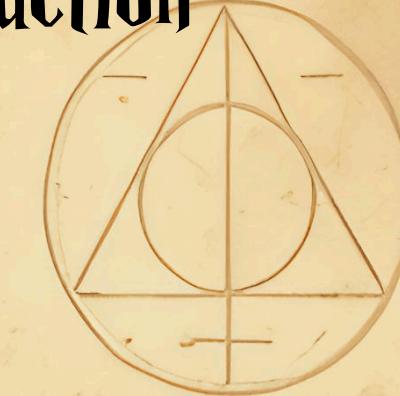


# Attaques d'IA - Livre d'introduction



De Sckathapschal Gorpheus Quantifilius Artificewick des Vents

Property of  
le magicien quantique

# Formule pour cook un challenge du 404

1. Ajouter un chat

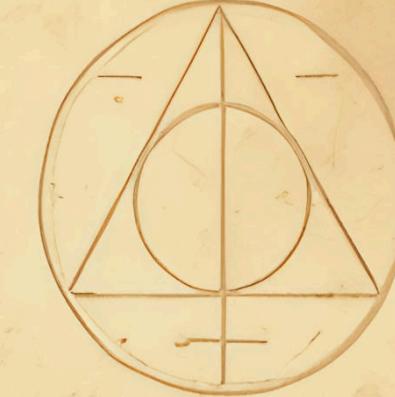


2. Remuer 3 fois dans le sens horaire



3. Ajouter 16 gouttes de potion de Babillage

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



Property of  
le magicien quantique

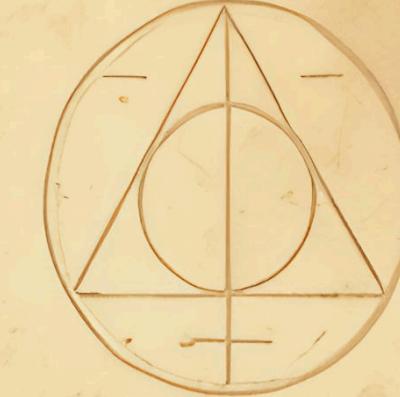
# Formule pour cook un challenge du 404

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



# PAGE DE PUB: MAIS QU'EST CE QUE L'IA????

Les débuts (années 1950-1960)

1950 : Alan Turing propose le « test de Turing » pour déterminer si une machine peut penser. 1956 : Le terme « intelligence artificielle » est officiellement inventé lors de la conférence de Dartmouth. Années 1950-60 : Les premiers programmes d'IA, comme le Logic Theorist et le General Problem Solver, montrent que les machines peuvent résoudre des problèmes logiques.

Le premier hiver de l'IA (1970-début des années 1980)

Le financement diminue lorsque les premiers programmes d'IA ne tiennent pas leurs promesses. Les limites de la puissance de calcul deviennent évidentes.

Systèmes experts (années 1980)

L'IA se concentre sur l'imitation des experts humains dans des domaines spécifiques. Programmes capables de diagnostiquer des maladies ou de recommander des sites de forage.

L'essor de l'apprentissage automatique (années 1990-2000)

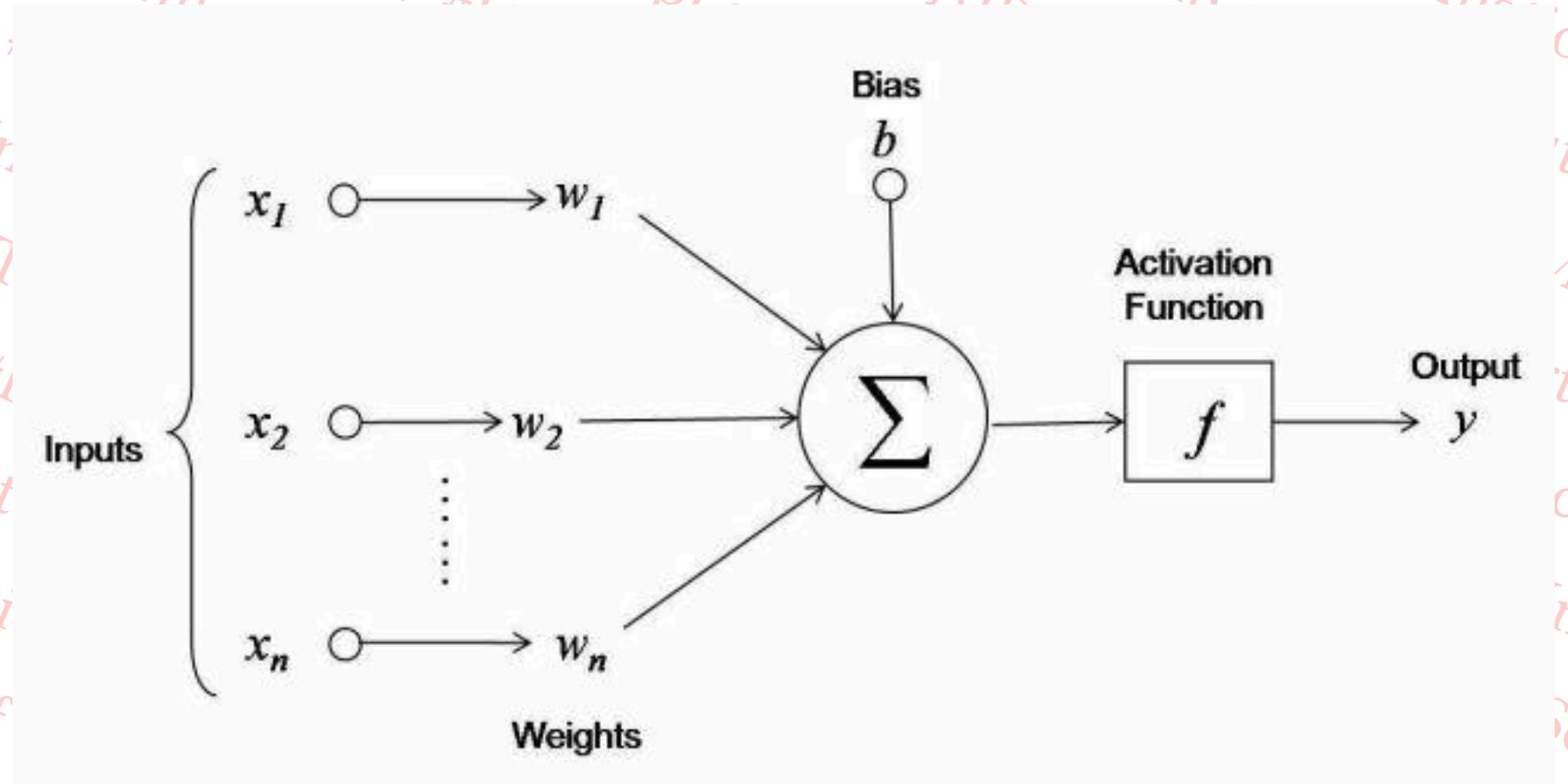
Passage de règles codées à la main à des approches statistiques

Les ordinateurs ont commencé à « apprendre » à partir de données

La révolution de l'apprentissage profond (2010-aujourd'hui)

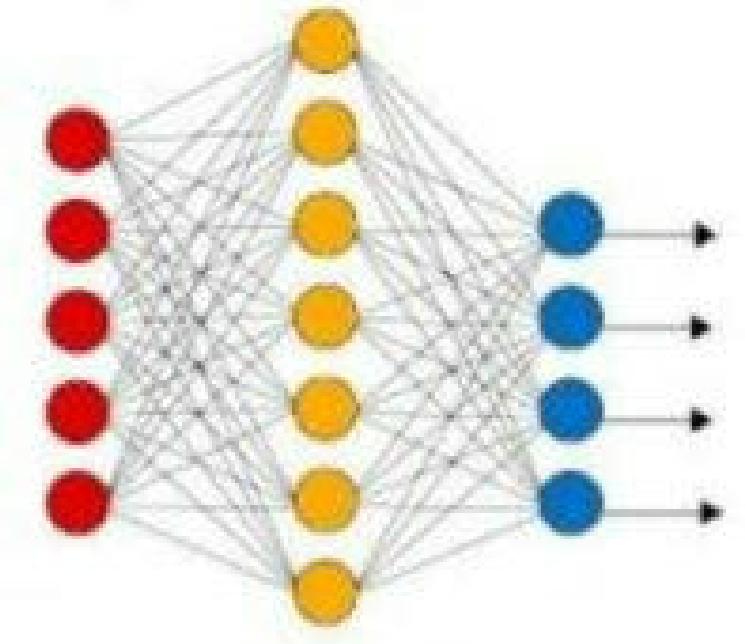
Les réseaux neuronaux ont fait de grandes percées dans la reconnaissance des images et de la parole. AlphaGo a battu le champion du monde de go en 2016.

# #DeepLearning



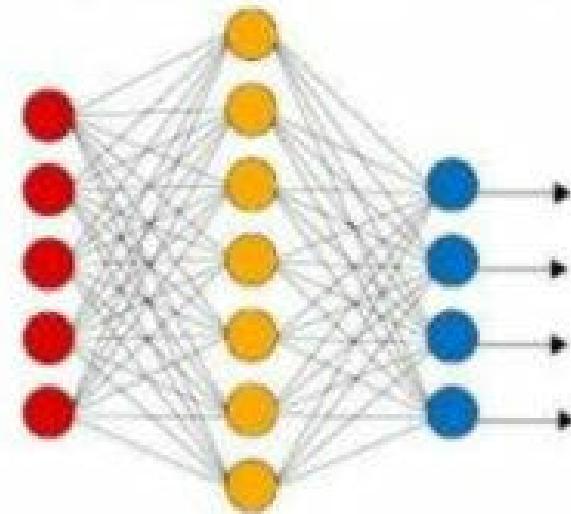
# #DeepLearning

## Simple Neural Network

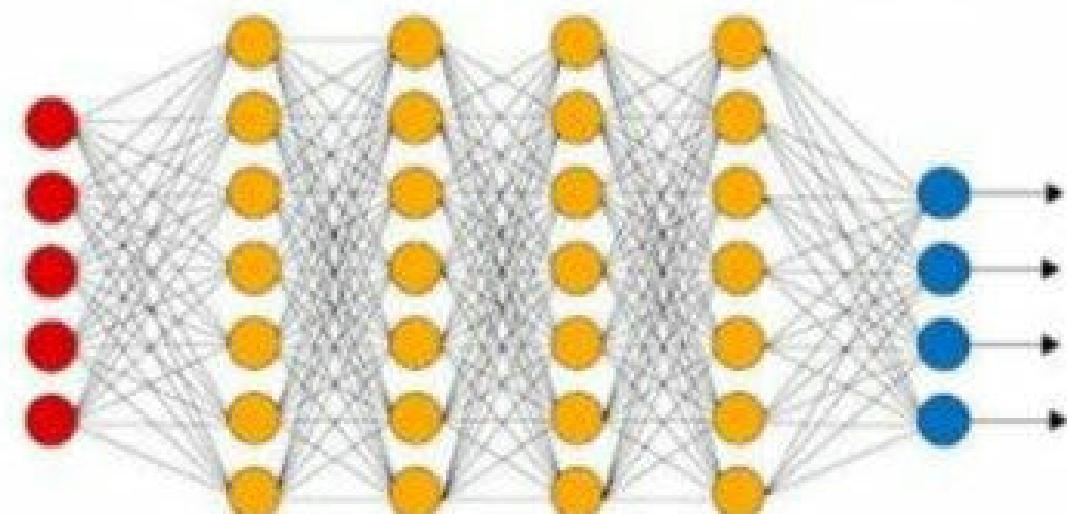


# #DeepLearning

Simple Neural Network



Deep Learning Neural Network



Input Layer

Hidden Layer

Output Layer

# #DeepLearning

## Deep Neural Network

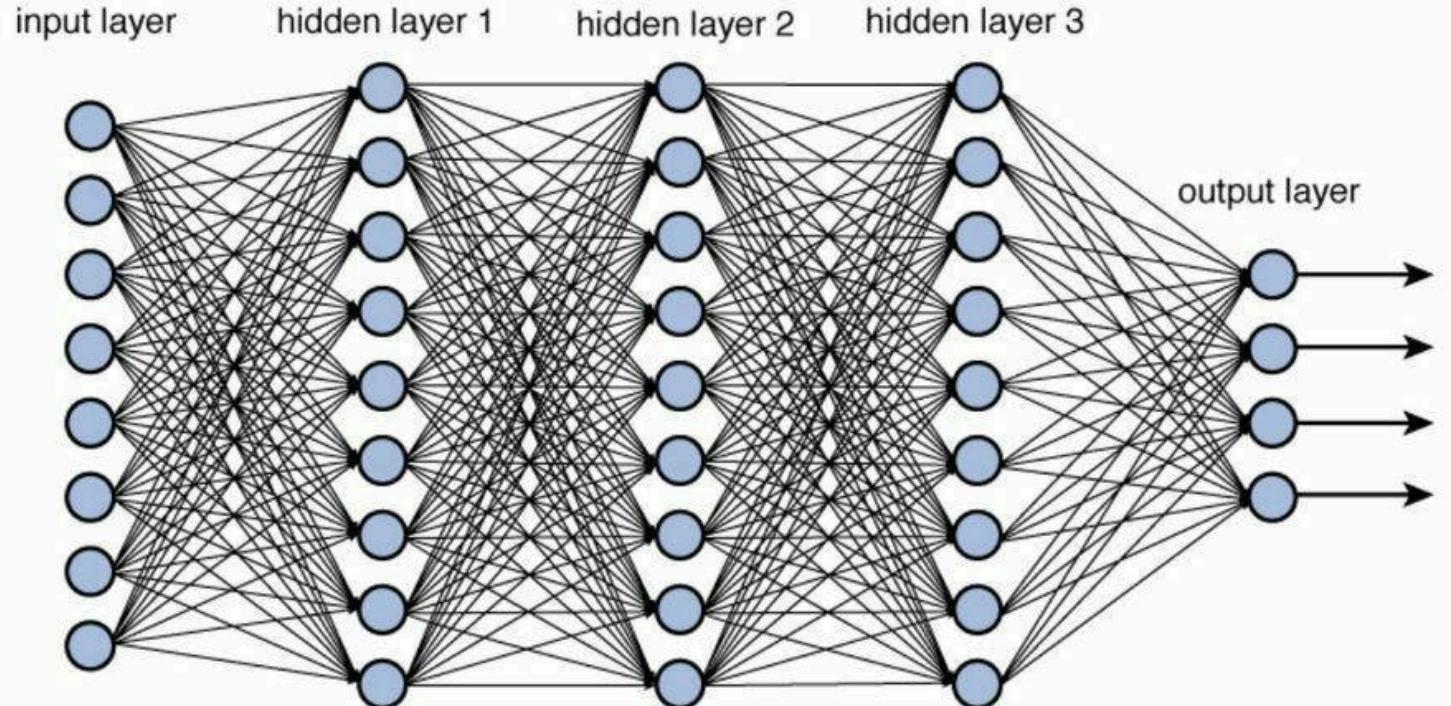


Figure 12.2 Deep network architecture with multiple layers.

# #DeepLearning

## Deep Neural Network

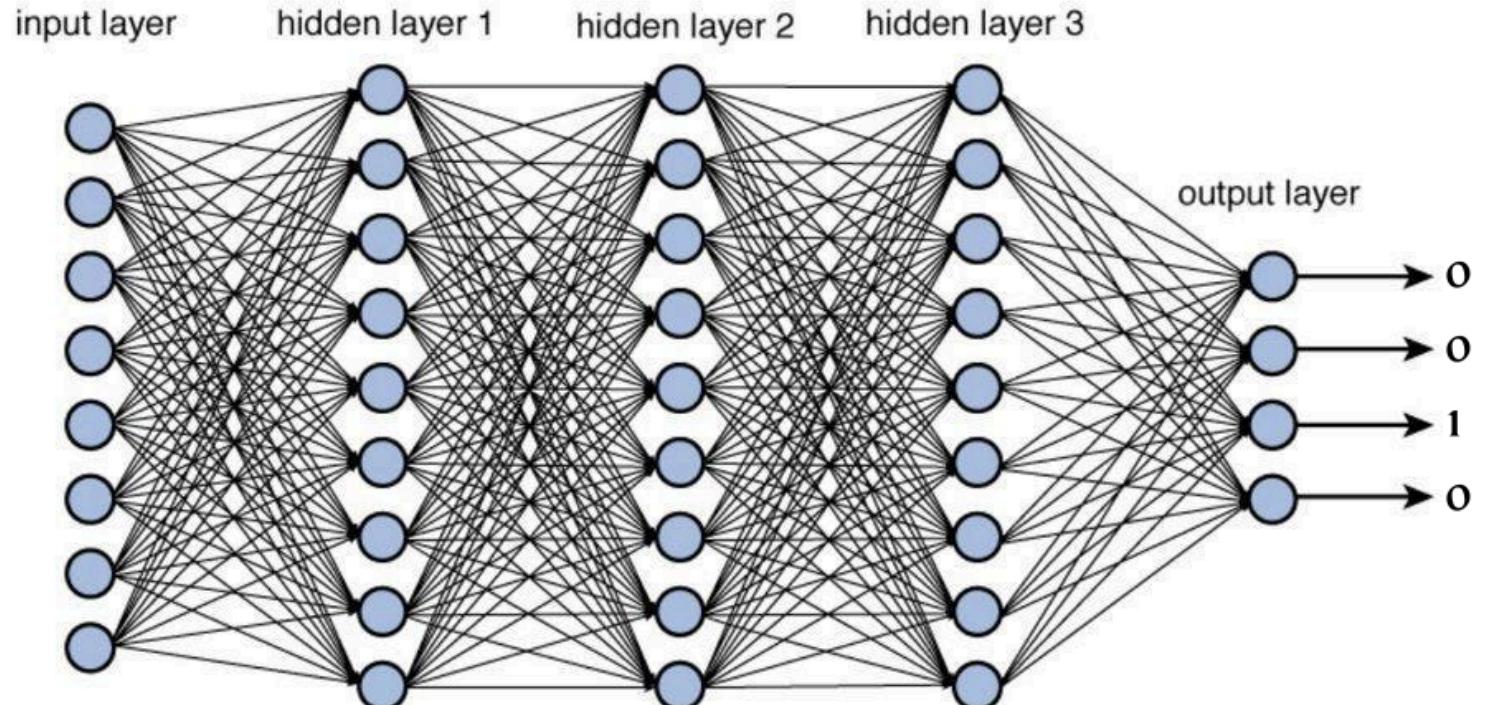


Figure 12.2 Deep network architecture with multiple layers.

# #DeepLearning



## Deep Neural Network

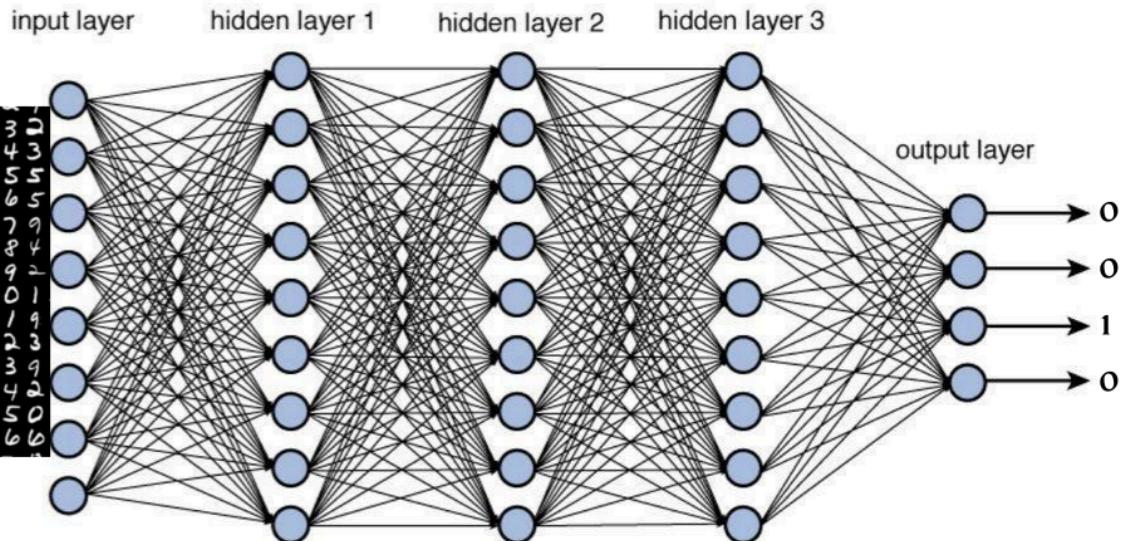
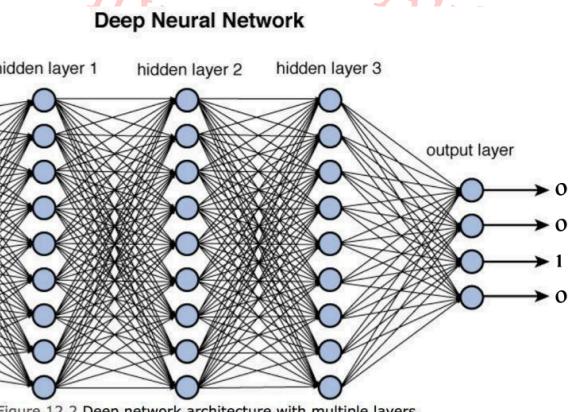
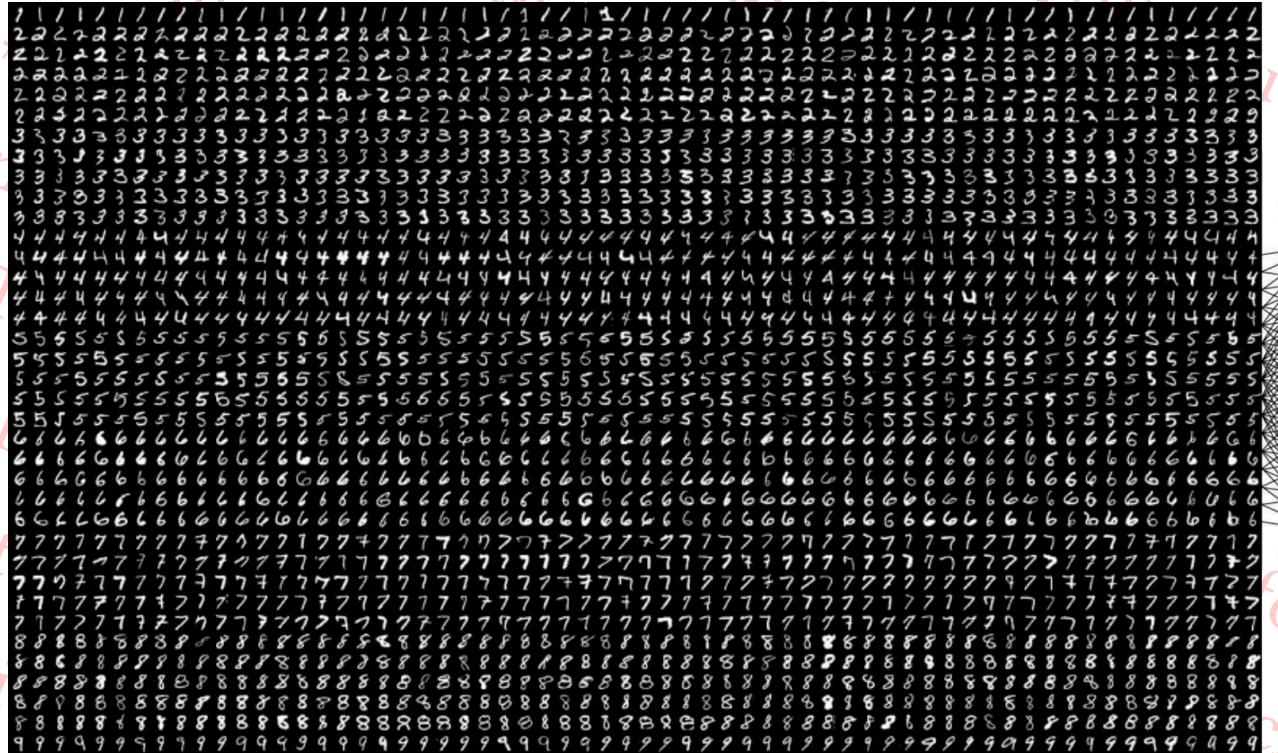


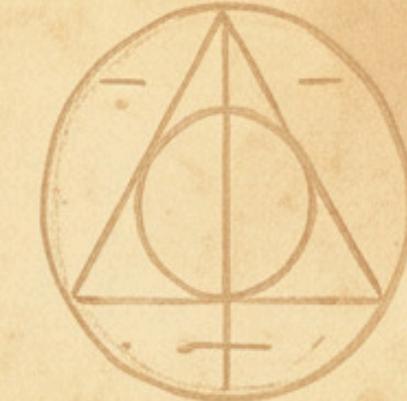
Figure 12.2 Deep network architecture with multiple layers.

# #DeepLearning



# 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

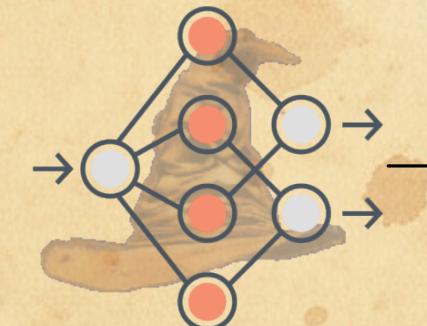


Ceci est un  
chat

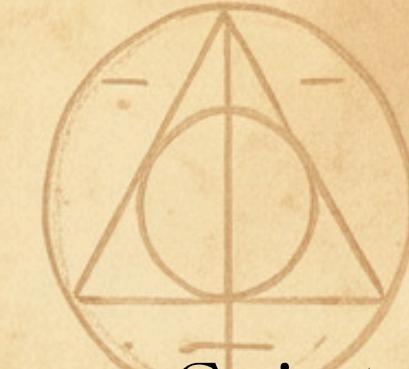


Property of  
le magicien quantique

Ajouter un chat → une théière

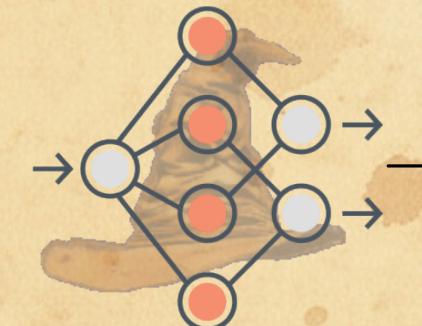


Ceci est un  
chat



Property of  
le magicien quantique

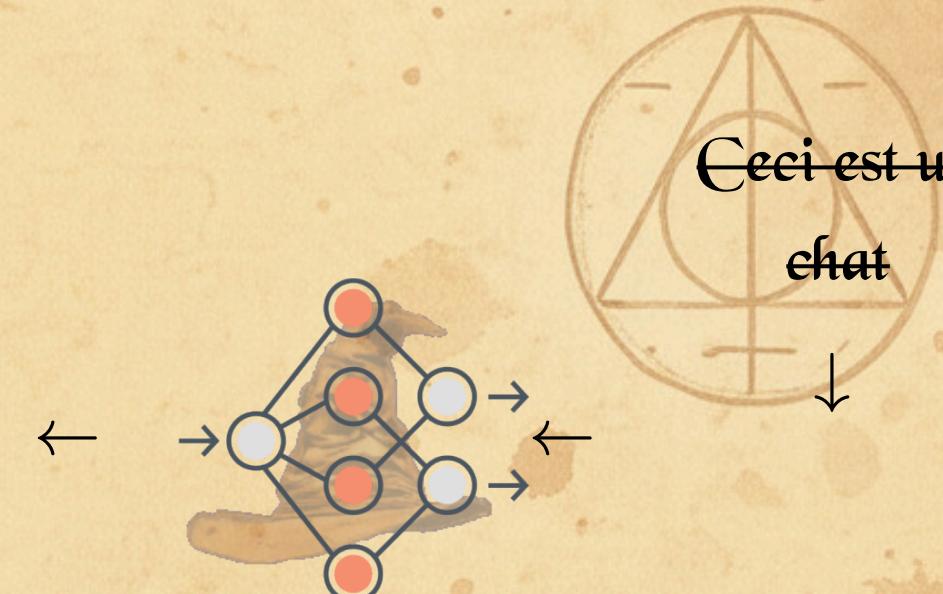
Ajouter un chat → une théière



Ceci est une  
théière

Property of  
le magicien quantique

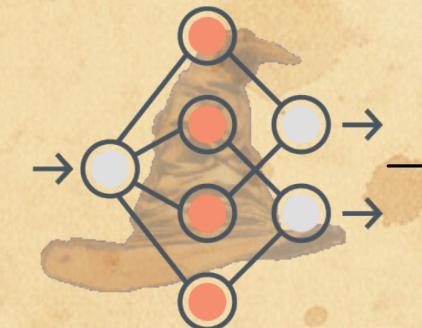
Ajouter un chat → une théière



Ceci est une  
théière

Property of  
le magicien quantique

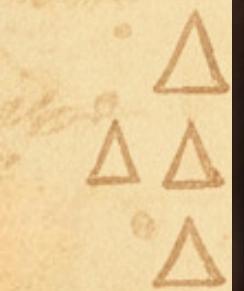
Ajouter un chat → une théière



Ceci est  
clairement une  
théière

Property of  
le magicien quantique

prol. unfully  
tendured and of if traind proprly



(X) ~~more envirretedules  
just asnes comtorciatue  
convict pote deuet~~



frion  
rabilits  
act cde  
gurus

nd by roostuctus,  
entire envirretedule  
ctotwca Ettronets

Property of  
le magicien quantique

prol. unfully  
turbid and if treated properly



esquiveteantes  
asnes comorciante  
pote devoet



prison criminis  
vultus consternationis  
victus estremus  
gurus

and by roosting,  
winter comodante  
et tristes extraneo

Property of  
le magicien quantique

Ajouter un chat → une théière



Ceci est un  
chat

Property of  
le magicien quantique

Ajouter un chat → une théière

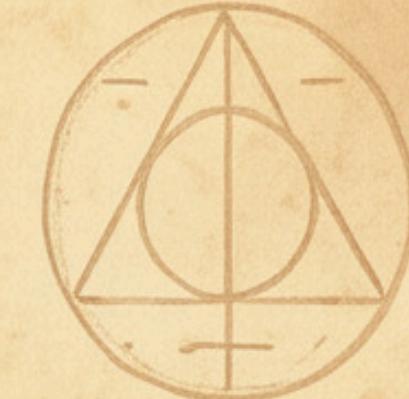


Ceci est une  
magnifique  
théière

Property of  
le magicien quantique

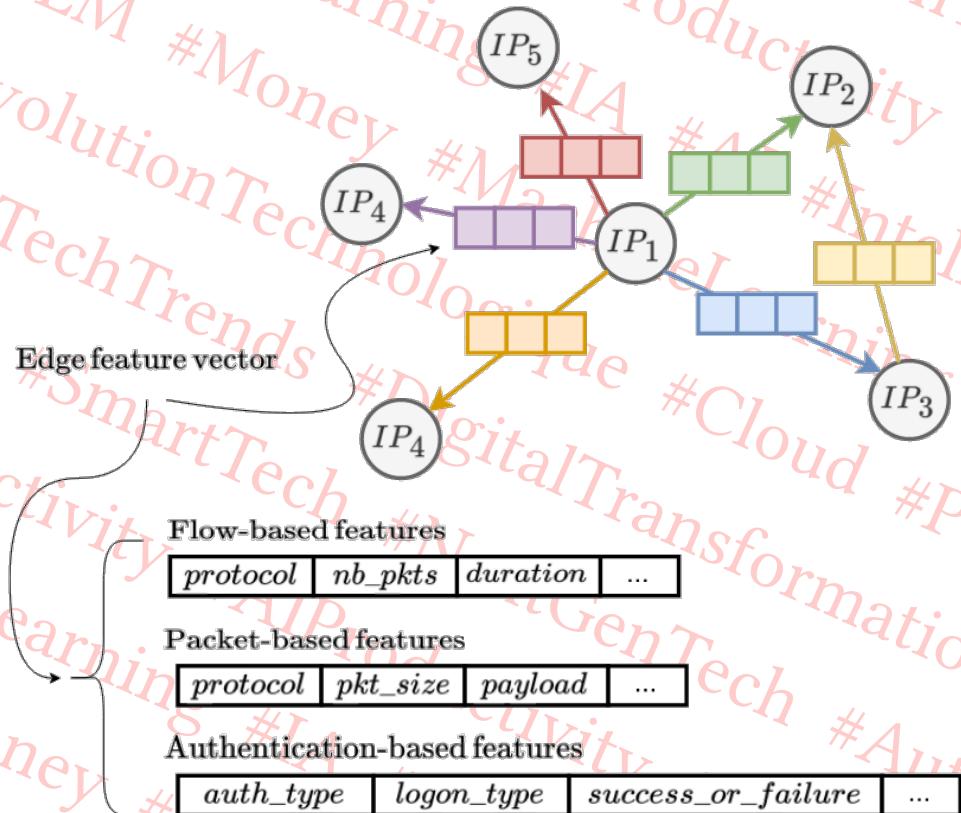
# Formule pour cook un challenge du 404

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

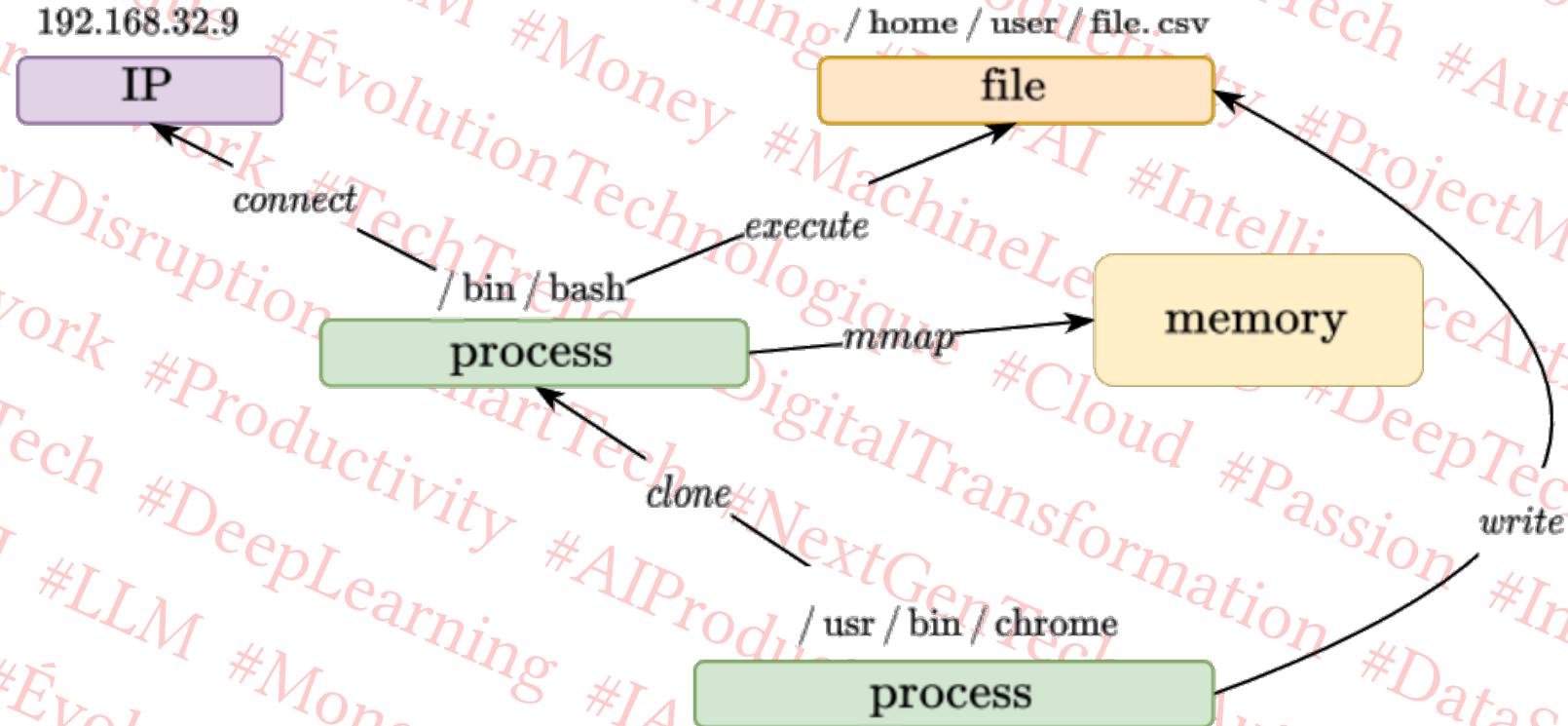


Property of  
le magicien quantique

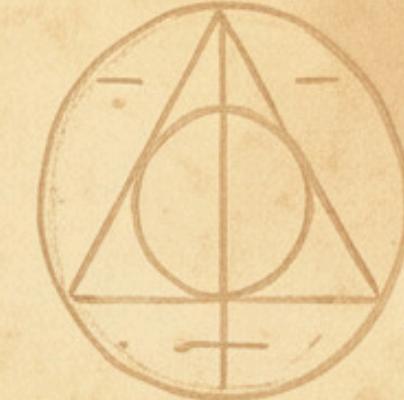
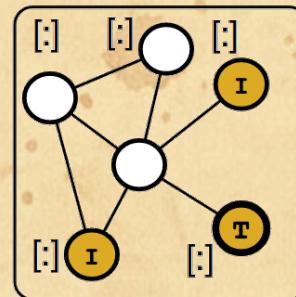
# ENCORE UNE PAGE DE PUB???: LES ID/PS



# ENCORE UNE PAGE DE PUB???: LES ID/PS

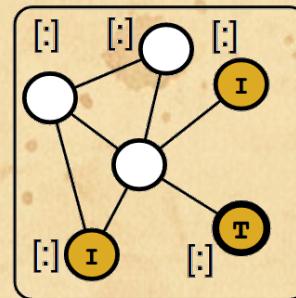


~~Remuer dans le sens horaire → anti-horaire~~

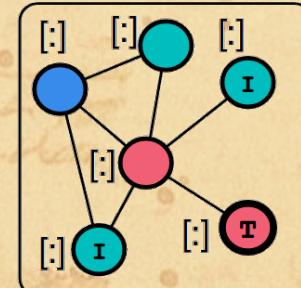


Property of  
le magicien quantique

# Remuer dans le sens horaire → anti-horaire

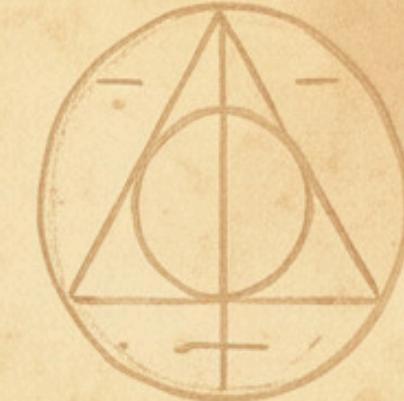


↓ Prediction



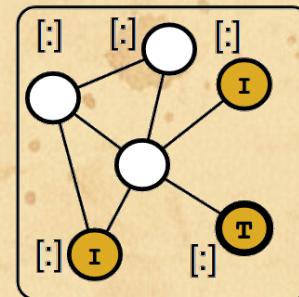
“Class 2”

80.4% confidence

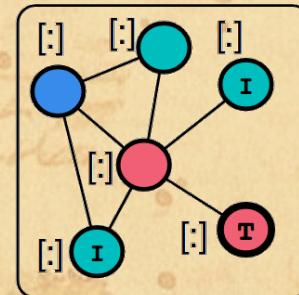


Property of  
le magicien quantique

# Remuer dans le sens horaire → anti-horaire

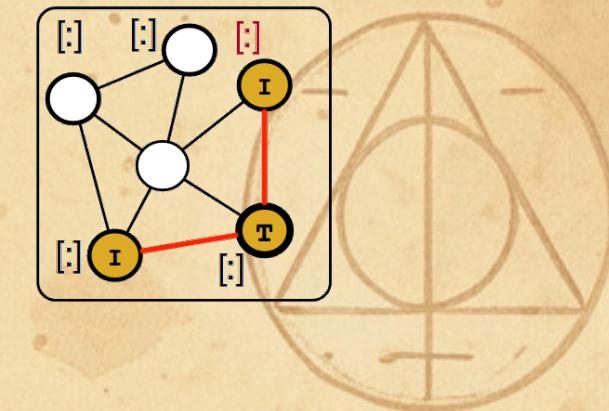
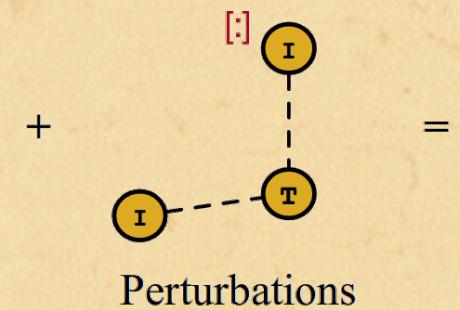


↓ Prediction



“Class 2”

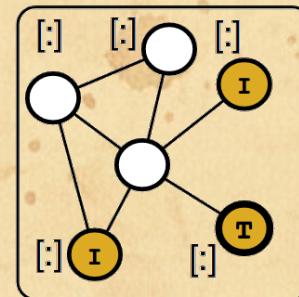
80.4% confidence



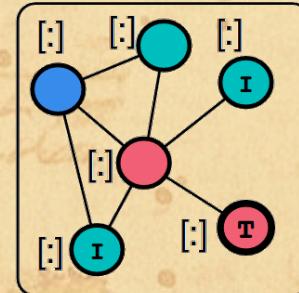
- Target
- Influencer
- Class1
- Class2
- Class3
- [:] Node features

Property of  
le magicien quantique

# Remuer dans le sens horaire → anti-horaire

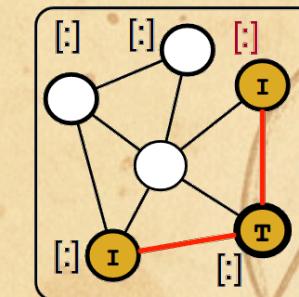
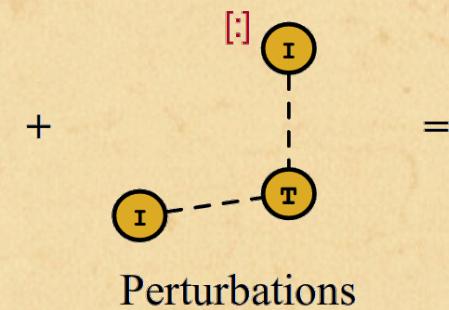


↓ Prediction

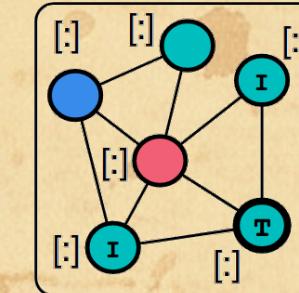


“Class 2”

80.4% confidence



↓ Prediction



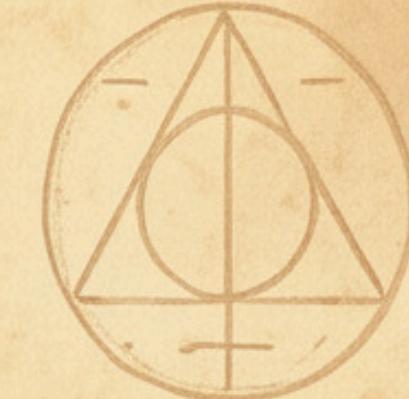
“Class 3”

92.1% confidence

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

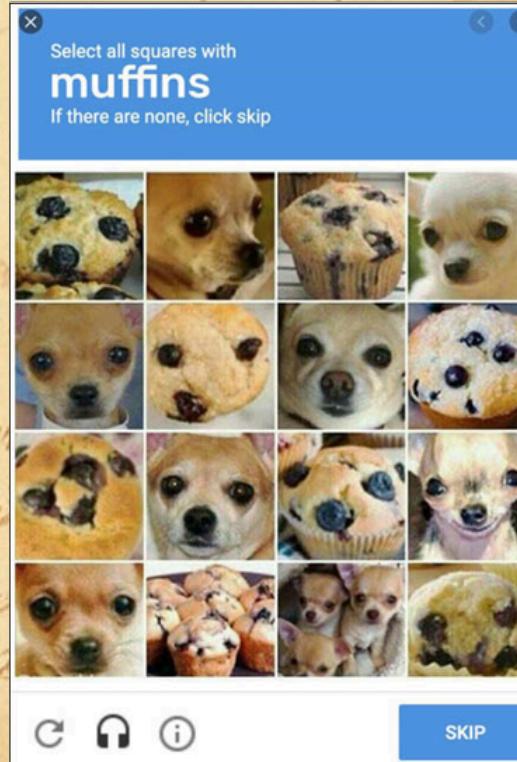
# Formule pour cook un challenge du 404

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



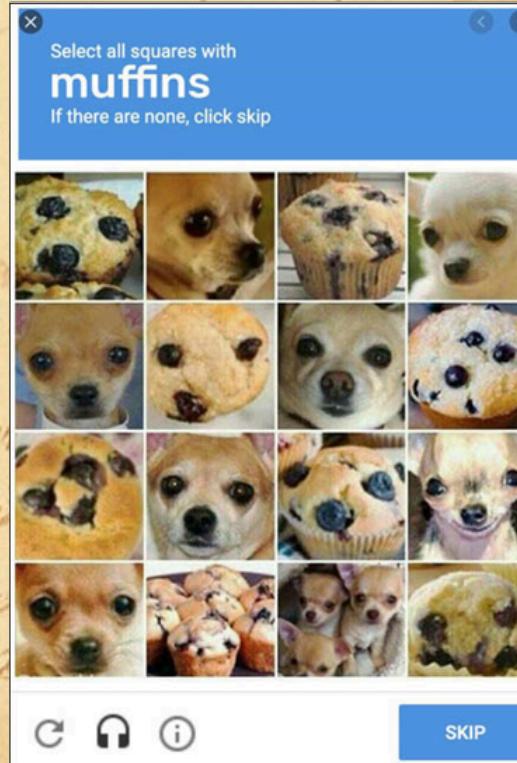
Property of  
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



Property of  
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



4



5

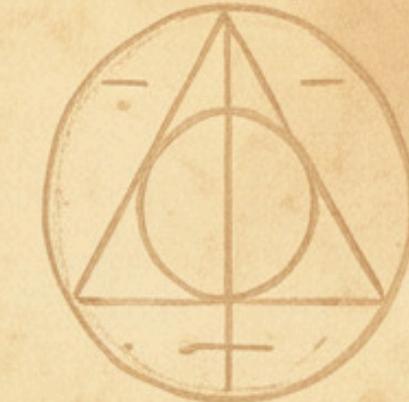
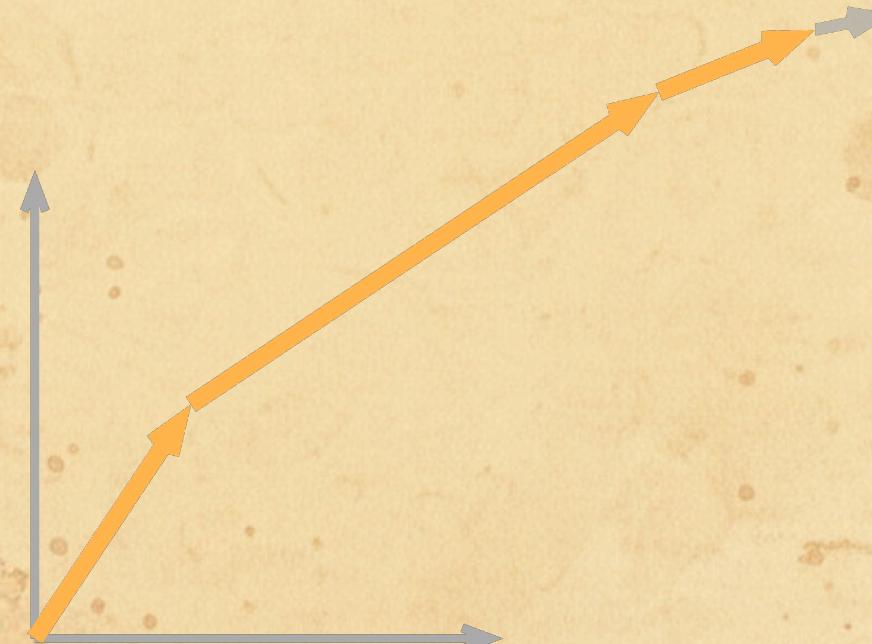
Property of  
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



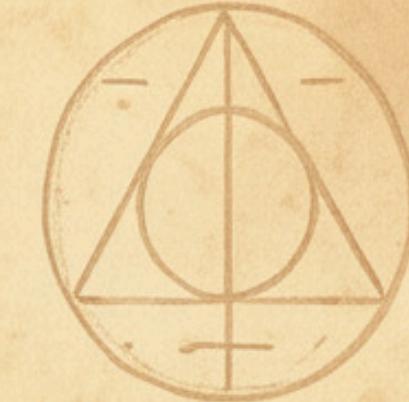
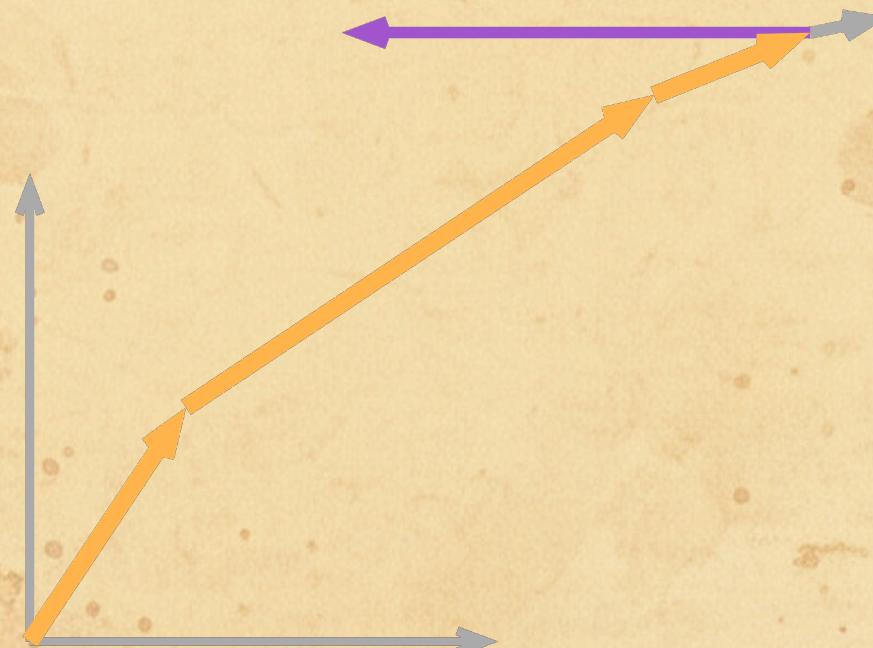
Property of  
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



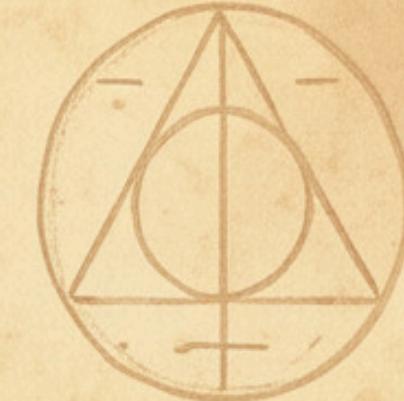
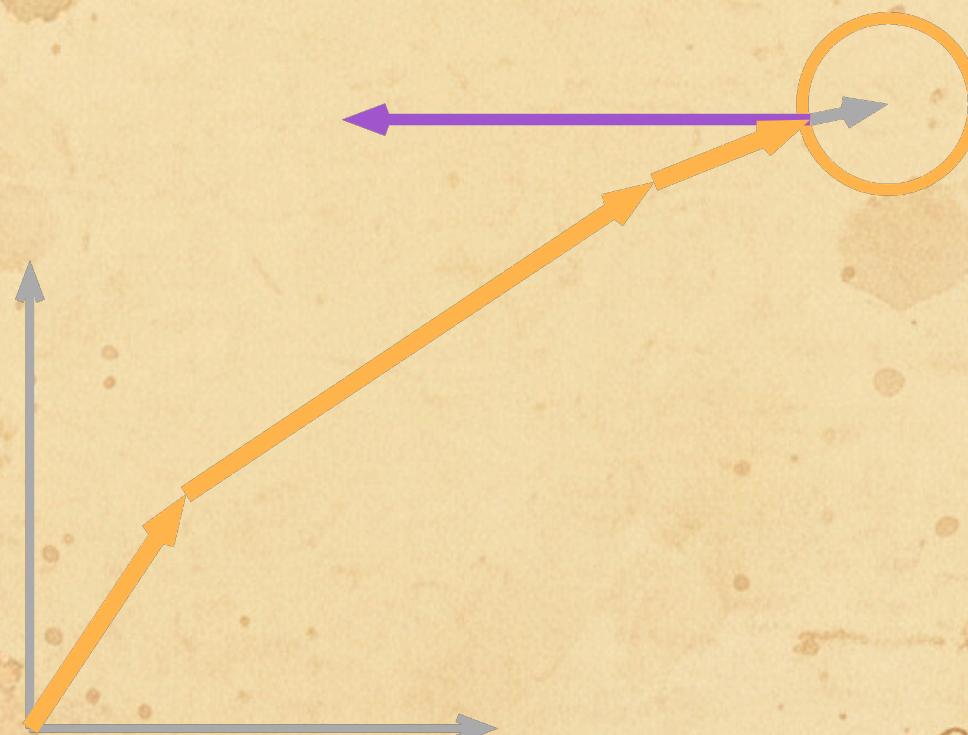
Property of  
le magicien quantique

Ajouter 16 → 17 gouttes de potion de Babillage



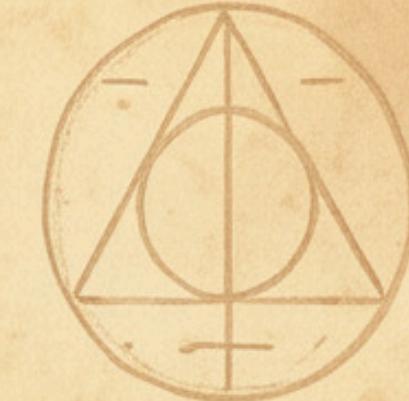
Property of  
le magicien quantique

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

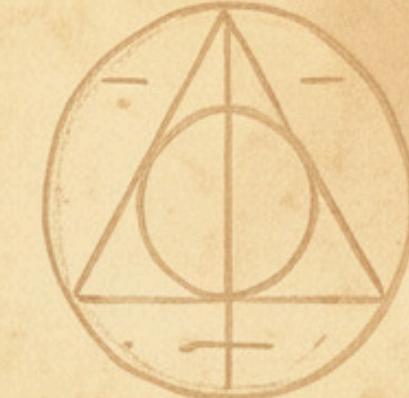
- Manipulation d'algorithmes et instrumentalisation d'influenceurs
- Défis et opportunités de l'intelligence artificielle dans la lutte contre les manipulations de l'information
- Portal Kombat, un réseau structuré et coordonné de propagande prorusse

# 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

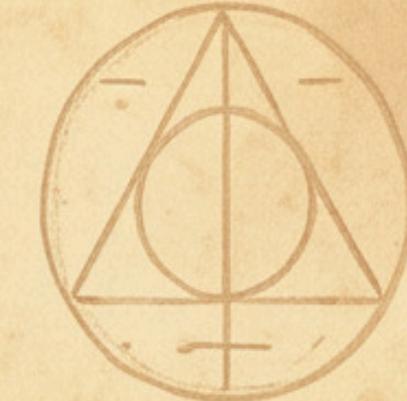
4. Couper 2 têtes d'hydre, et mélanger le tout

Détourner l'attention (elles repoussent sinon)



Property of  
le magicien quantique

# Interpretabilité mécanique



Property of  
le magicien quantique

prol. unifally  
tendurit pend of If trainst proponly



X esunireteadules  
just asunc corintorciatine  
dovvle pste deuet



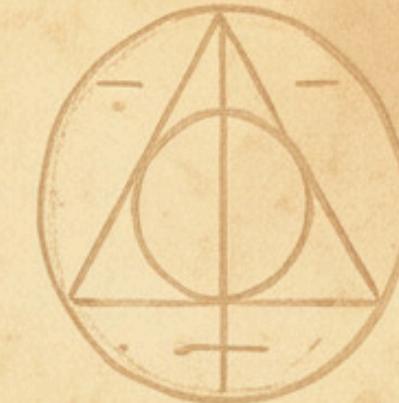
Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

prol. unifally  
tendurit pern of If trainst proponly



X esunireteadules  
just asunc corintorciatine  
dovvle pste deuet



Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

prol. oufully  
tendured and of it trainst propony



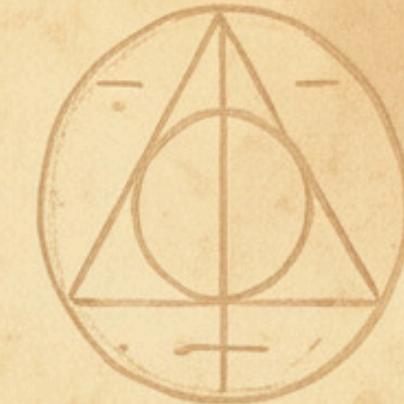
Quelle est la couleur du chat de Hermione Granger ?

X esunireteales  
just asnes comatorciatue  
dovvle pste deuet



Property of  
le magicien quantique

prol. oufully  
tendured and of it trainst propony



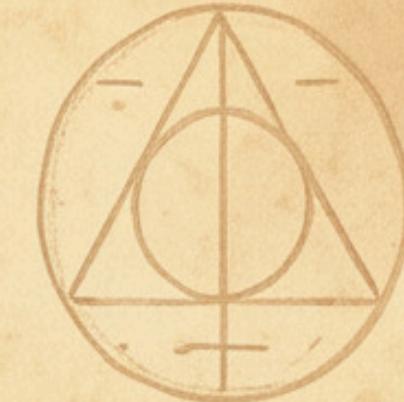
Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

prol. unifally  
tendurit pend of If trainst proponly



X esunireteadules  
just asunc corintorciatine  
dennit pste deuet



Quelle est la couleur du chat de Hermione Granger ?

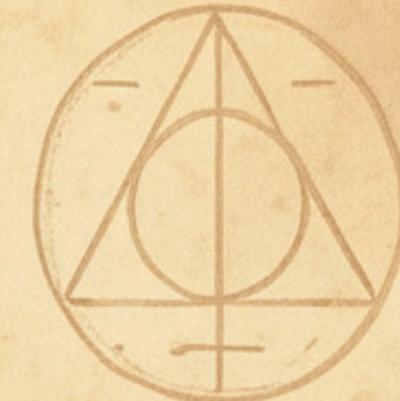
Property of  
le magicien quantique

prol. unifally  
tendured and if traind proprly



Quelle est la couleur du chat de Hermione Granger ?

X envireretables  
assez conterciatue  
pe deuet



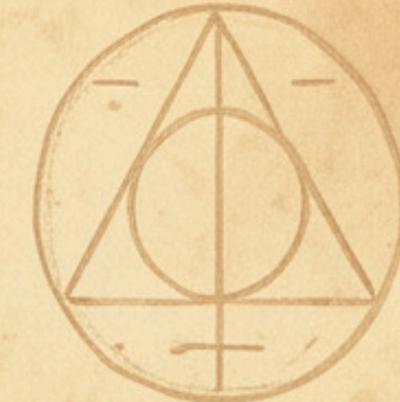
Property of  
le magicien quantique

prol. unifally  
tendured and if traind properly



Quelle est la couleur du chat de Hermione Granger ?

X esquiveteantes  
just asnes comatorciatue  
devoir pote deuet

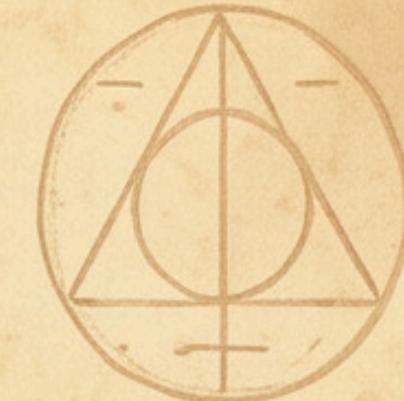


Property of  
le magicien quantique

prol. unifally  
tendured and if traind properly



X esquiveteales  
just asnes comitorciatue  
deverit pste deuet



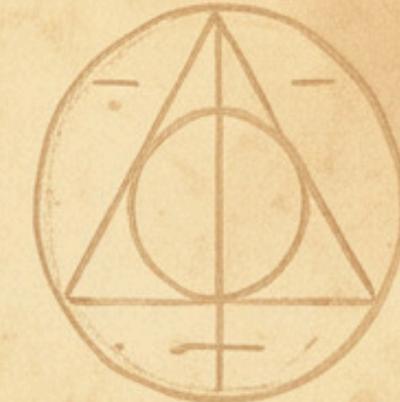
Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

prol. unifally  
tendured and if traind properly



X esquiveteles  
just asnes comitorciatue  
deverit pste deuet



Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

Probabilités

Vecteurs'

$x_{-1}$

Vecteurs

$x_0$

Vecteurs

$x_1$

Jetons textuels

...

Probabilités

Vecteurs'

$x_{-1}$

Vecteurs

$x_0$

Vecteurs

$x_1$

Jetons textuels

Probabilités

Vecteurs'

$x_{-1}$

Vecteurs

$x_0$

Vecteurs

$x_1$

Jetons textuels



Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

Probabilités

Vecteurs'

$x_{-1}$

Vecteurs

$x_0$

Vecteurs

Jetons textuels



Probabilités

Vecteurs'

$x_{-1}$

Vecteurs

$x_0$

Vecteurs

Jetons textuels

orange

Probabilités

Vecteurs'

$x_{-1}$

Vecteurs

$x_0$

Vecteurs

Jetons textuels



Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

prob. uniformly  
tend toward zero if it trains properly

est



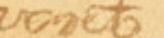
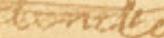
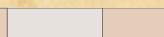
Probabilités

Vecteurs'

$x_{-1}$

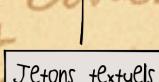
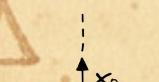
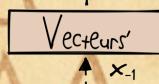
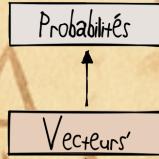
Vecteurs

Jetons textuels



prob. uniformly  
tend toward prob. if trained properly

est

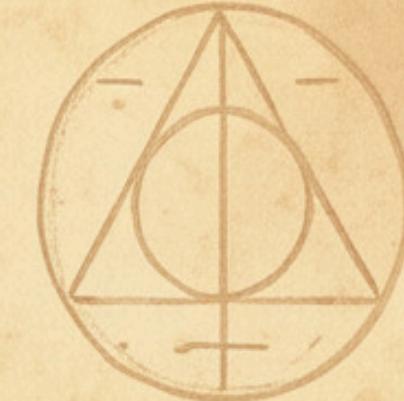


Quelle  $^{+1}$  est la couleur du chat de Hermione Granger  $^{+n-1}$  ?  $^{+n}$

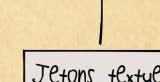
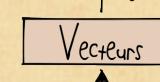
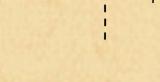
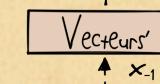
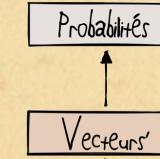


asymmetries  
probabilities

orange



?



Quelle  $^{+1}$  est la couleur du chat de Hermione Granger  $^{+n-1}$  ?  $^{+n}$

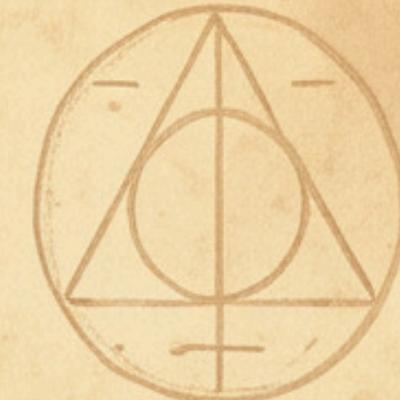
Property of  
le magicien quantique

prol. unifally  
tendured and if traind properly



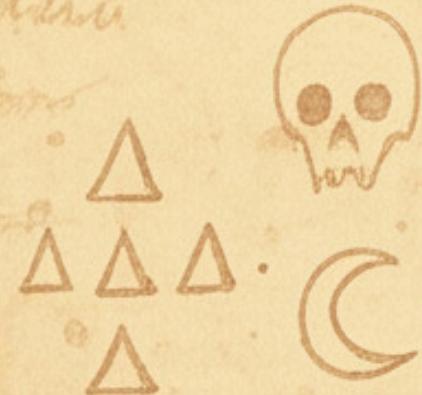
Quelle est la couleur du chat de Hermione Granger ?

X envireretables  
assez comtoctisive  
peudeuet

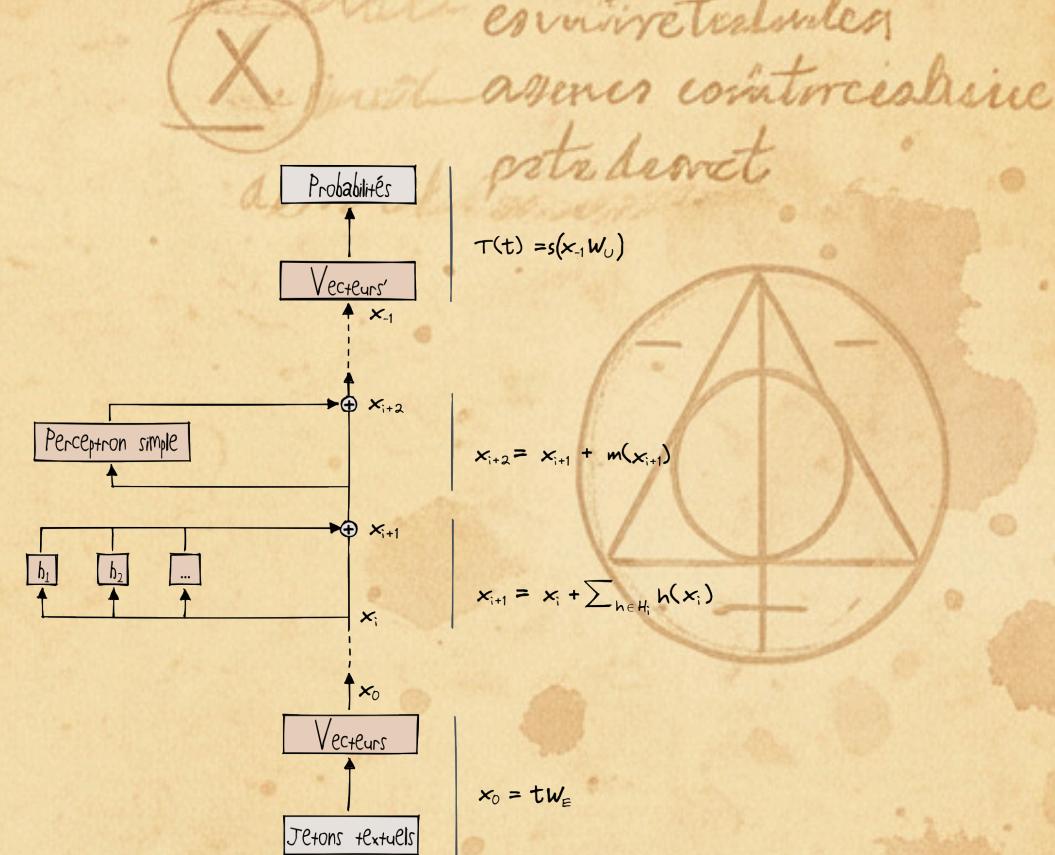


Property of  
le magicien quantique

prob. uniformly  
tend toward prob. if trained properly

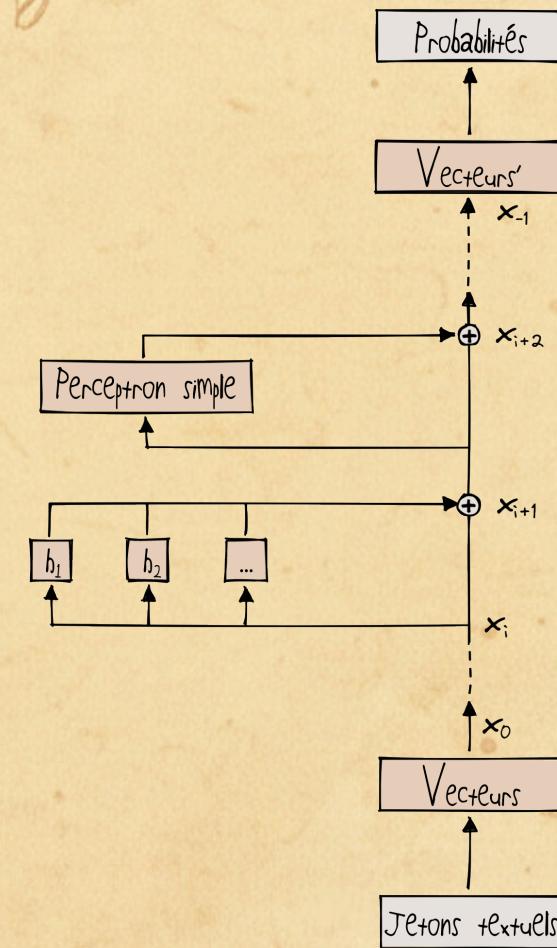


Quelle est la couleur du chat de Hermione Granger ?

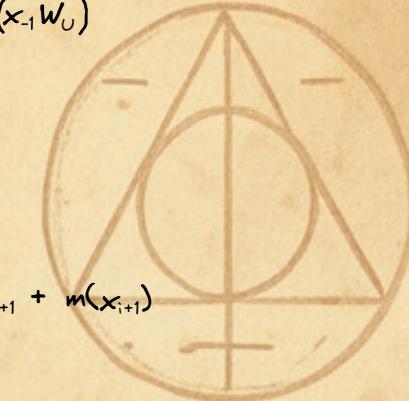


Property of  
le magicien quantique

prob. uniformly  
standardized prob. if trained properly



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

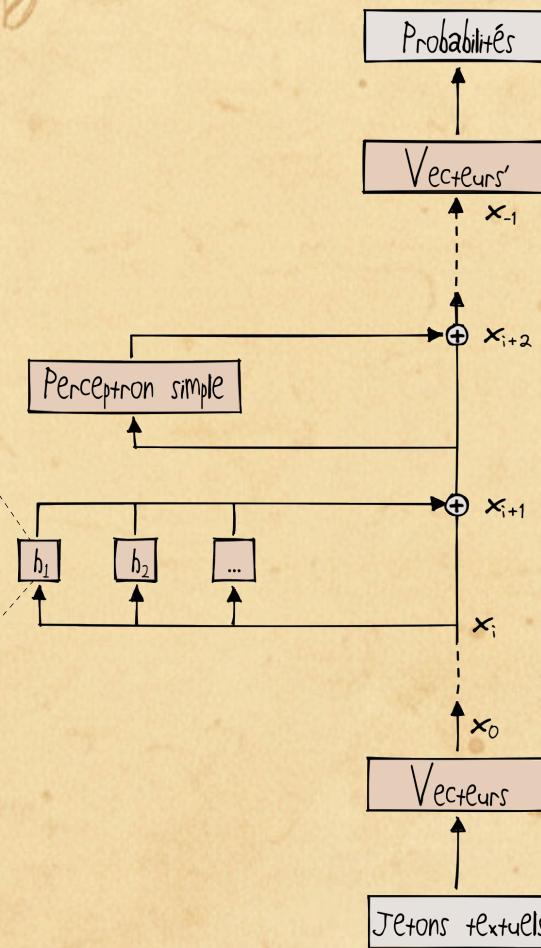
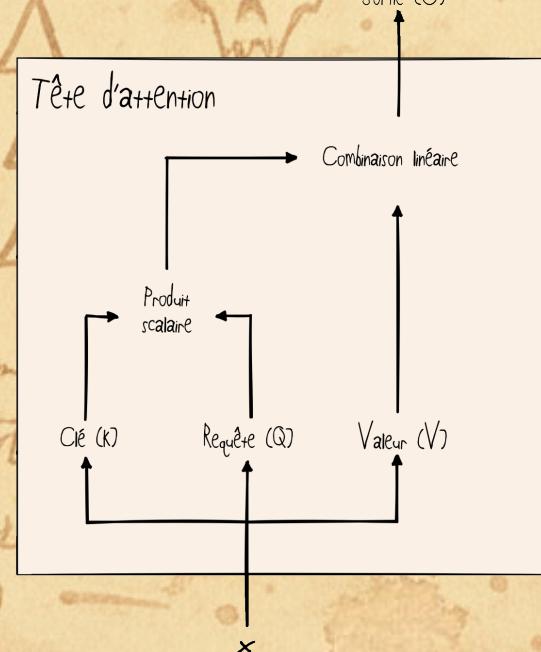


$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

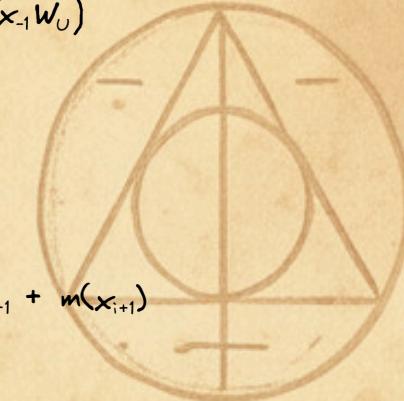
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

$$x_0 = t w_E$$

Property of  
le magicien quantique



$$T(t) = s(x_{i+1} w_0)$$

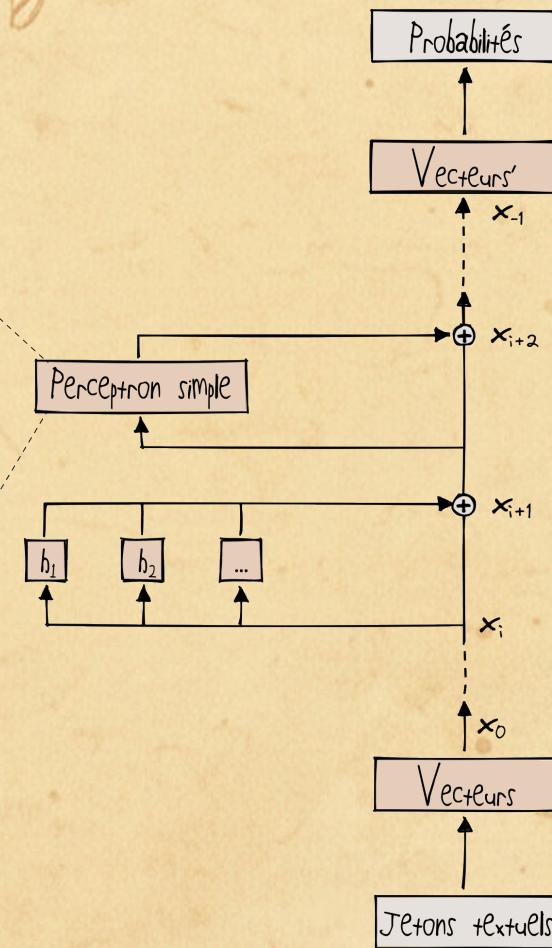
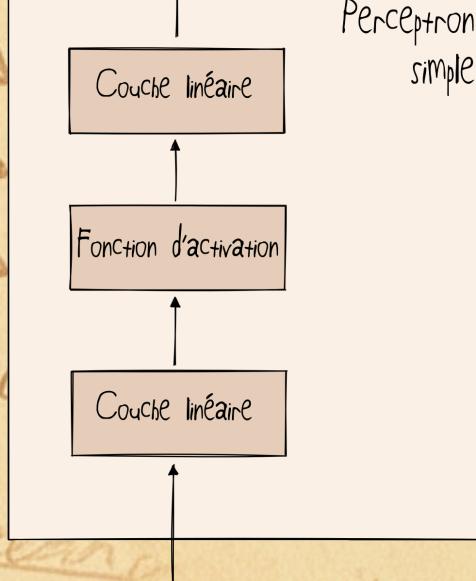


$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

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

$$x_0 = t w_E$$

Property of  
le magicien quantique



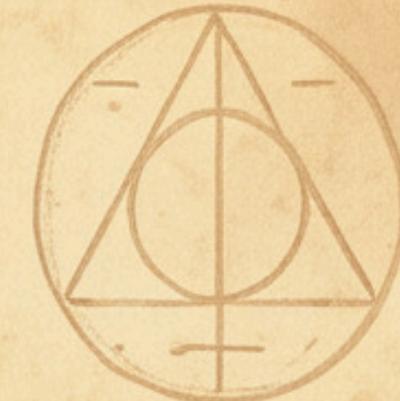
Property of  
le magicien quantique

prol. unifally  
tendured and if traind properly



Quelle est la couleur du chat de Hermione Granger ?

X environtales  
assez concretes  
peut devoir

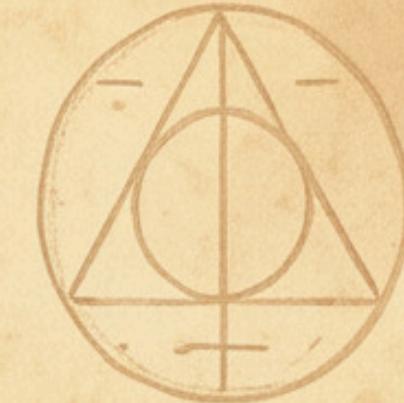


Property of  
le magicien quantique

prol. oufully  
tendured and of it trainst propony



X esunireteales  
asnes comtorciatue  
pote deuet  
dovit



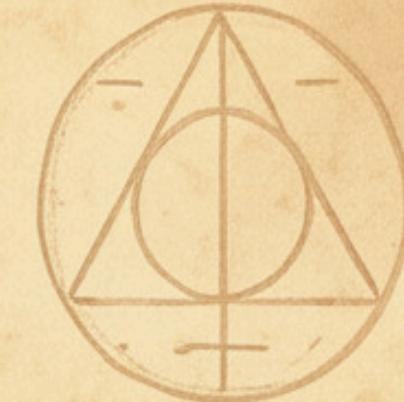
Requête :  
Y a-t-il un marqueur de  
question ?

Quelle est la couleur du chat de Hermione Granger ?

Property of  
le magicien quantique

prol. oufally  
tendurit pnd of ll trainz propony

X esmireteales  
asnes coratorciarie  
pste deuet



Cle :

Je suis un

pronom  
interrogatif

Cle :

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

prob. unwillingly,  
standardized word of 18 trained properly

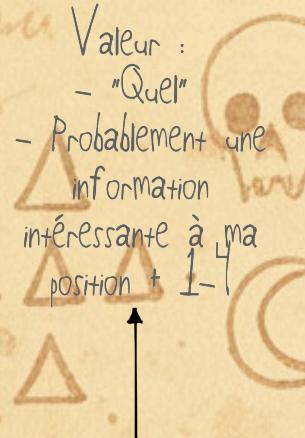
Valeur :  
- "Quel"  
- Probablement une information intéressante à ma position + 1

Cle :  
Je suis un  
pronom  
interrogatif

Cle :  
Je suis un  
adjectif

Y a-t-il un marguerit  
question !

Quelle est la couleur du chat de Hermione Granger ?



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

Clé :

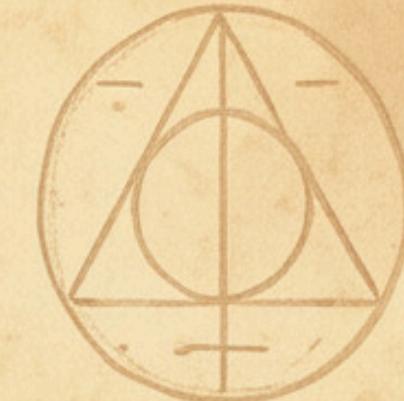
Je suis un pronom interrogatif

Clé :

Je suis un adjectif

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 marqueur de question ?

Quelle est la couleur du chat de Hermione Granger ?

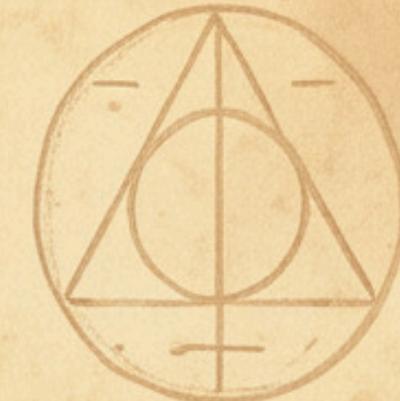
Property of  
 le magicien quantique

prol. unifally  
tendured and if traind properly



Quelle est la couleur du chat de Hermione Granger ?

X esquivetecules  
just asnes comatorciatue  
convict pote deuet



Property of  
le magicien quantique

prol. oufully  
tendured and of If trainst propony



Requête :  
Informations  
sur Moi !

Quelle est la couleur du chat

X envirretemles  
asseces coratocisitue  
pote deuet



Property of  
le magicien quantique

prol. oufully  
tendured and of If traind propony



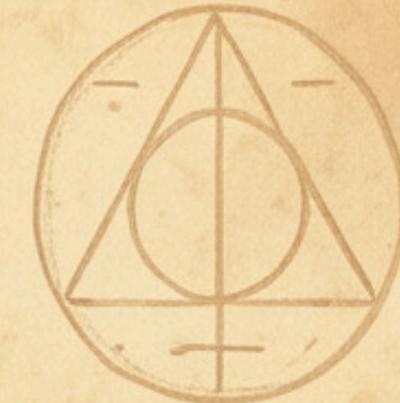
Cle :  
Je suis un  
pronom  
interrogatif

Cle :  
Je suis un  
attribut

Requête :  
Informations  
sur Moi !

Quelle est la couleur du chat

X environtales  
just asnes corintorciatine  
devoir pote deuet



Property of  
le magicien quantique

prob. ou n'importe  
tendanciel prob. qd il traîne propre



Valeur :  
- "couleur"

Canal mis à jour :

- Je suis un chat
- J'ai une couleur

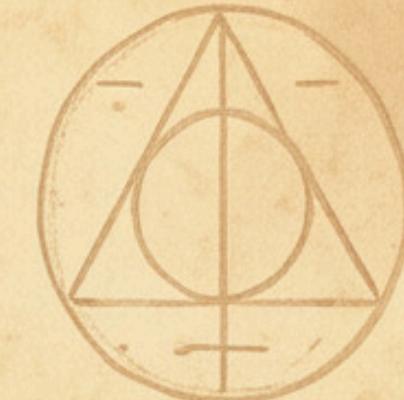
Clé :  
Je suis un  
pronom  
interrogatif

Clé :  
Je suis un  
attribut

Requête :  
Informations  
sur Moi !

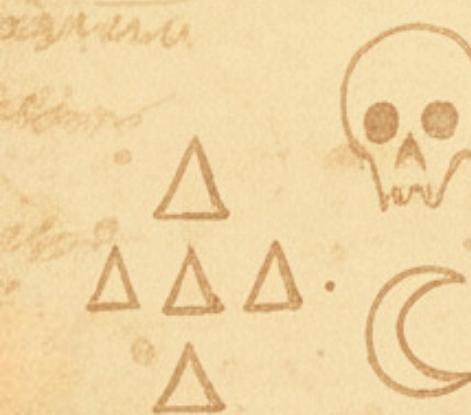
Quelle est la couleur du chat

X ~~je suis un chat et j'ai une couleur~~  
~~donc je suis un chat et j'ai une couleur~~



Property of  
le magicien quantique

prol. ouïefully  
tendurid pend of If traient proprely

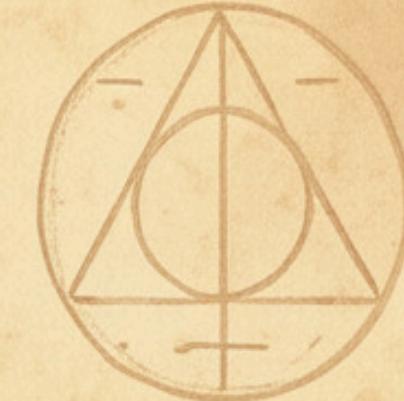


Valeur :  
- "couleur"

Canal mis à jour :

- Je suis un chat
- J'ai une couleur

→ Jeton suivant ?



Cle :  
Je suis un  
pronom  
interrogatif

Cle :  
Je suis un  
attribut

Requête :  
Informations  
sur Moi !

Quelle est la couleur du chat

Property of  
le magicien quantique

prob. ou nulles.  
tendrait pas à faire progresser



Valeur :  
- "couleur"

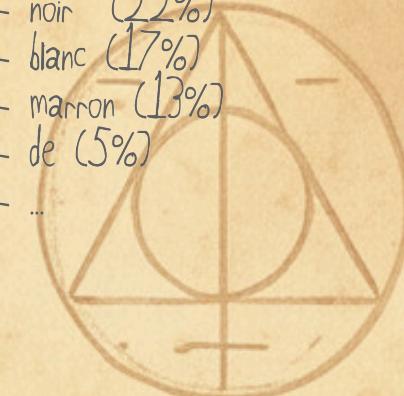
Canal mis à jour :

- Je suis un chat
- J'ai une couleur

X ~~je suis une entité de la nature~~  
~~je suis une entité de la nature~~  
~~je suis une entité de la nature~~  
~~je suis une entité de la nature~~

→ Jeton suivant ? →

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



Cle :  
Je suis un  
pronom  
interrogatif

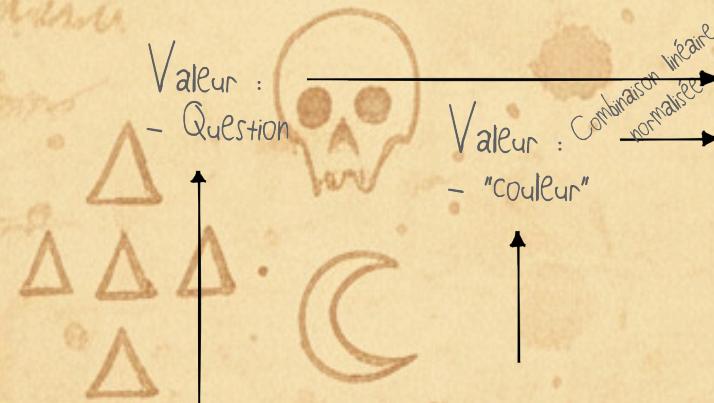
Cle :  
Je suis un  
attribut

Requête :  
Informations  
sur Moi !

Quelle est la couleur du chat

Property of  
le magicien quantique

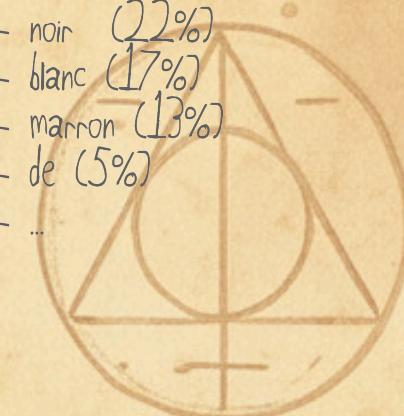
prob. ou nulles.  
tendrait pas à faire progresser



Canal mis à jour :

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

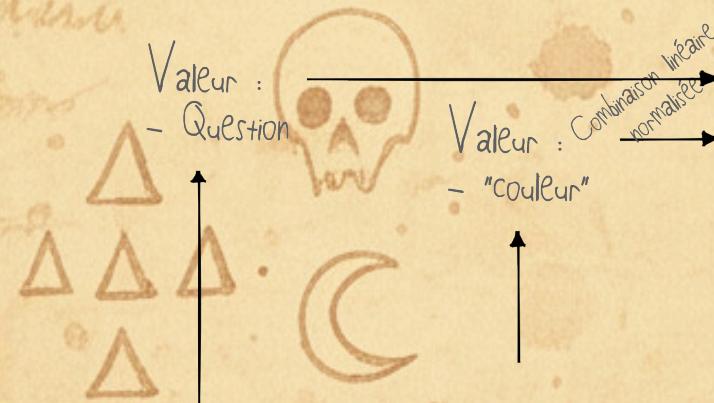
- noir (22%)
- blanc (17%)
- marron (13%)
- de (5%)
- ...



Quelle est la couleur du chat

Property of  
le magicien quantique

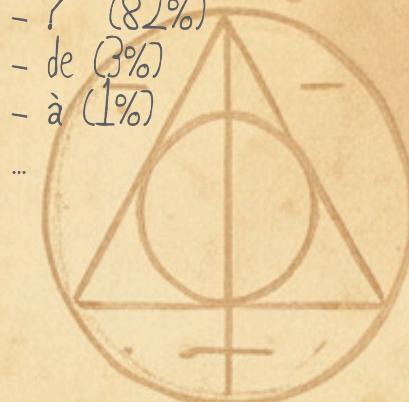
prob. ou nulles.  
tendrait pas à faire progresser



Canal mis à jour :

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

- ? (82%)
- de (7%)
- à (1%)



Clé : Je suis un pronom interrogatif

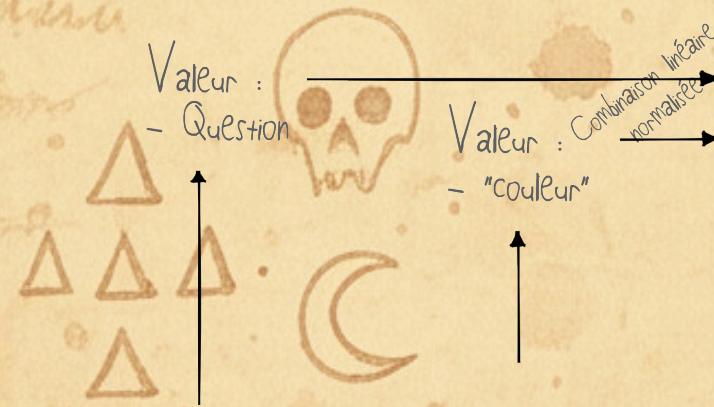
Clé : Je suis un attribut

Requête : Informations sur Moi !

Quelle est la couleur du chat

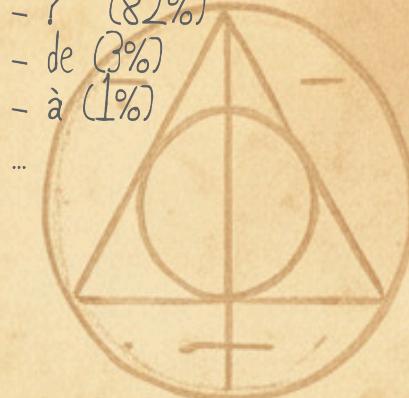
Property of  
le magicien quantique

prob. ou nulles.  
tendrait pas à faire progresser



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

- ? (82%)
- de (0%)
- à (1%)



Clé : Je suis un pronom interrogatif  
Clé : Je suis un attribut  
Requête : Informations sur Moi !

Quel est la couleur du chat

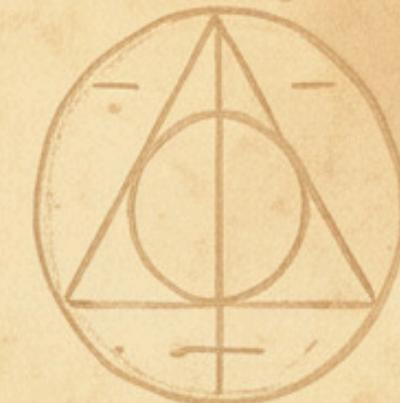
Property of  
le magicien quantique

prol. unifally  
tendured and if traind properly



Quelle est la couleur du chat de Hermione Granger ?

X environtales  
assez concretes  
peut devoir



Property of  
le magicien quantique

prol. unfully  
tundured and of if trainst proprly



X esunireteadles  
just asnes comtorciatine  
dovrde pte deuet



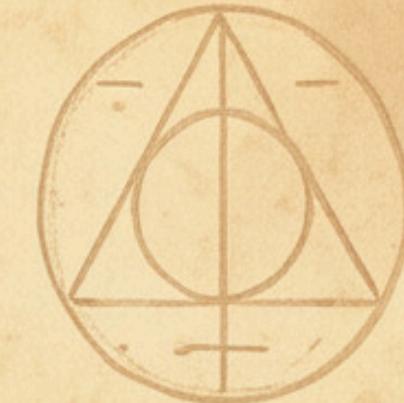
Hermione Granger

Property of  
le magicien quantique

prol. unifally  
tendured and of If traind proponly



X esunireteadles  
just asnes comatorciatue  
donvle pste deuet



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

Hermione Granger

Property of  
le magicien quantique

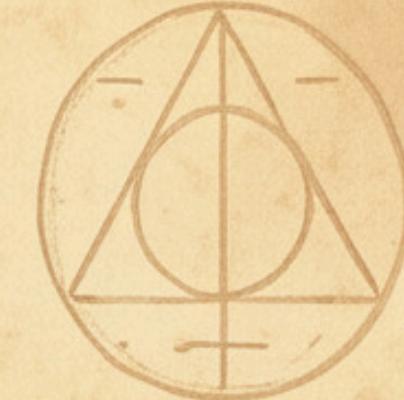
## Espace du perceptron (mémoire)

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 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



## Espace du perceptron (mémoire)

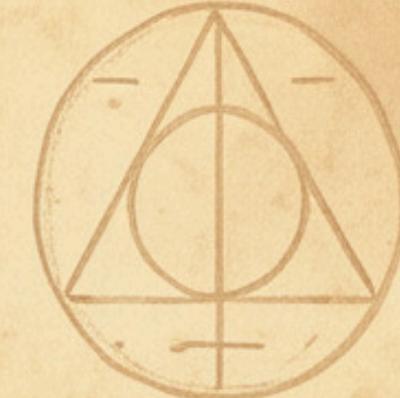
Hermione Granger :

- 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



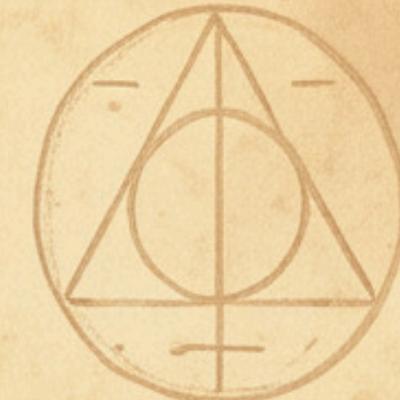
Property of  
le magicien quantique

prol. unifally  
tendured and if traind properly



Quelle est la couleur du chat de Hermione Granger ?

X envireretables  
assez comtoctisive  
peudeuet



Property of  
le magicien quantique

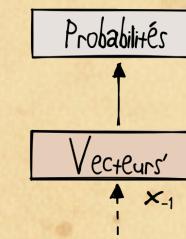
prol. oufally,  
tenduret pend of Il traient proprely



Quelle est la couleur du chat de Hermione Granger ?

X envirretemles  
just asnes comatorciatue  
devoir le pte deuet

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



Property of  
le magicien quantique

prol. uniformly  
standard and if trained properly



Tp

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

X  
esquiveteantes  
asnes comorciante  
pote devoet



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/>)