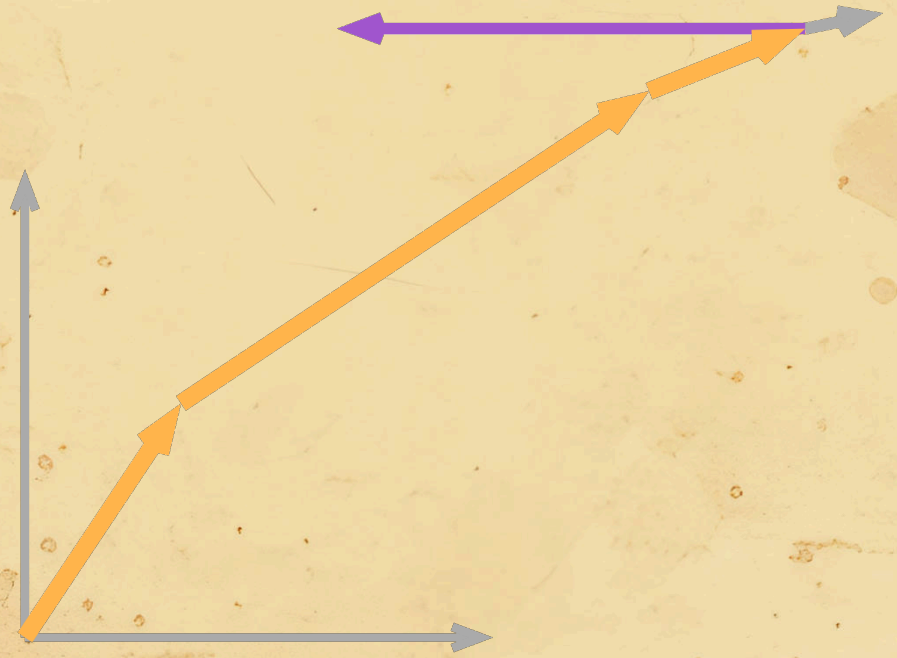


Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage

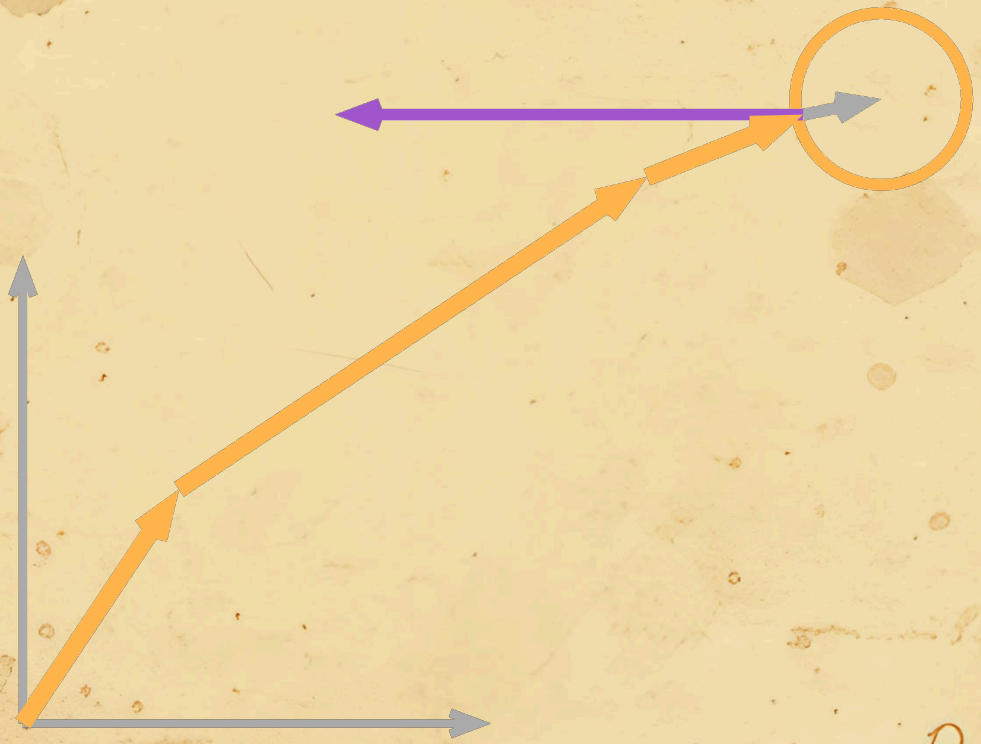
Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage



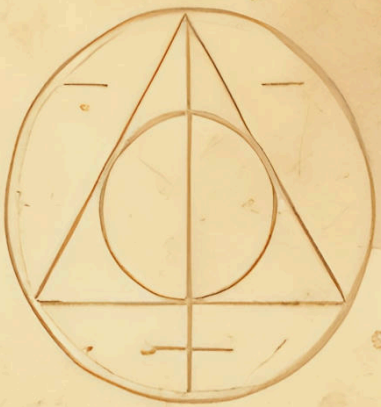
Property of
le magicien quantique

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



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

→ Manipulation d'algorithmes et instrumentalisation d'influenceurs

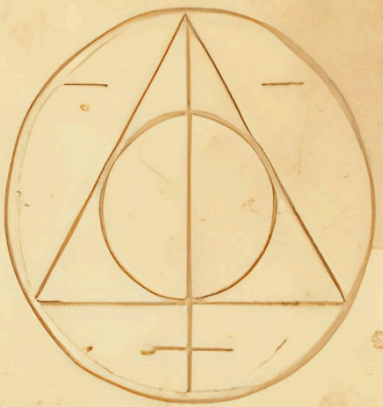
→ Défis et opportunités de l'intelligence artificielle dans la lutte contre les manipulations de l'information

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

Formule pour cook un challenge du 404

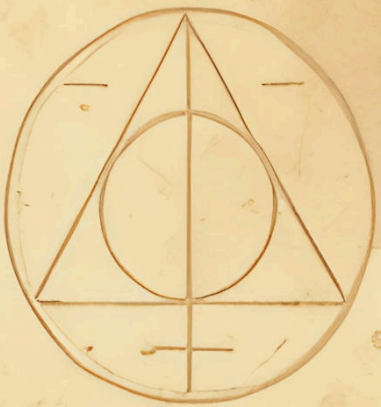
1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le ~~sens horaire~~ → sens anti-horaire
3. Ajouter ~~16~~ → 17 gouttes de potion de Babillage
4. ~~Couper~~ 2 têtes d'hydre, et mélanger le tout

Détourner l'attention (elles repoussent sinon)

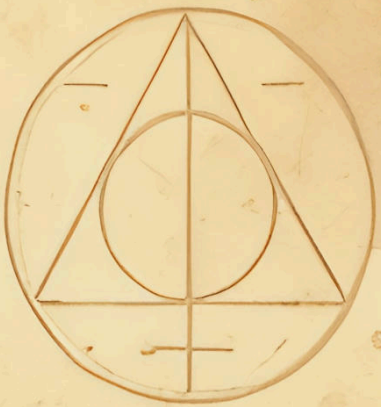


Property of
le magicien quantique

Interpretabilite mecanique

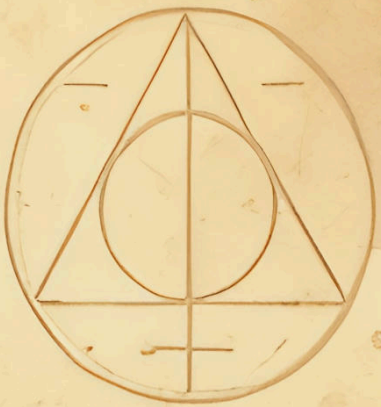


Property of
le magicien quantique



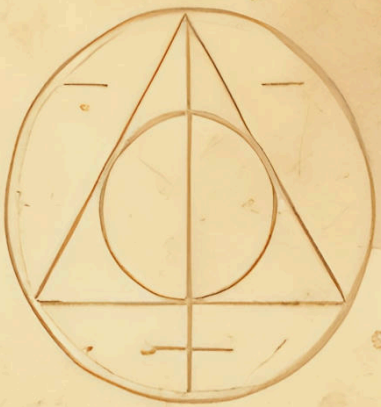
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



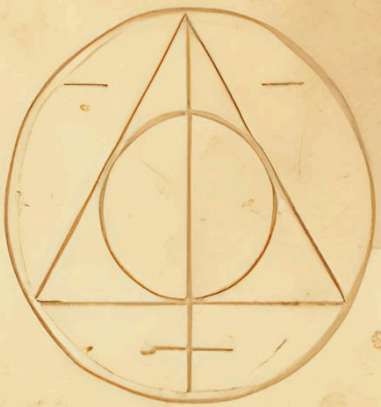
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



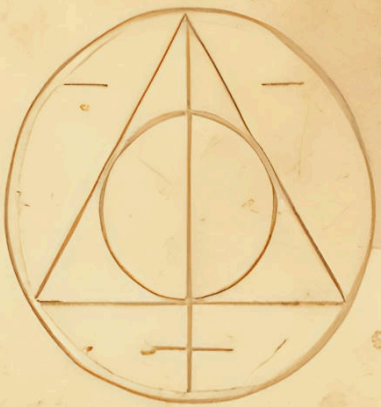
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



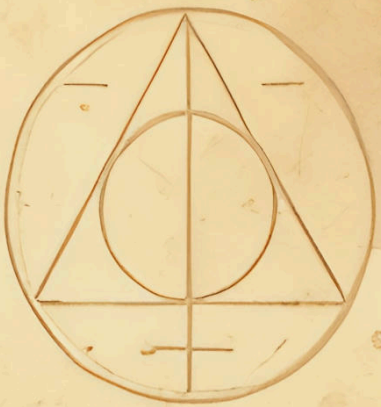
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



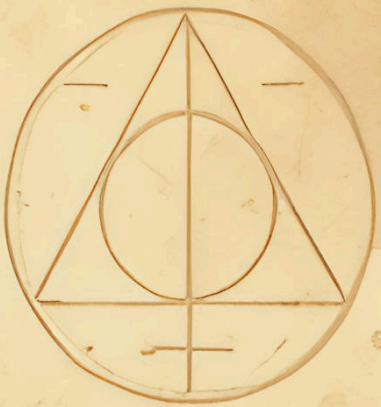
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



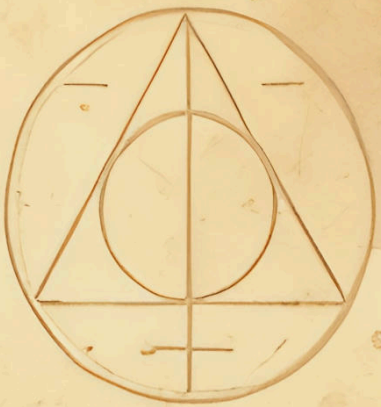
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



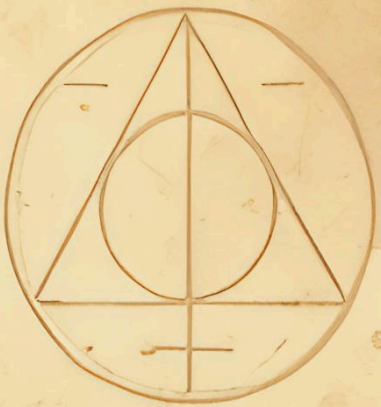
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



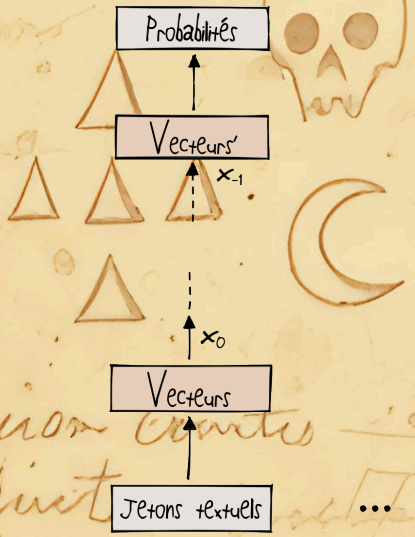
Quelle est la couleur du chat de Hermione Granger ?

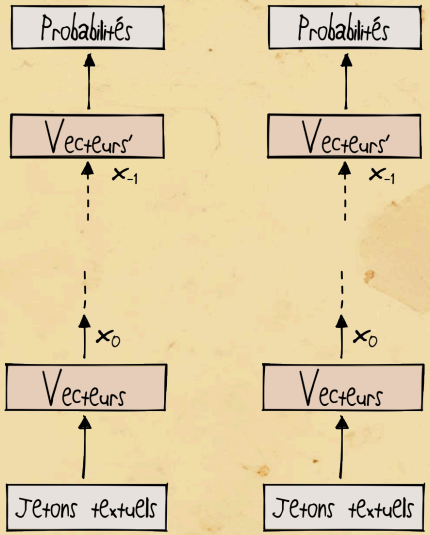
Property of
le magicien quantique

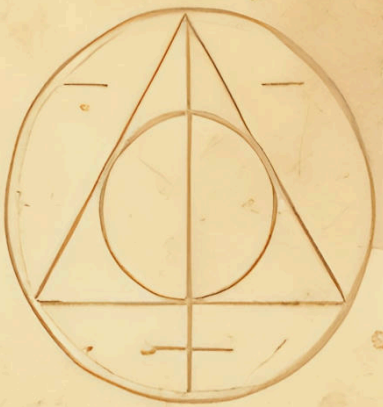


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

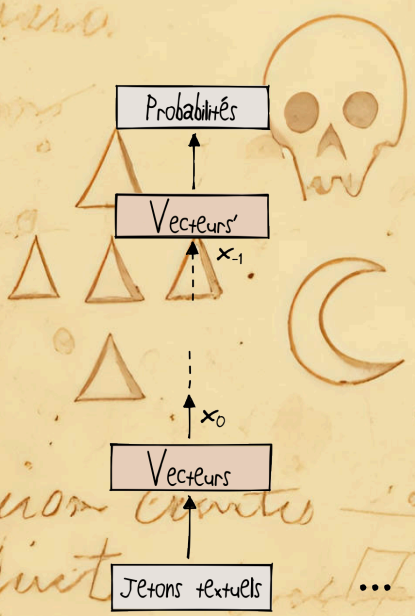


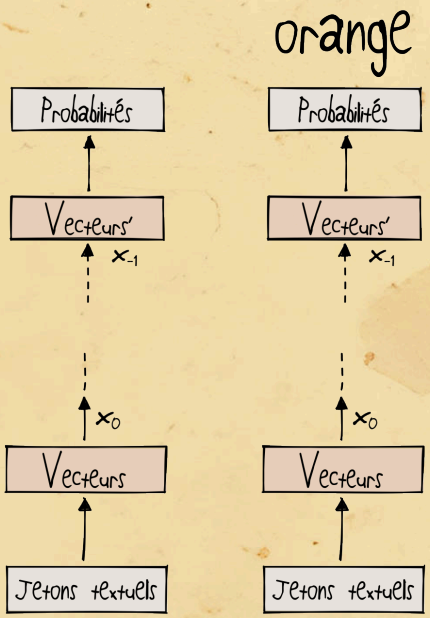


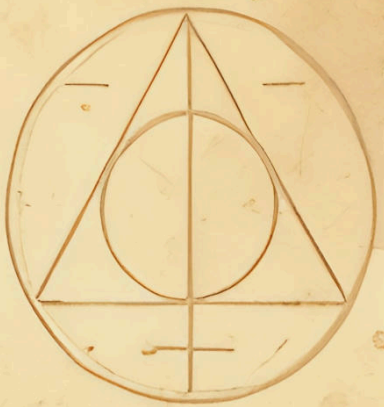


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

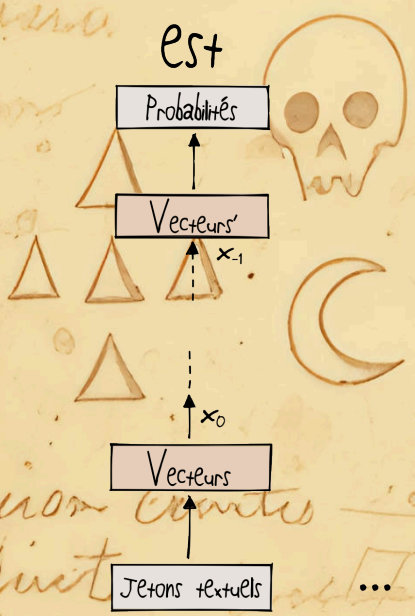


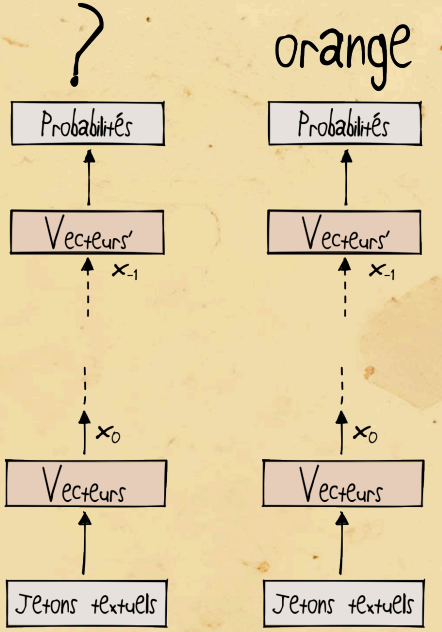


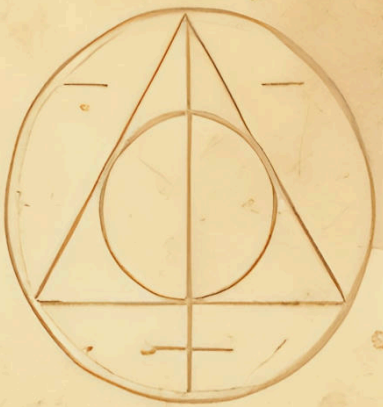


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

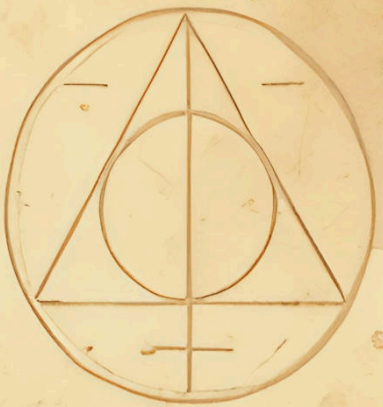




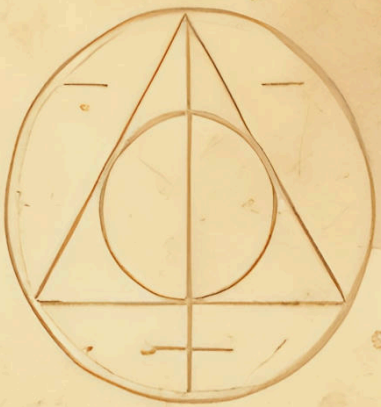


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

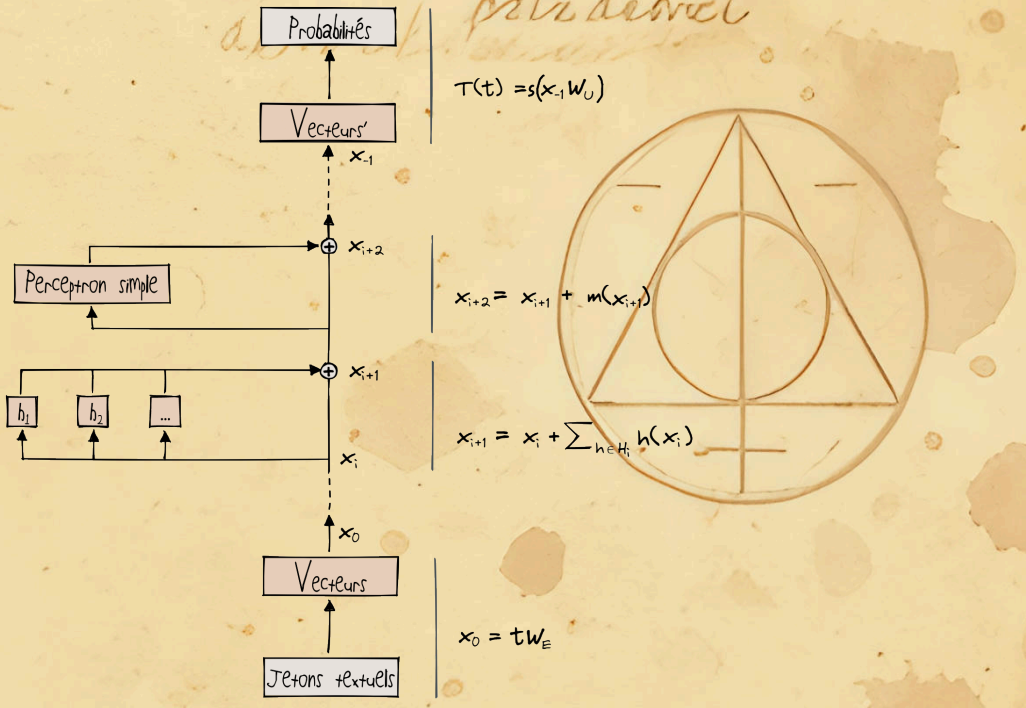


Property of
le magicien quantique



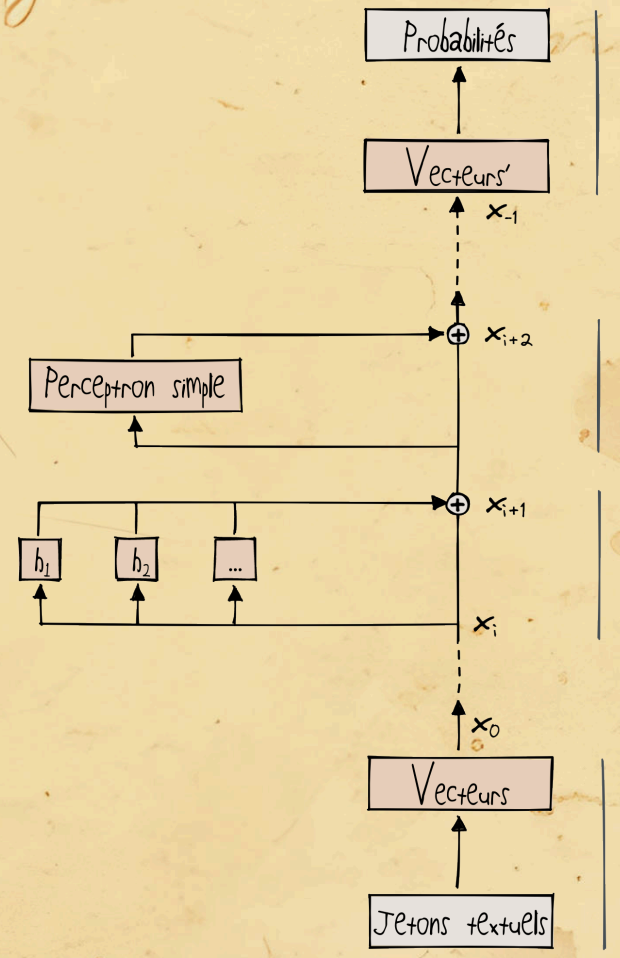
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



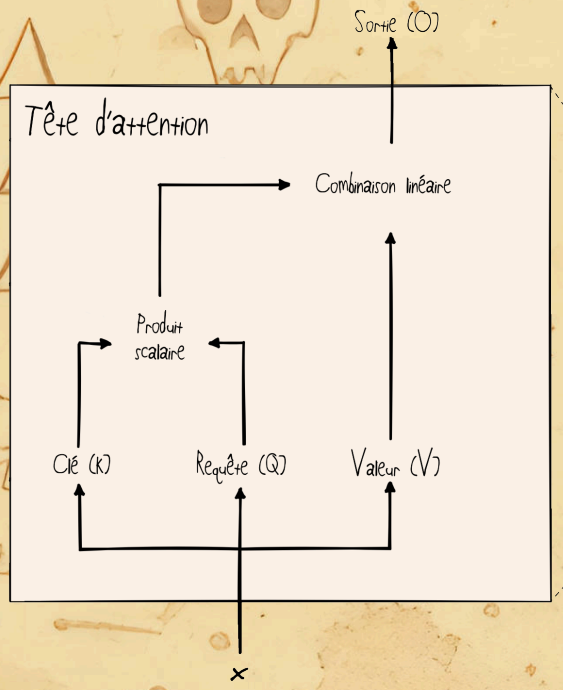
$T(t) = s(x_{-1} W_U)$

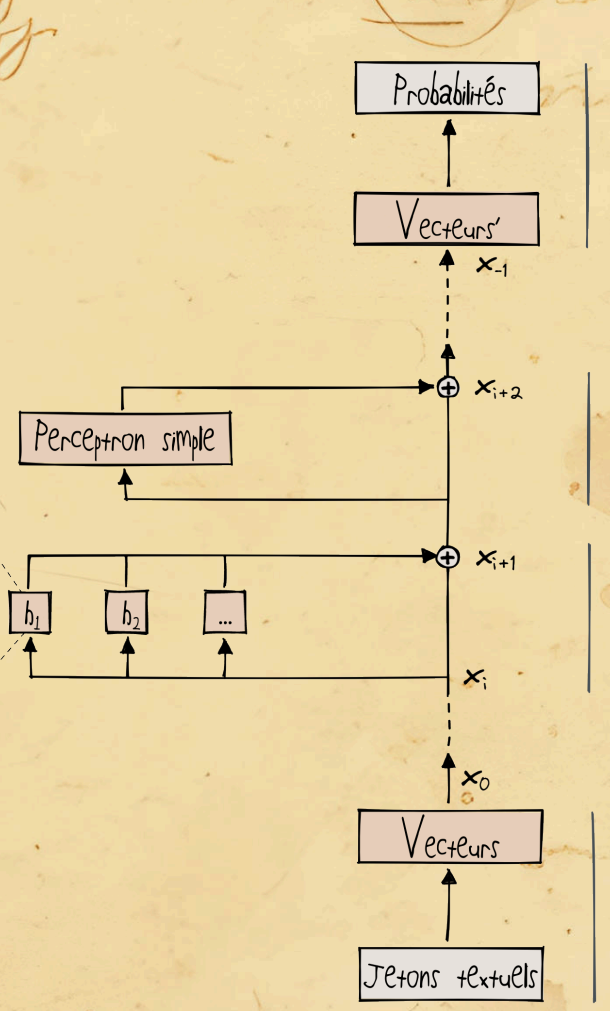
$x_{i+2} = x_{i+1} + m(x_{i+1})$

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

$x_0 = t W_E$

Property of
le magicien quantique





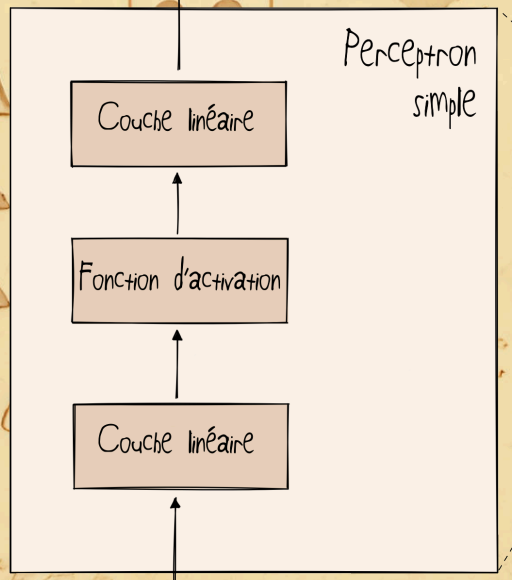
$$T(t) = s(x_{-1} W_U)$$

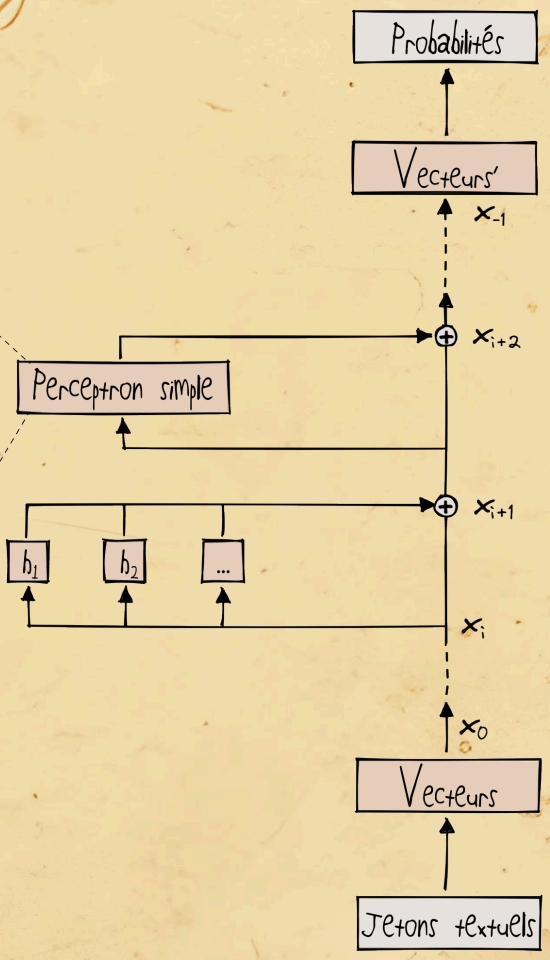
$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

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

$$x_0 = tW_E$$

Property of
le magicien quantique





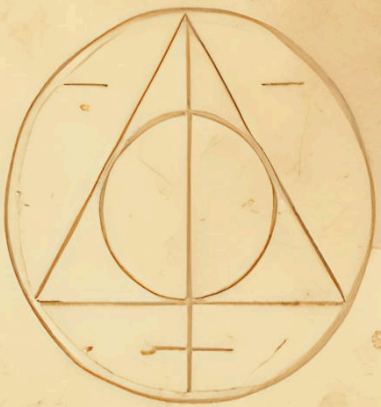
$T(t) = s(x_{-1} W_U)$

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

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

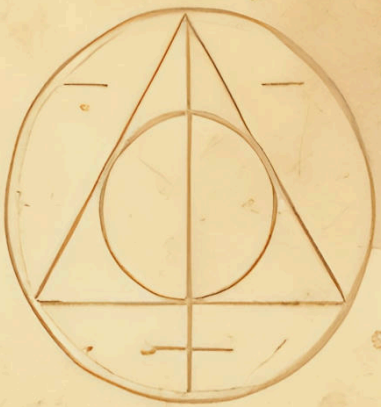
$x_0 = t W_E$

Property of
le magicien quantique



Quelle est la couleur du chat de Hermione Granger ?

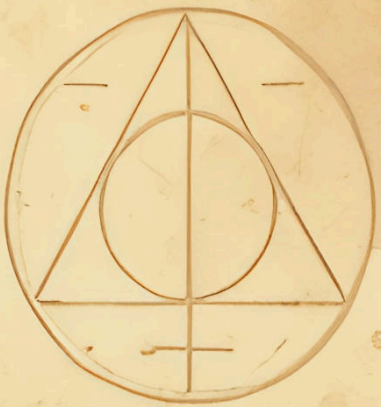
Property of
le magicien quantique



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

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

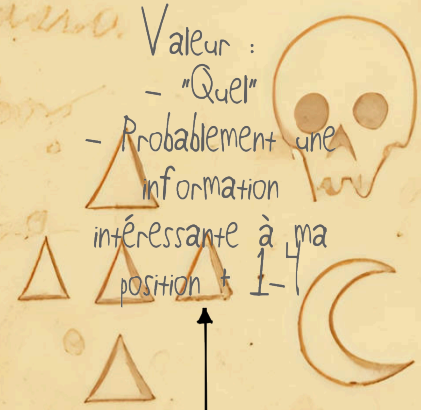


Clé : Je suis un pronom interrogatif ... Clé : Je suis un adjectif ...

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

Quelle	est	la	couleur	du	chat	de	Hermione Granger	?

Property of le magicien quantique



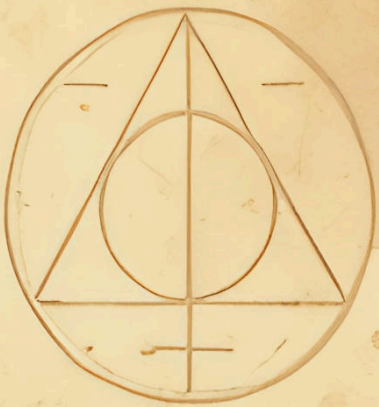
Valeur :
- "Quel"
- Probablement une information intéressante à ma position

Clé :
Je suis un pronom interrogatif

...

Clé :
Je suis un adjectif

...



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

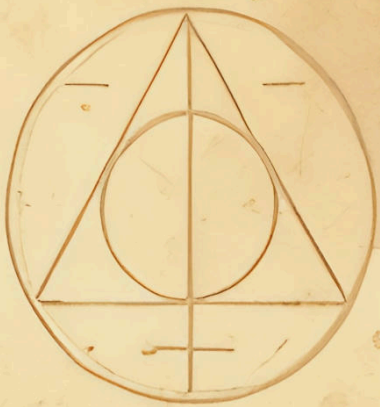
Quelle	est	la	couleur	du	chat	de	Hermione Granger	?

Property of
le magicien quantique

Valeur :
- "Quel"
- Probablement une information intéressante à ma position + 1-4

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

?



Clé :
Je suis un pronom interrogatif

...

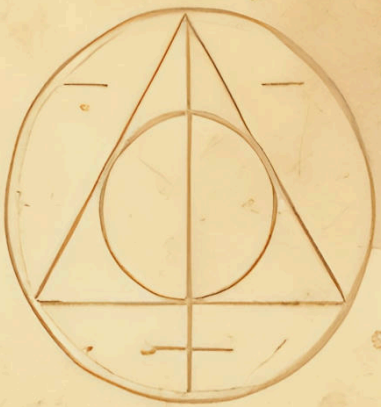
Clé :
Je suis un adjectif

...

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

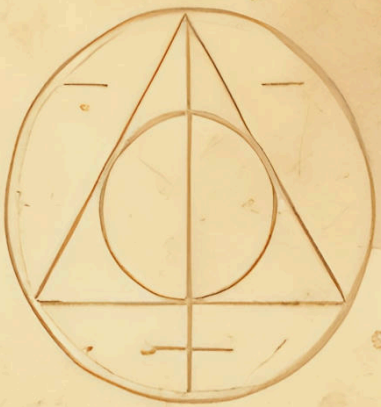
Quelle	est	la	couleur	du	chat	de	Hermione Granger	?

Property of
le magicien quantique



Quelle est la couleur du chat de Hermione Granger ?

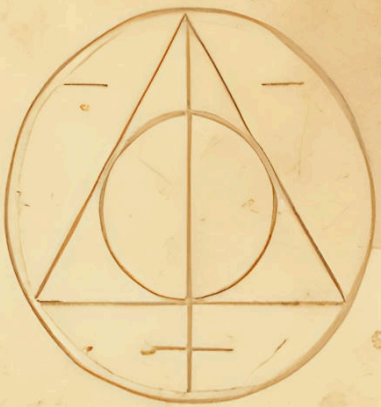
Property of
le magicien quantique



Requête :
Informations
sur moi ?

Quelle est la couleur du chat [redacted]

Property of
le magicien quantique



Clé : Je suis un... pronom interrogatif

Clé : Je suis un attribut

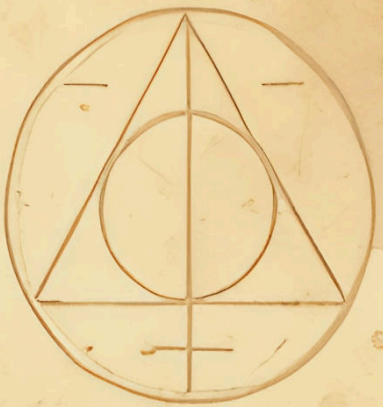
Requête : Informations sur moi ?

Quelle	est	la	couleur	du	chat	

Property of
le magicien quantique

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 ?

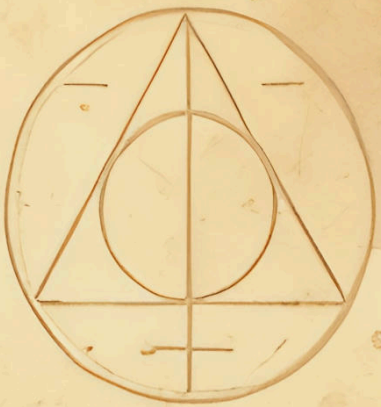
Quelle	est	la	couleur	du	chat	

Property of
le magicien quantique

Valeur :
- "couleur"

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

→ Jeton suivant ?



Clé :
Je suis un...
pronom interrogatif

Clé :
Je suis un attribut

Requête :
Informations sur moi ?

Quelle	est	la	couleur	du	chat	

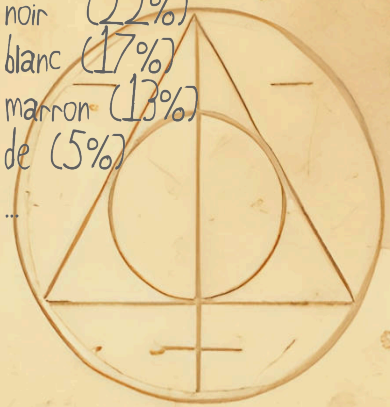
Property of
le magicien quantique

Valeur :
- "couleur"

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

Jeton suivant ?

- noir (72%)
- blanc (17%)
- marron (1%)
- de (5%)
- ...



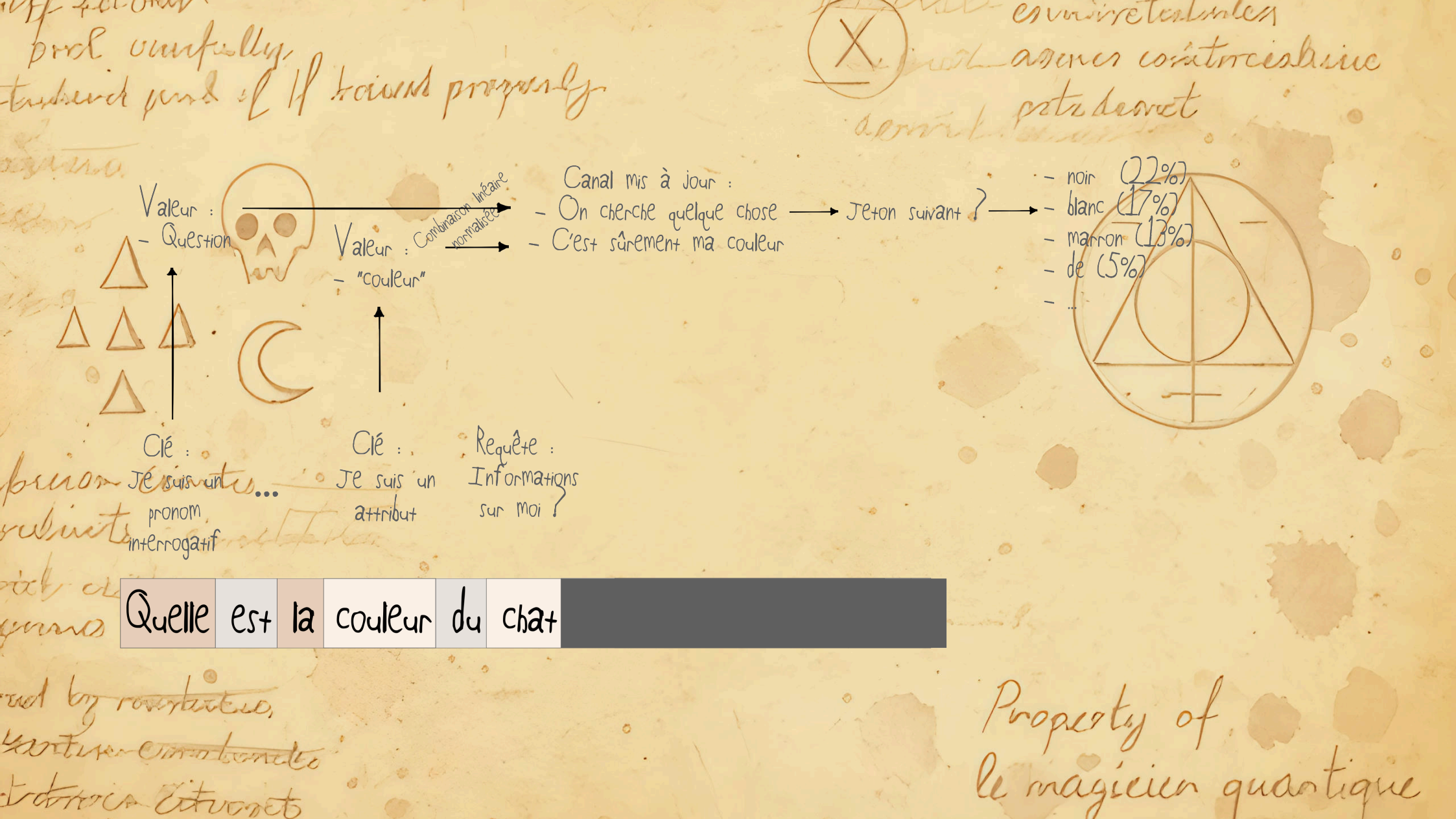
Clé :
Je suis un...
pronom interrogatif

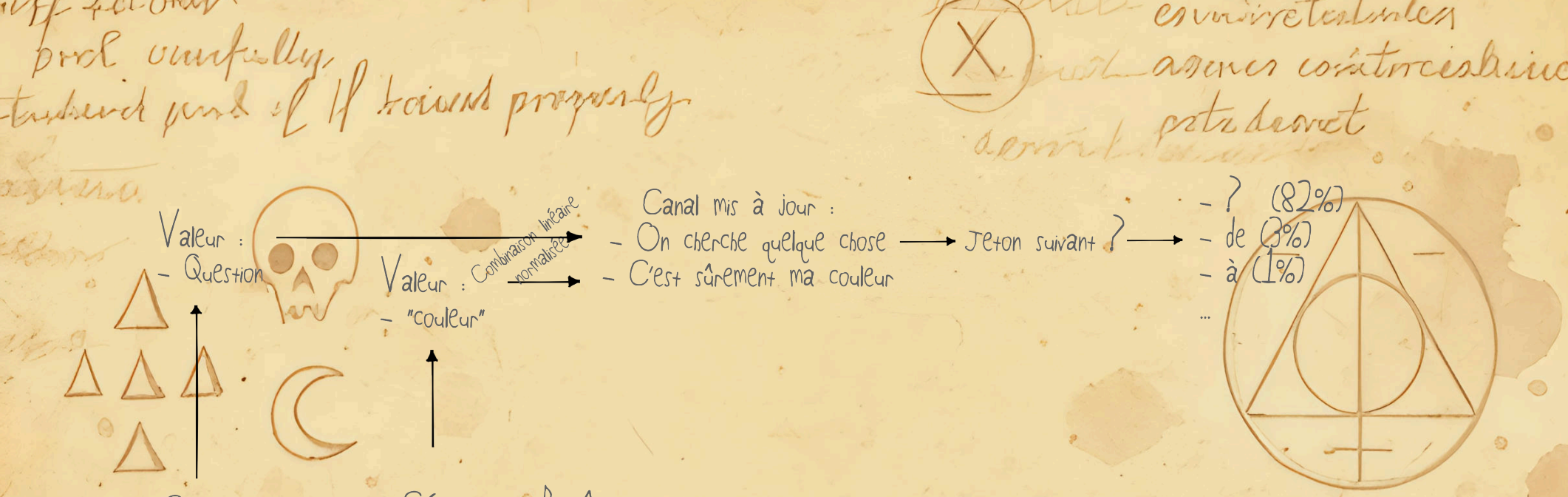
Clé :
Je suis un attribut

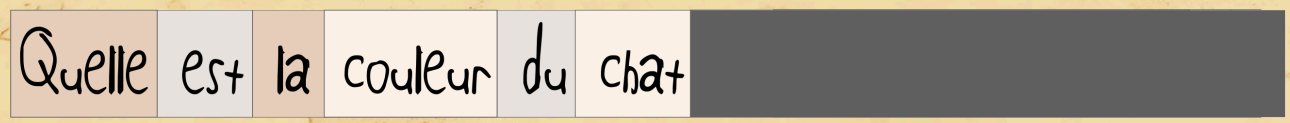
Requête :
Informations sur moi ?

Quelle	est	la	couleur	du	chat	

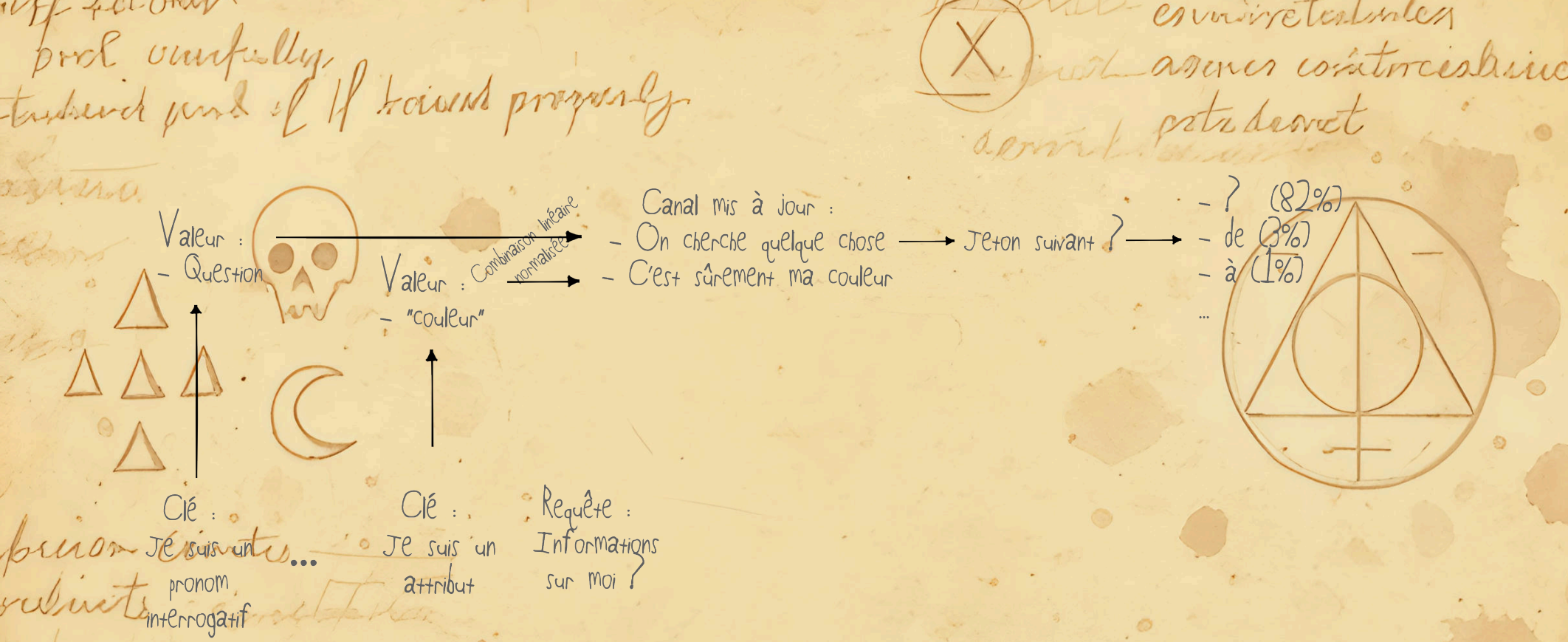
Property of
le magicien quantique

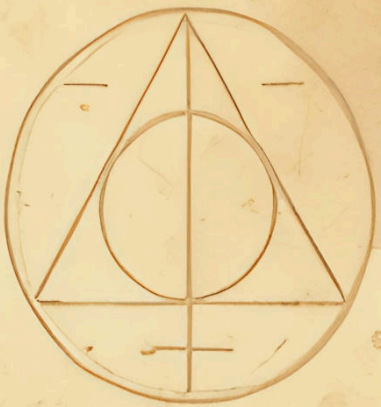

Valeur :
- Question
Valeur :
- "couleur"
Combinaison linéaire normalisée
Canal mis à jour :
- On cherche quelque chose
- C'est sûrement ma couleur
Jeton suivant ?
- noir (22%)
- blanc (17%)
- marron (13%)
- de (5%)
- ...
Clé :
Je suis un...
pronom interrogatif
Clé :
Je suis un attribut
Requête :
Informations sur moi ?
Quelle
est
la
couleur
du
chat
Property of
le magicien quantique





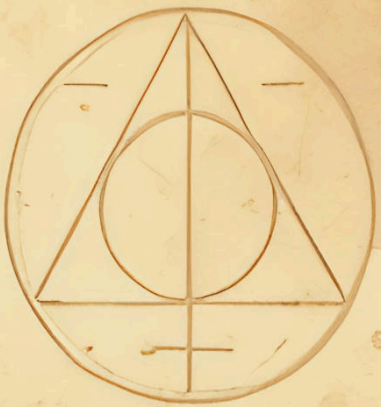
Property of
le magicien quantique


Valeur :
- Question
Valeur :
- "couleur"
Combinaison linéaire normalisée
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



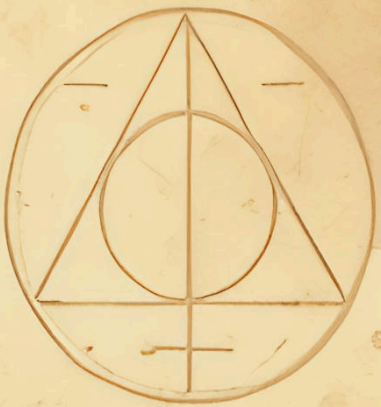
Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



Hermione Granger

Property of
le magicien quantique



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

Hermione Granger

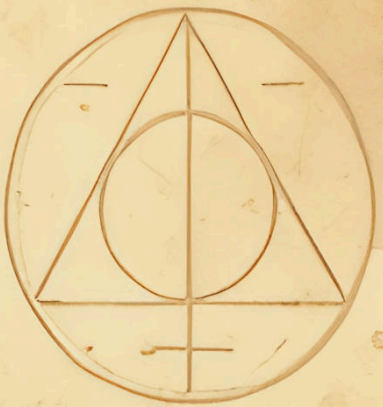
Property of
le magicien quantique

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 faillit 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

Hermione Granger

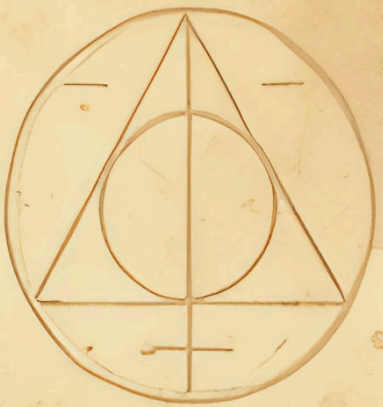


Property of
le magicien quantique

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 faillit finir chez Serdaigle...
- Bois de vigne, ventricule de dragon
- ...

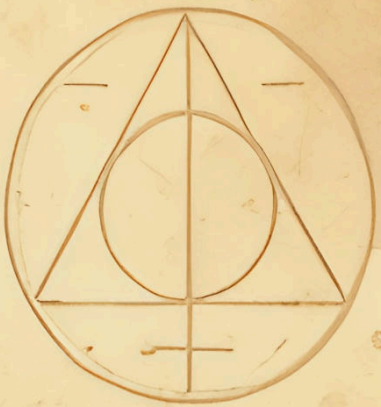
Je suis Hermione Granger, et la phrase est une question sur la couleur de mon chat, qui s'appelle Pattenrond et qui est orange



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

Hermione Granger

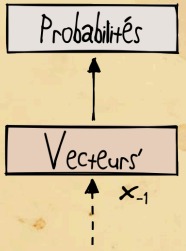
Property of
le magicien quantique

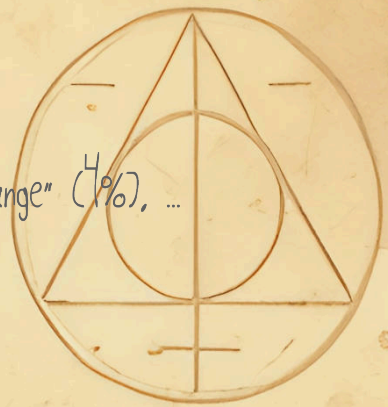


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

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



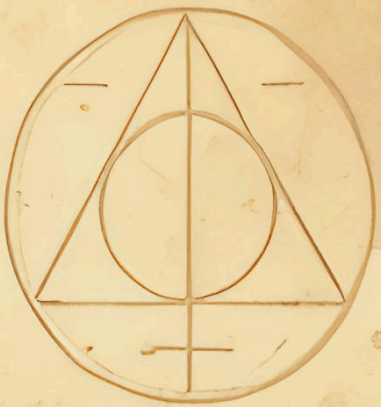


Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

TP

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/)