

Principes généraux d'intelligence artificielle & Expériences de traduction texte-image-texte avec GPT-4 et DALL•E 3

Adrien Delière

adrien.deliege@uliege.be

Postdoc ULiège (math, computer vision)

Séminaire international de sémiotique à Paris 2023-2024

14 février 2024

Principes généraux d'intelligence artificielle
&
Expériences de traduction texte-image-texte
avec GPT-4 et DALL•E 3

Adrien Delière

adrien.deliege@uliege.be

Postdoc ULiège (math, computer vision)

Séminaire international de sémiotique à Paris 2023-2024

14 février 2024

Principes généraux d'intelligence artificielle & Expériences de traduction texte-image-texte avec GPT-4 et DALL•E 3

Adrien Delière

adrien.deliege@uliege.be

Postdoc ULiège (math, computer vision)

Séminaire international de sémiotique à Paris 2023-2024

14 février 2024

Principes généraux d'intelligence artificielle & Expériences de traduction texte-image-texte avec GPT-4 et DALL•E 3

Adrien Delière

adrien.deliege@uliege.be

Postdoc ULiège (math, computer vision)

Séminaire international de sémiotique à Paris 2023-2024

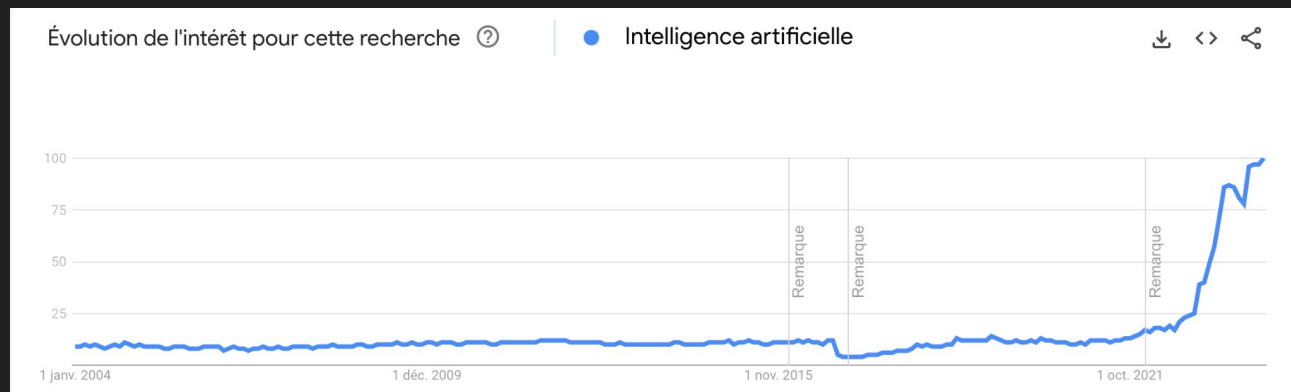
14 février 2024

Principes généraux d'intelligence artificielle

L'IA, c'est quoi ?

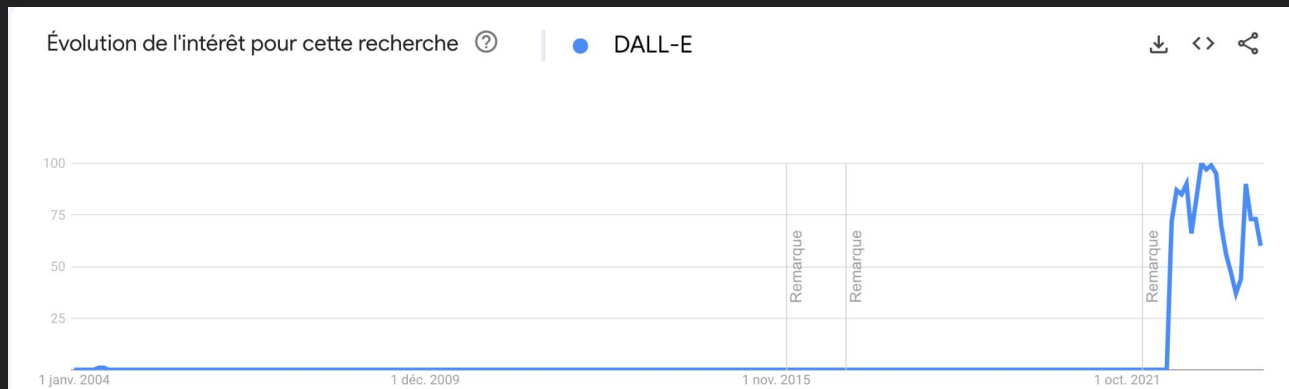
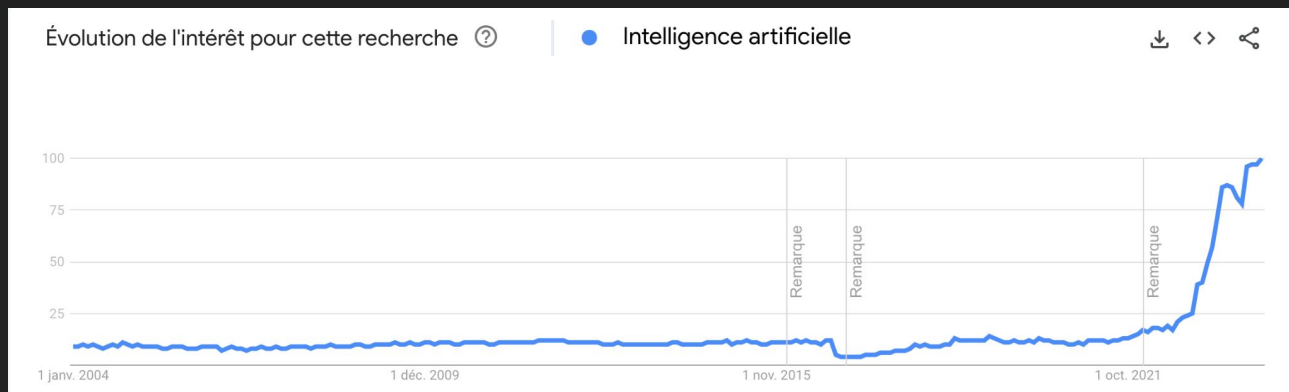
L'explosion de l'IA

Sources : Google Trends



L'explosion de l'IA

Sources : Google Trends



Courbes similaires pour:

- Midjourney
- Stable Diffusion
- Modèle génératif

Une définition de l'IA

IA = un système informatique qui effectue des tâches traditionnellement associées à l'intelligence humaine

Larousse : ensemble de théories et de techniques mises en oeuvre en vue de réaliser des machines capables de simuler l'intelligence humaine

Robert : ensemble des théories et des techniques développant des programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine (raisonnement, apprentissage...)

Une définition de l'IA

IA = un système informatique qui effectue des tâches traditionnellement associées à l'intelligence humaine

Indépendamment :

- des méthodes, algorithmes, opérations utilisés par l'IA

Une définition de l'IA

IA = un système informatique qui effectue des tâches traditionnellement associées à l'intelligence humaine

Indépendamment :

- des méthodes, algorithmes, opérations utilisés par l'IA
- du processus utilisé pour créer l'IA

Une définition de l'IA

IA = un système informatique qui effectue des tâches traditionnellement associées à l'intelligence humaine

Indépendamment :

- des méthodes, algorithmes, opérations utilisés par l'IA
- du processus utilisé pour créer l'IA
- des performances de l'IA

Une définition de l'IA

IA = un système informatique qui effectue des tâches traditionnellement associées à l'intelligence humaine

Indépendamment :

- des méthodes, algorithmes, opérations utilisés par l'IA
- du processus utilisé pour créer l'IA
- des performances de l'IA

Focus sur la nature de la tâche effectuée et l'intelligence humaine requise

L'IA, une réelle nouveauté ?



L'IA, une réelle nouveauté ?

- Des IA existent depuis longtemps
- Abus d'utilisation du terme "IA"
- Récente résurgence via la popularité de nouvelles tâches accessibles par IA

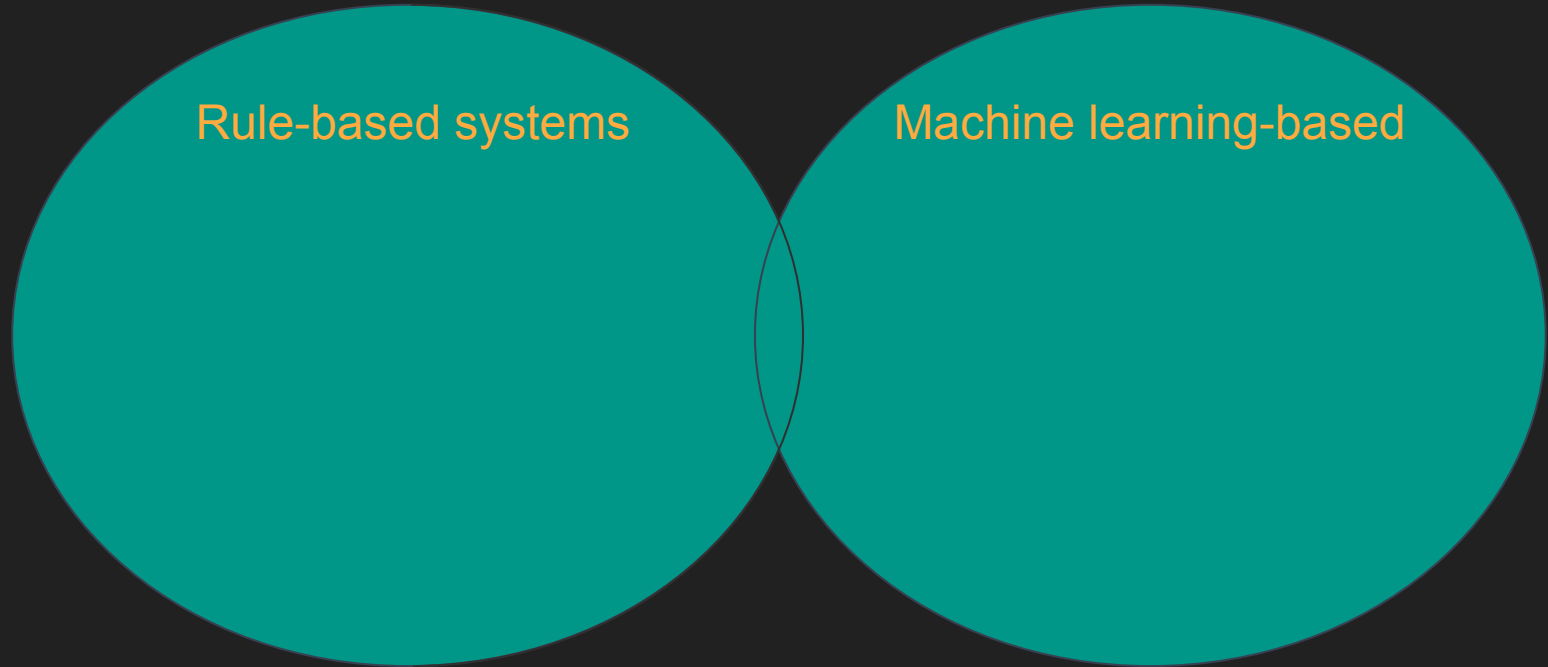
"grand public"



Principes généraux d'intelligence artificielle

Comment on crée une IA ?

Rule-based vs Machine learning



Rule-based vs Machine learning

Rule-based systems

Un expert décide de toutes les opérations à effectuer

Machine learning-based

Le système détermine lui-même les opérations

Rule-based vs Machine learning

Rule-based systems

Un expert décide de toutes les opérations à effectuer

L'expert encode ses connaissances

Machine learning-based

Le système détermine lui-même les opérations

Le système apprend via les données disponibles, essai-erreur et feedback

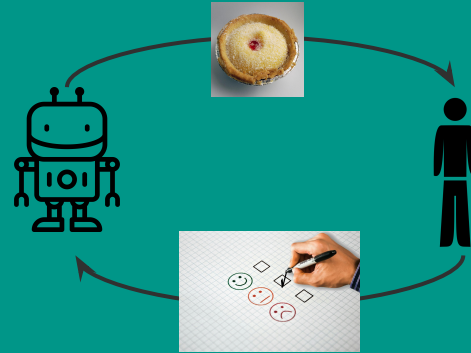
Rule-based vs Machine learning

Rule-based systems



Recette du chef

Machine learning-based



Recette de l'IA

Rule-based vs Machine learning

Rule-based systems

- Contrôle des opérations
- Experts trop peu nombreux
- Experts pas assez forts...
- ... même pour tâches "simples"

Machine learning-based

- ~~Contrôle des opérations~~
- Acquérir données : "facile"
- Essai-erreur : "facile"
- Donner feedback : "facile"

Rule-based vs Machine learning

ML : Techniques pour apprendre des modèles (“IA”) à partir de données

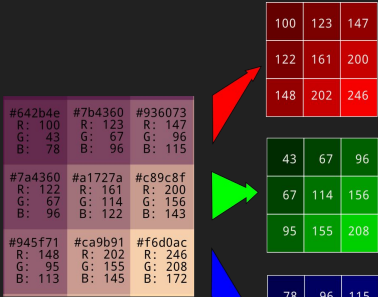
Rule-based systems

- Contrôle des opérations
- Experts trop peu nombreux
- Experts pas assez forts...
- ... même pour tâches “simples”

Machine learning-based

- ~~Contrôle des opérations~~
- Acquérir données : “facile”
- Essai-erreur : “facile”
- Donner feedback : “facile”

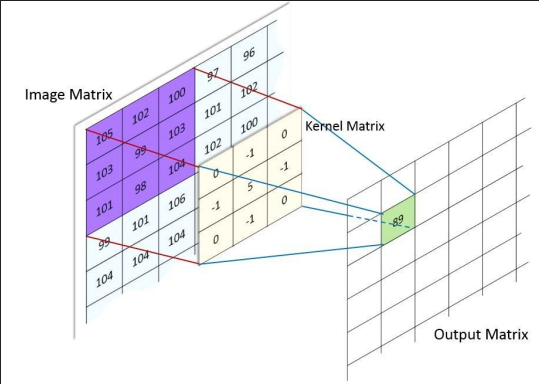
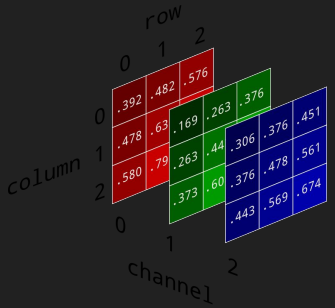
Entrainement d'un modèle par ML : Rappel : Image = Pixels



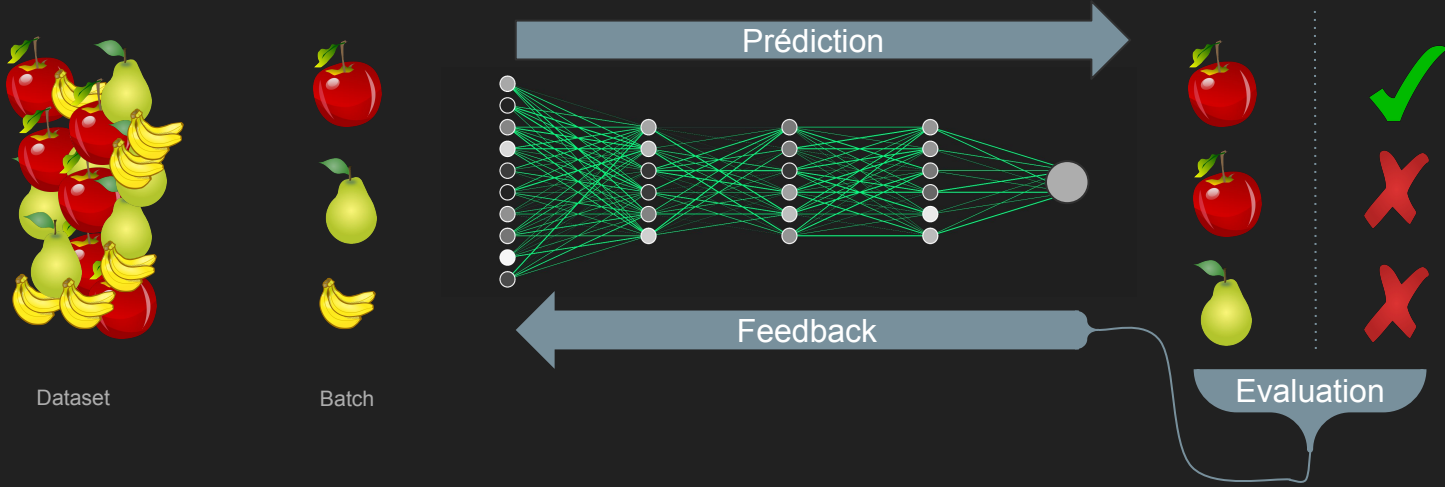
100	123	147
122	161	200
148	202	246

43	67	96
67	114	156
95	155	208

78	96	115
96	122	143
113	145	172



Entrainement d'un modèle par ML

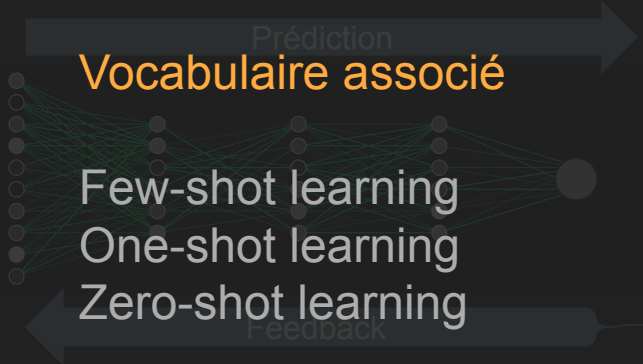


Données

Modèle

Evaluation

Entrainement d'un modèle par ML



Quantité

Provenance

Données

Modèle

Entrainement d'un modèle par ML

Vocabulaire associé

Vision model
Language model
Multitask model

Tâche

Neural network
Transformer

Architecture

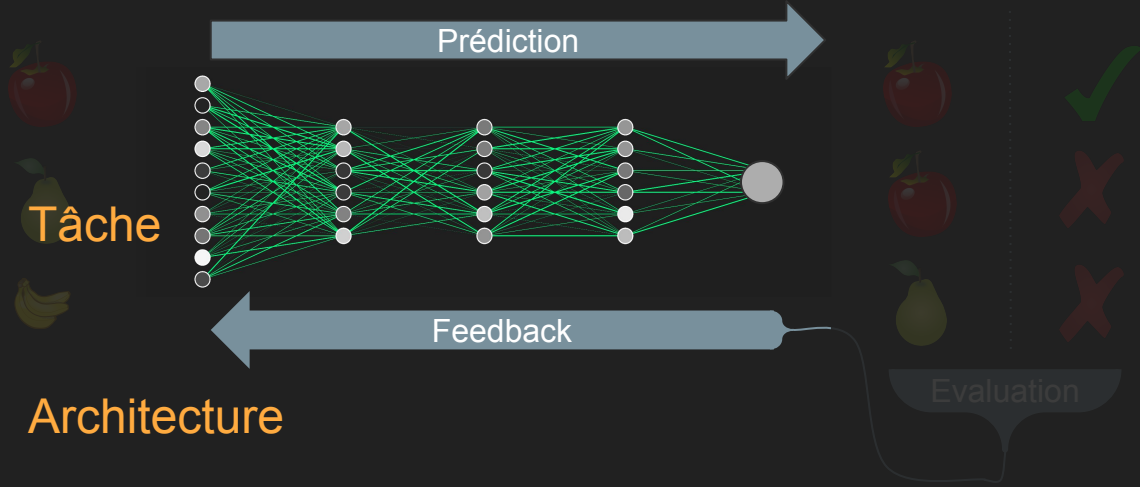
Deep learning

Large model
Foundation model

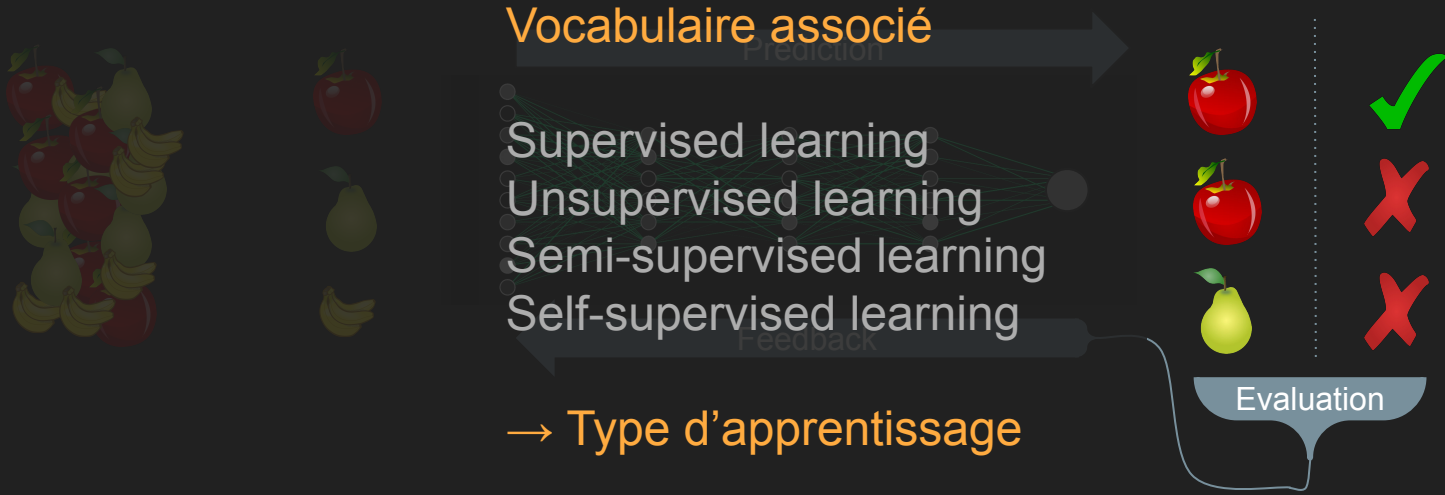
Taille

Modèle

Evaluation



Entrainement d'un modèle par ML

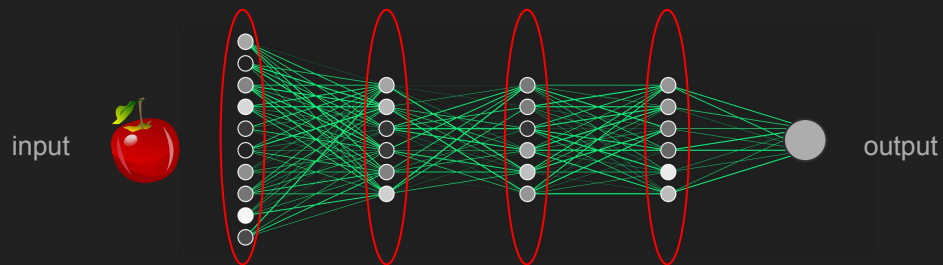


Données

Modèle

Evaluation

Poids d'un modèle et embeddings



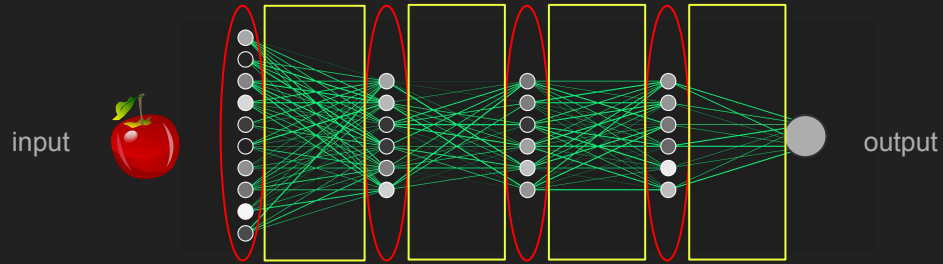
Représentations numériques de l'input

Varie d'un input à l'autre

"Cases vides" tant qu'on ne soumet pas d'input

Dernière représentation : souvent appelée "embedding"

Poids d'un modèle et embeddings



Représentations numériques de l'input

Varie d'un input à l'autre

"Cases vides" tant qu'on ne soumet pas d'input

Dernière représentation : souvent appelée "embedding"

Poids/paramètres du modèle

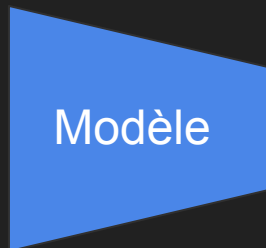
Appliqués de la même façon à tous les inputs

Déterminent les "règles de calcul" du modèle

Ajustés grâce au feedback durant l'apprentissage

Représentation schématique

Après entraînement



Liste de poids
+ architecture

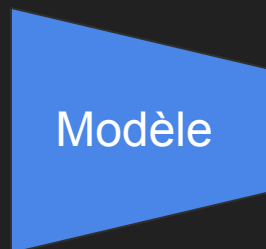
Les connaissances du modèle sont accumulées
dans ses poids

Le modèle a compressé le dataset dans ses poids

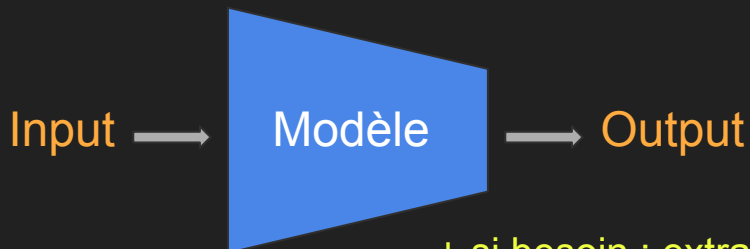


Représentation schématique

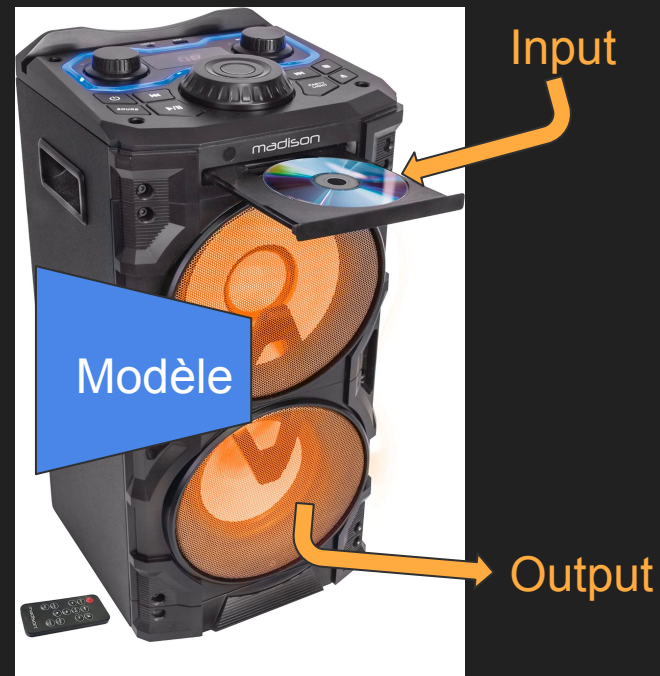
Après entraînement



Utilisation (inférence)



+ si besoin : extraction d'un embedding de l'input



Principes généraux d'intelligence artificielle

Le cas des modèles génératifs texte-image

Avant-propos

Les modèles génératifs texte-image :

- Basés sur **datasets énormes** de textes et d'images
- Contiennent plusieurs **milliards de poids**
- Nécessitent **puissance de calcul** colossale
- Nécessitent **milliers d'heures** d'entraînement

Avant-propos

Les modèles génératifs texte-image :

- Basés sur **datasets énormes de texte et d'images**
- Contiennent plusieurs milliards de poids
- Nécessitent puissance de calcul colossale
- Nécessitent milliers d'heures d'entraînement



Représentent une
compression d'une grande
partie de la connaissance
humaine

Avant-propos

Les modèles génératifs texte-image :

- Basés sur datasets énormes de texte et d'images
- Contiennent plusieurs milliards de poids
- Nécessitent puissance de calcul colossale
- Nécessitent milliers d'heures d'entraînement

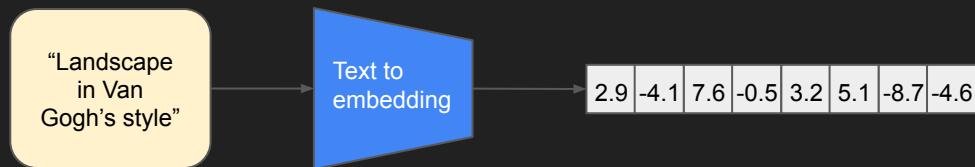
→ Développement dans une **poignée d'universités et entreprises** seulement

→ **Détails inconnus** mais **suppositions raisonnables** vu la littérature existante

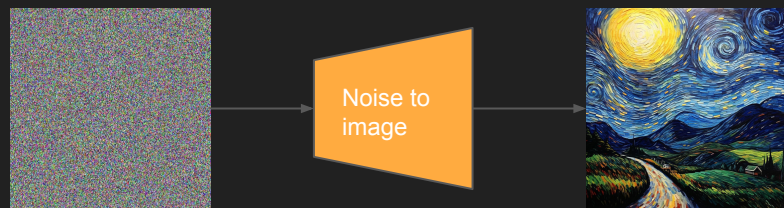
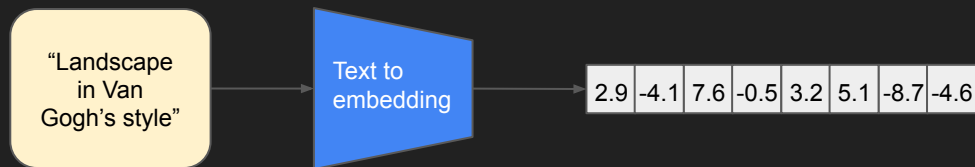


Représentent une
compression d'une grande
partie de la connaissance
humaine

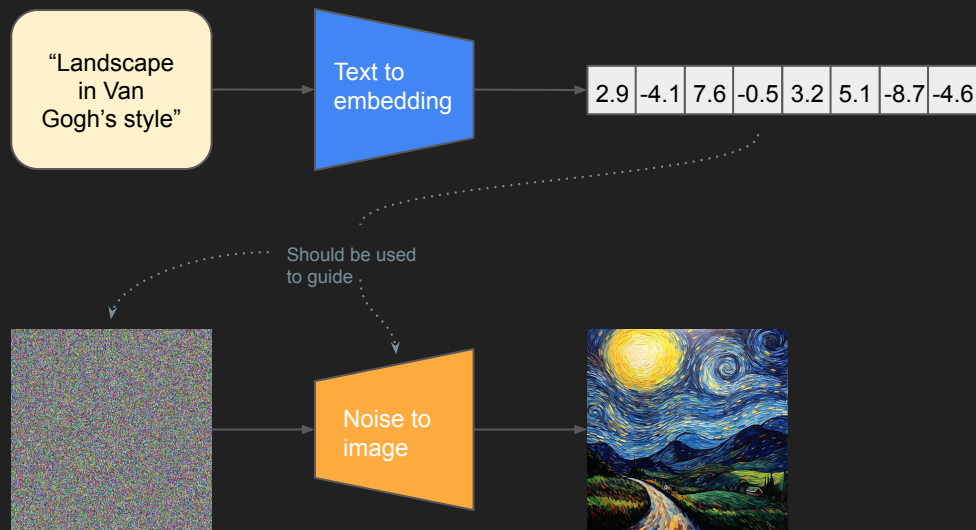
Une construction possible : éléments nécessaires



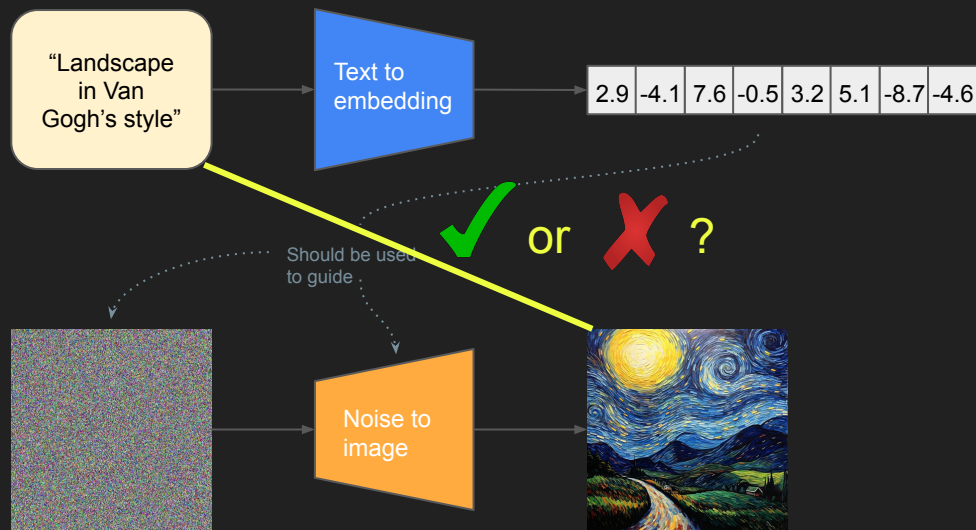
Une construction possible : éléments nécessaires



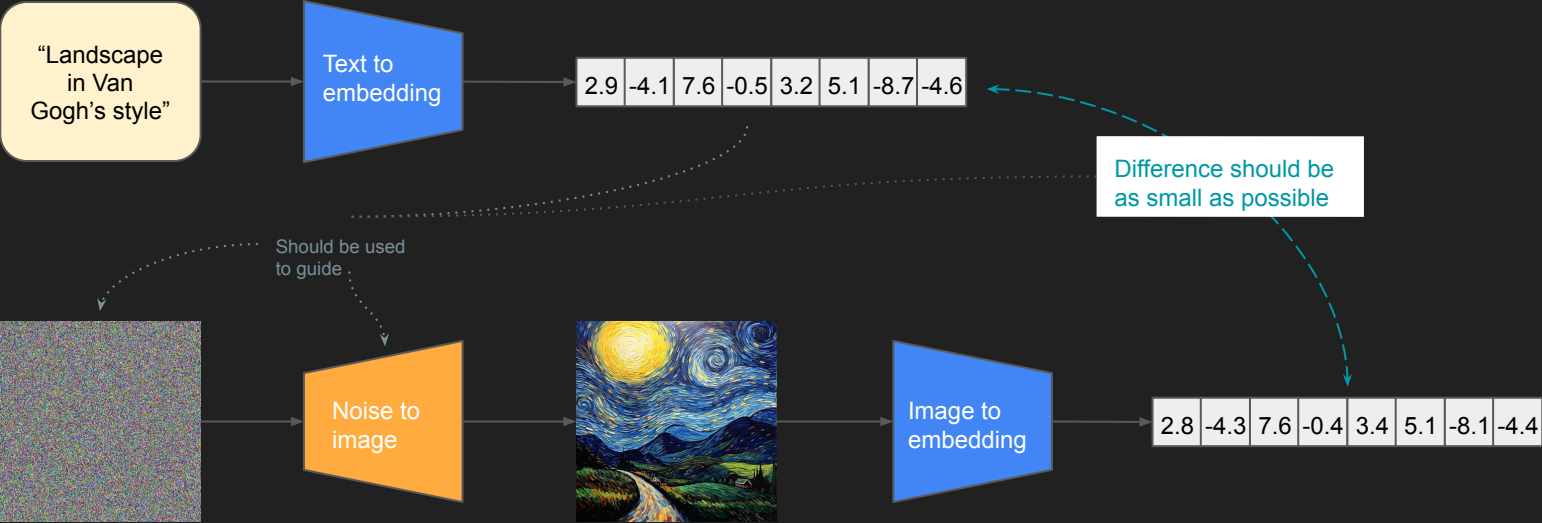
Une construction possible : éléments nécessaires



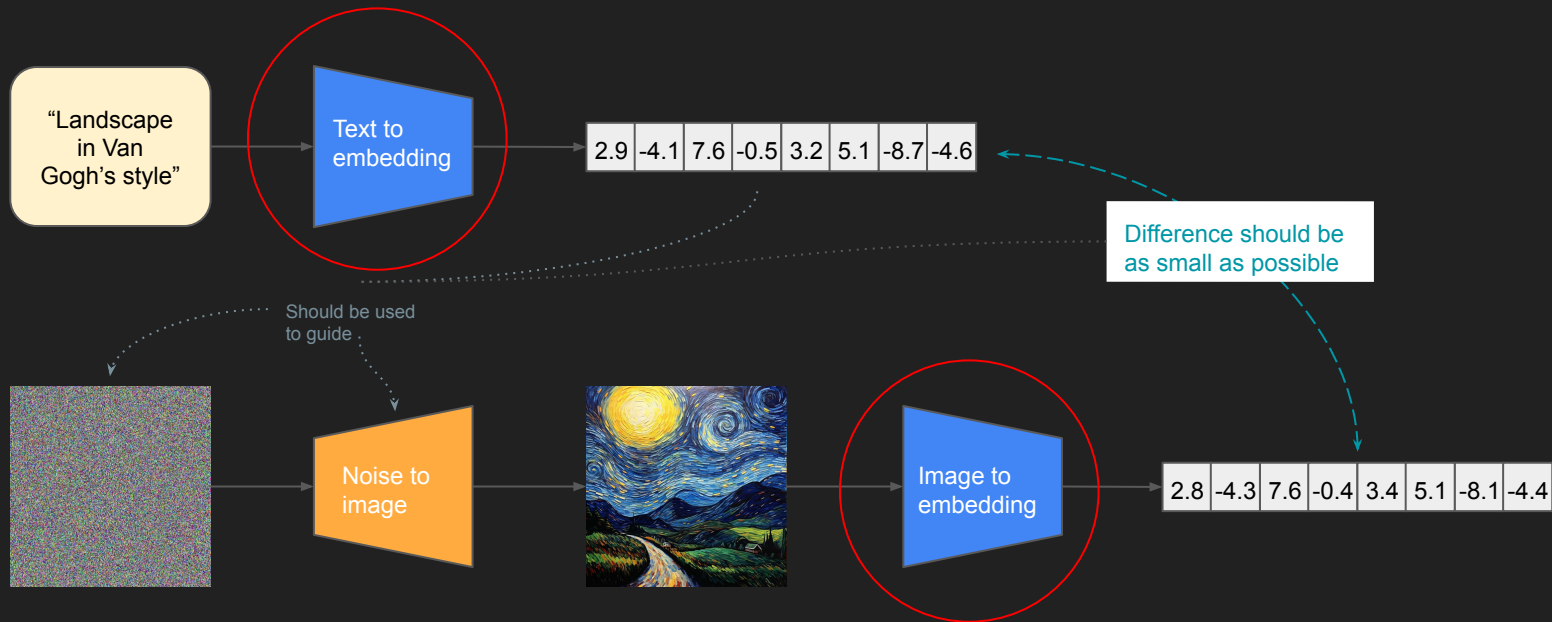
Une construction possible : éléments nécessaires



Une construction possible : éléments nécessaires



Une construction possible : éléments nécessaires



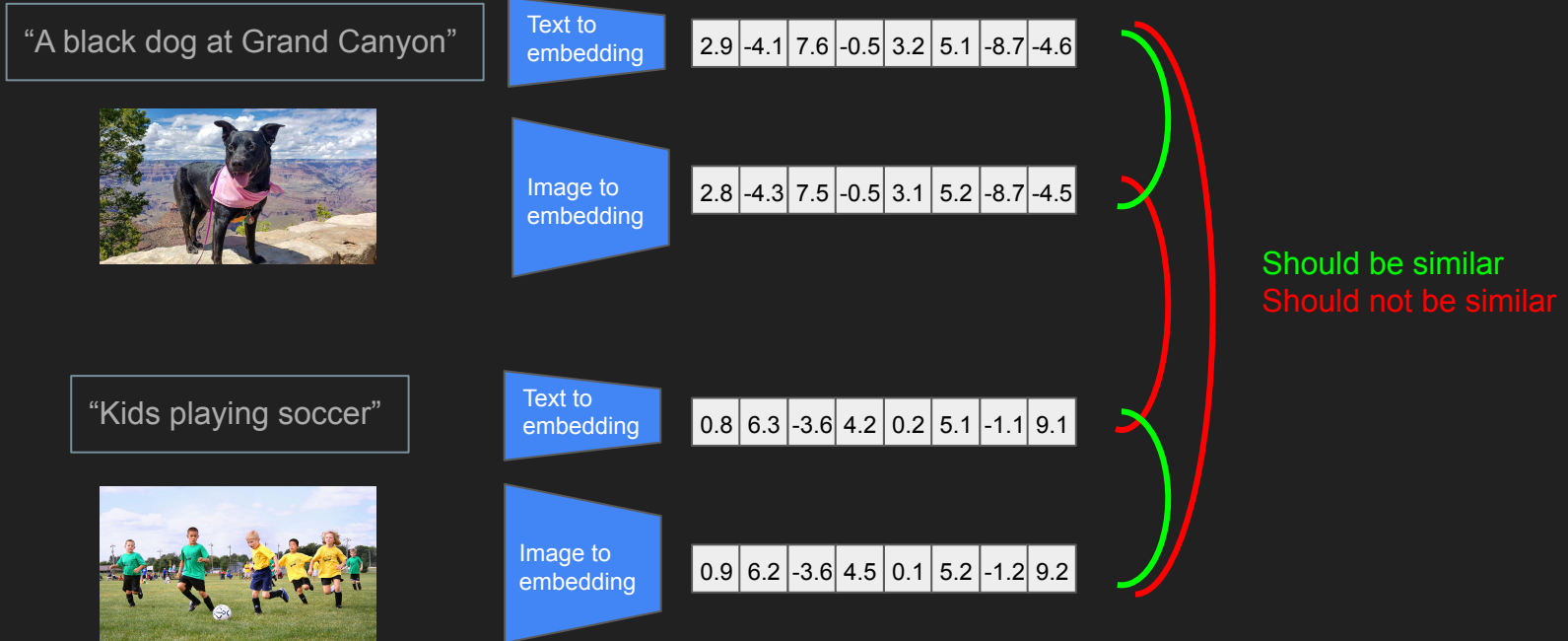
Doit avoir été entraîné sur des paires texte-images pour que la comparaison des embeddings ait du sens ! Ex. : CLIP

Une construction possible : étapes

1. Collecter des paires textes-images

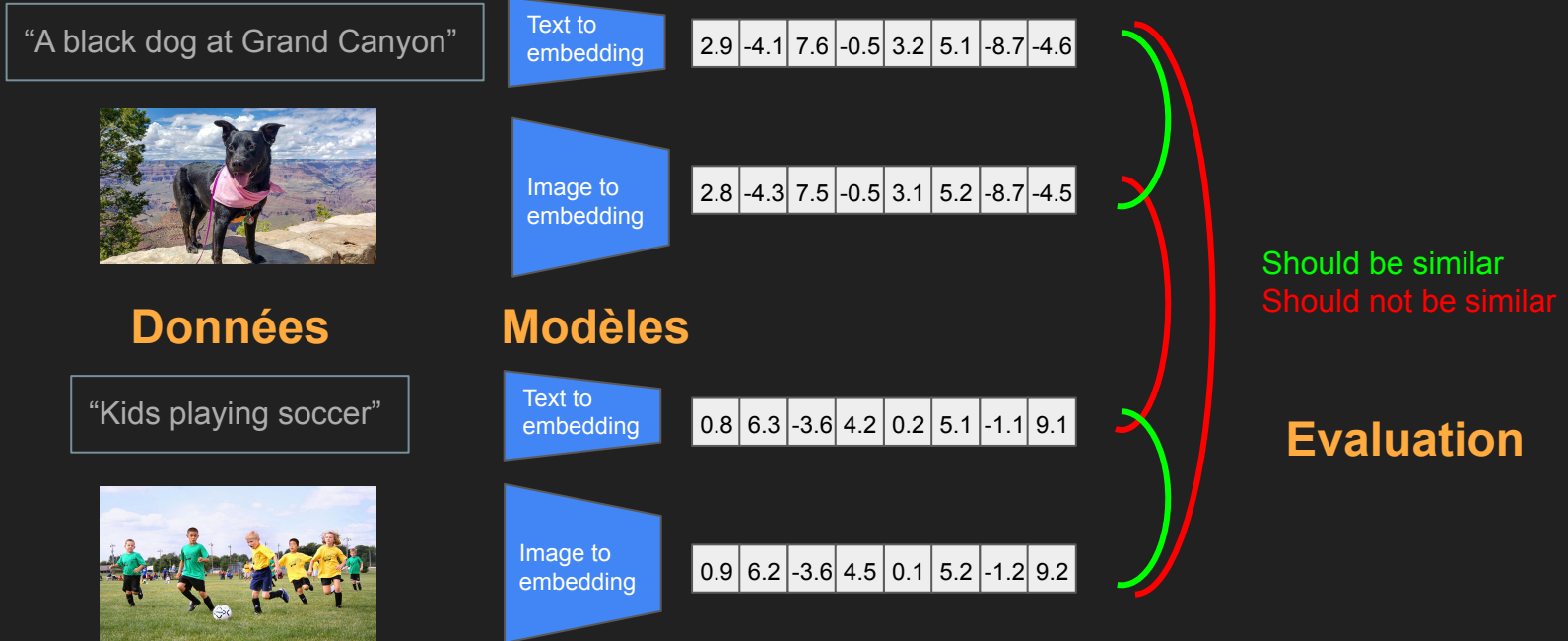
Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP



Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP



Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP

CLIP :

- Peut recevoir du texte
- Peut recevoir des images
- Entraîné pour produire des embeddings cohérents texte-image

"A black dog at Grand Canyon"



Données

"Kids playing soccer"



Text to embedding

2.9	-4.1	7.6	-0.5	3.2	5.1	-8.7	-4.6
-----	------	-----	------	-----	-----	------	------

Image to embedding

2.8	-4.3	7.5	-0.5	3.1	5.2	-8.7	-4.5
-----	------	-----	------	-----	-----	------	------

Modèles

Text to embedding

0.8	6.3	-3.6	4.2	0.2	5.1	-1.1	9.1
-----	-----	------	-----	-----	-----	------	-----

Image to embedding

0.9	6.2	-3.6	4.5	0.1	5.2	-1.2	9.2
-----	-----	------	-----	-----	-----	------	-----



Should be similar
Should not be similar

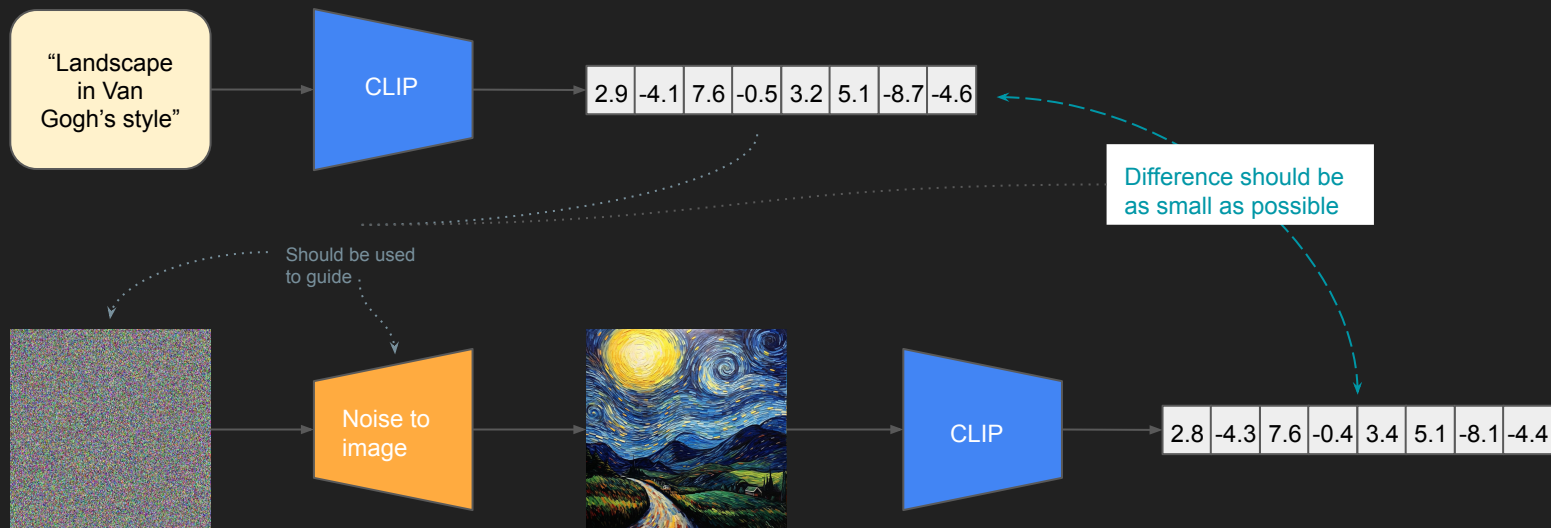
Evaluation

Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP
3. **Entraîner un modèle générateur d'images**

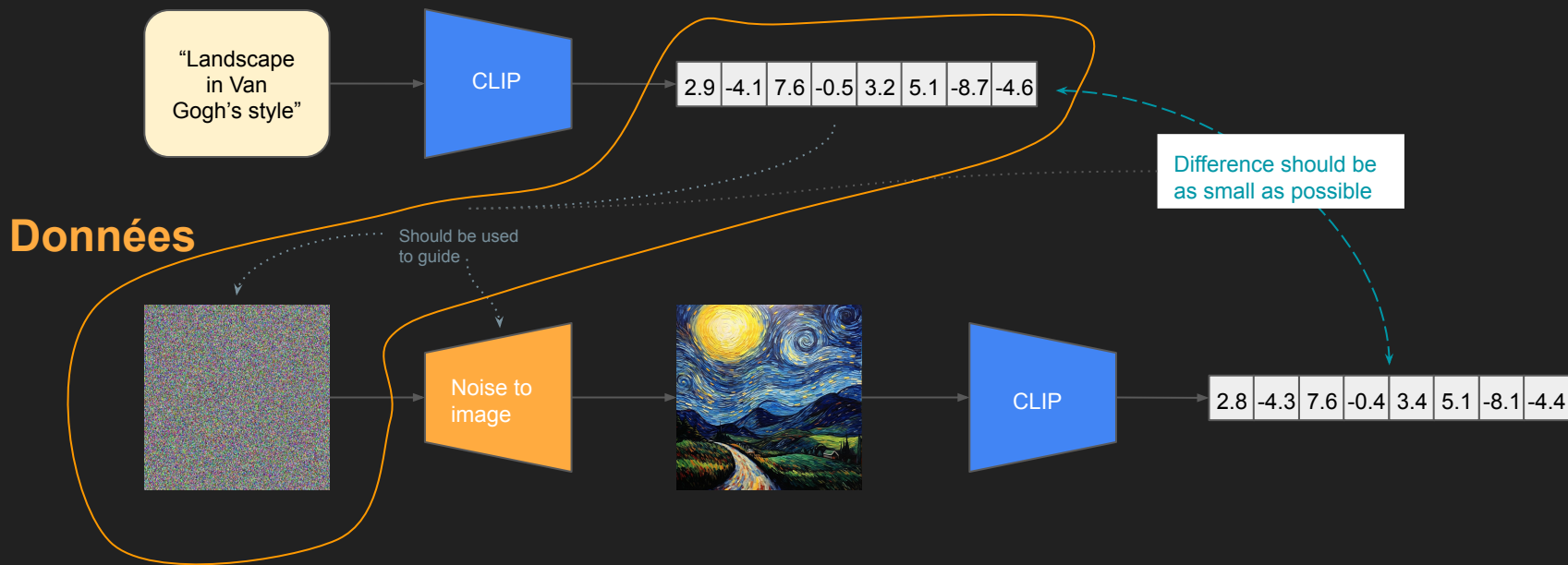
Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP
3. **Entraîner un modèle générateur d'images**



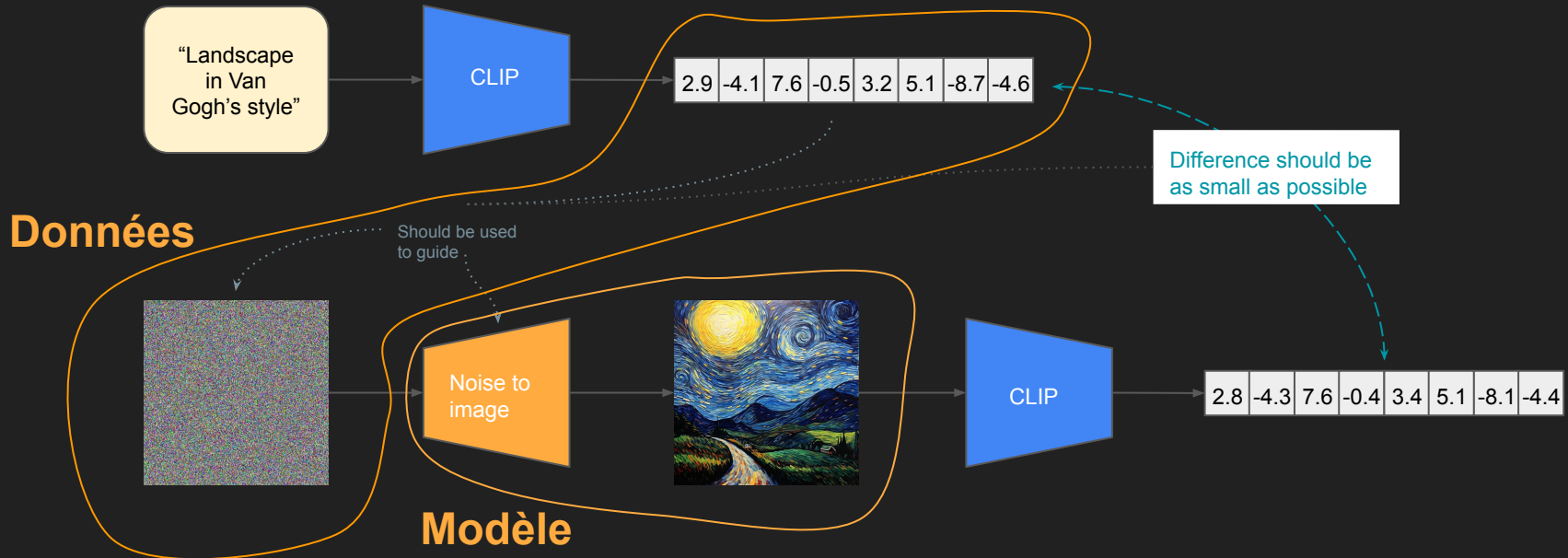
Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP
3. **Entraîner un modèle générateur d'images**



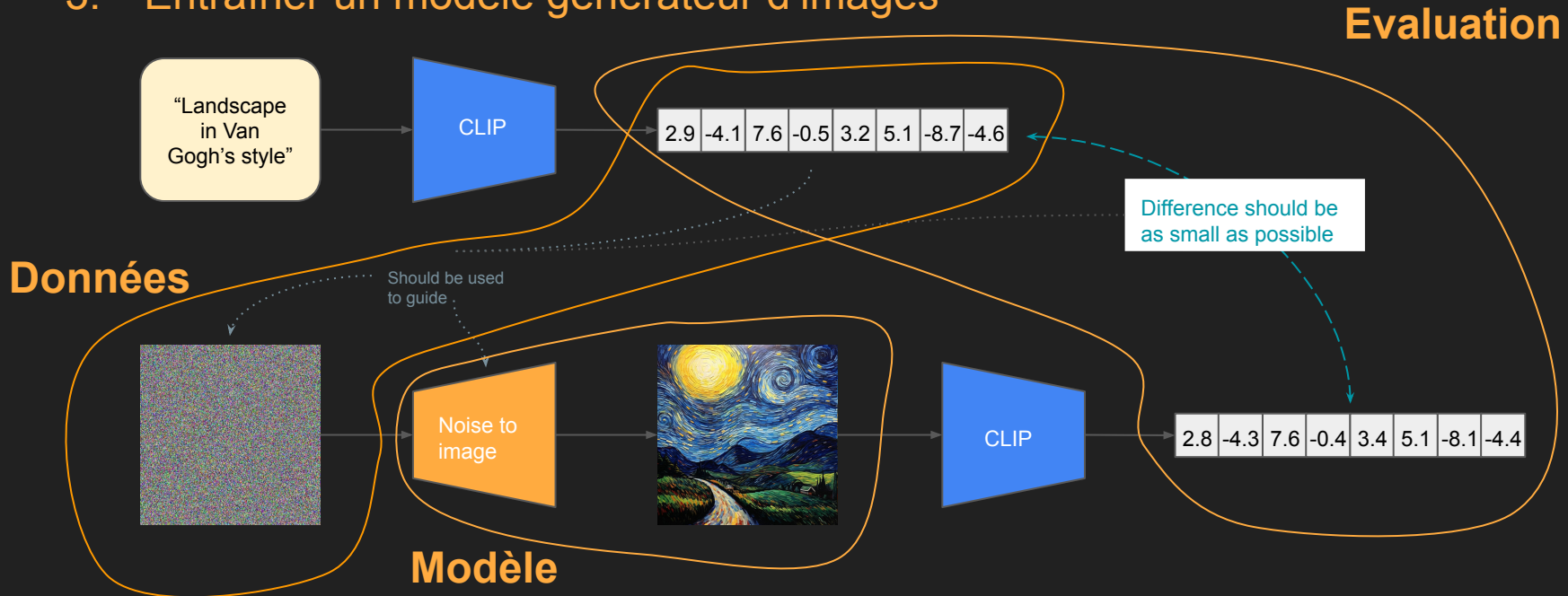
Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP
3. **Entraîner un modèle générateur d'images**

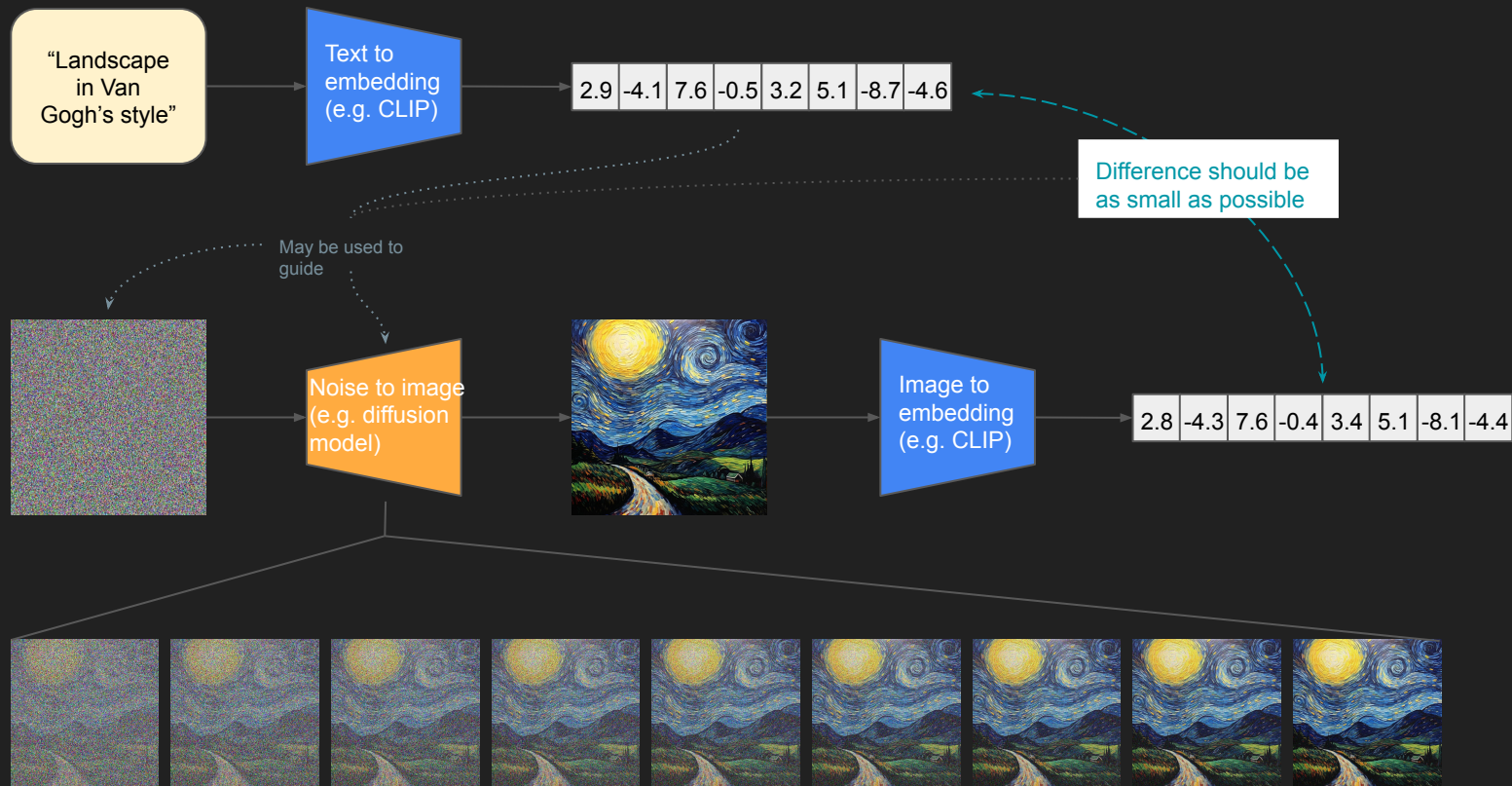


Une construction possible : étapes

1. Collecter des paires textes-images
2. Entraîner un modèle tel que CLIP
3. **Entraîner un modèle générateur d'images**



Noise to image : par ex. modèle de diffusion

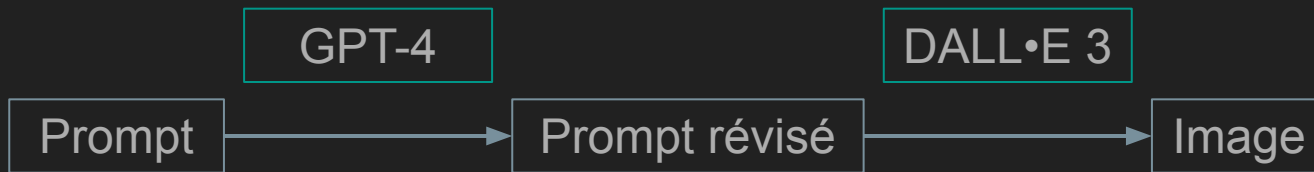
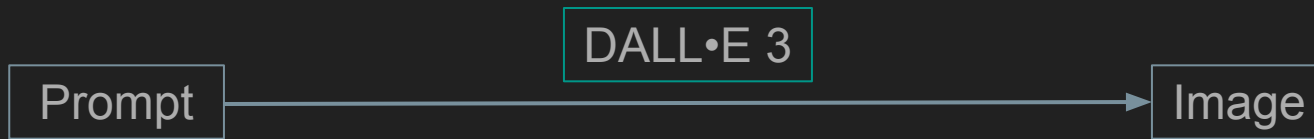


Expériences de traduction texte-image-texte avec GPT-4 et DALL·E 3

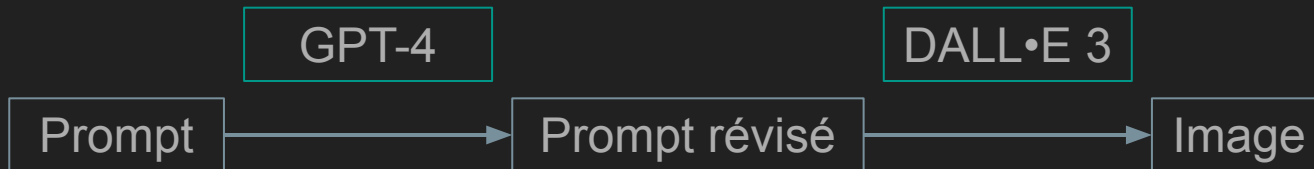
Part 1 : Texte → Image



Particularité d'OpenAI



Particularité d'OpenAI

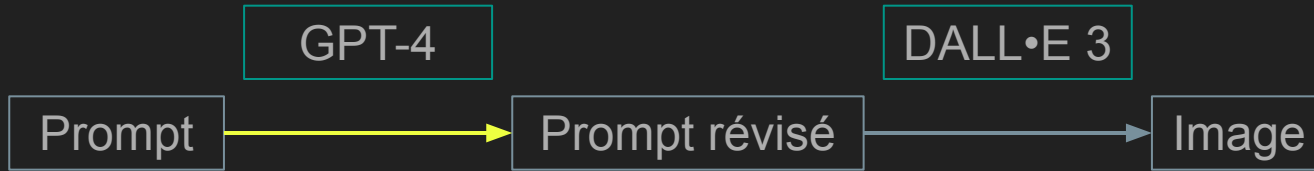


“for safety reasons, and to add more detail (more detailed prompts generally result in higher quality images)”

Empêche la génération de

- contenu violent, pour adultes, haineux
- biais blessants de sous/sur-représentation d'ethnicités
- figures publiques
- images dans le style d'un artiste vivant

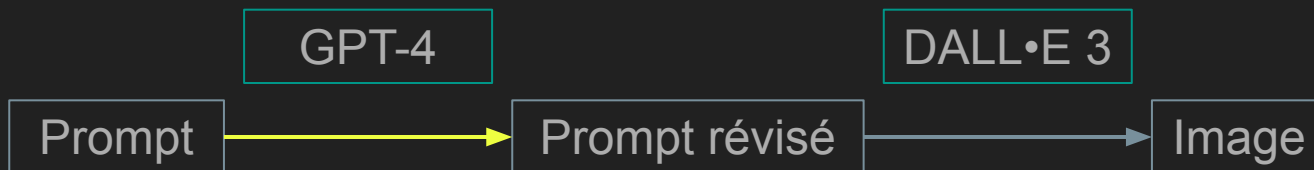
Expérience 1 : culture générale de l'IA ?



"A painting of Susanna and the Elders"

"A classical painting showcasing an old religious story from the Bible. Depict a scene with three individuals - a young Middle-Eastern woman named Susanna, dressed in beautiful, richly-colored fabrics, and two elderly men, one of them of Caucasian descent and the other of South Asian descent, looking intensely at her. The scene takes place in a lush garden, with hints of Greek architecture in the surroundings."

Expérience 1 : culture générale de l'IA ?



“A painting of Susanna and the Elders”

“A classical painting showcasing an **old religious story from the Bible**. Depict a scene with **three individuals - a young Middle-Eastern woman named Susanna**, dressed in beautiful, richly-colored fabrics, and **two elderly men**, one of them of **Caucasian** descent and the other of **South Asian** descent, **looking intensely at her**. The scene takes place in a **lush garden**, with hints of Greek architecture in the surroundings.”



GPT-4 connaît le thème

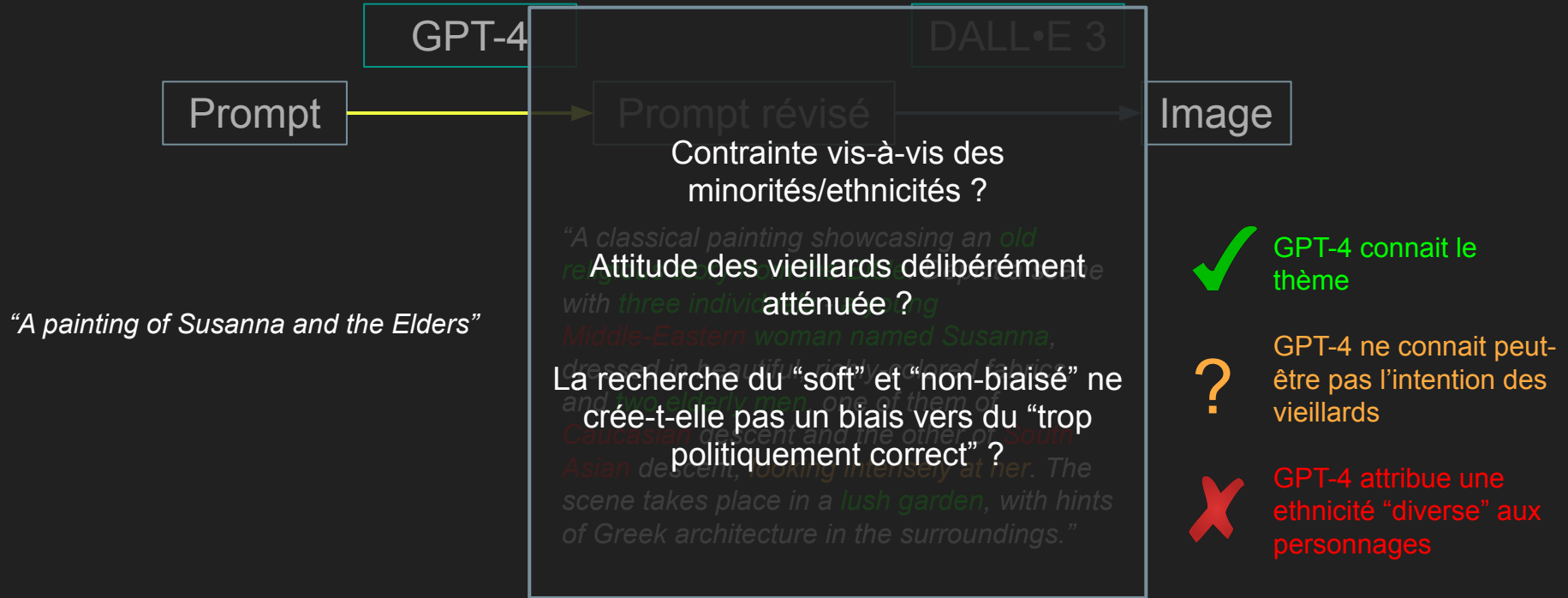


GPT-4 ne connaît peut-être pas l'intention des vieillards

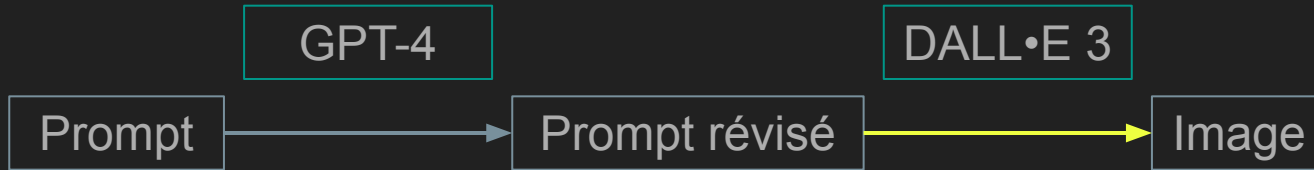


GPT-4 attribue une ethnicité “diverse” aux personnages

Expérience 1 : culture générale de l'IA ?



Expérience 1 : culture générale de l'IA ?



“A painting of Susanna and the Elders”

“A classical painting showcasing an old religious story from the Bible. Depict a scene with three individuals - a young Middle-Eastern woman named Susanna, dressed in beautiful, richly-colored fabrics, and two elderly men, one of them of Caucasian descent and the other of South Asian descent, looking intensely at her. The scene takes place in a lush garden, with hints of Greek architecture in the surroundings.”



Expérience 1 : culture générale de l'IA ?

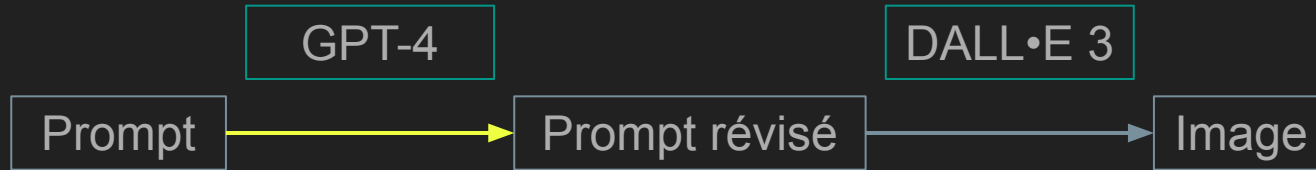
*“A classical painting showcasing an old religious story from the Bible. Depict a scene with **three** individuals - a young **Middle-Eastern** woman named Susanna, dressed in beautiful, richly-colored fabrics, and two elderly men, one of them of Caucasian descent and the other of **South Asian** descent, looking intensely at her. The scene takes place in a lush garden, with hints of Greek architecture in the surroundings.”*

Difficultés pour représenter

- un nombre d'instances donné
- des ethnicités
- des mains



Expérience 2 : prompt de Wikipédia (Tintoretto)

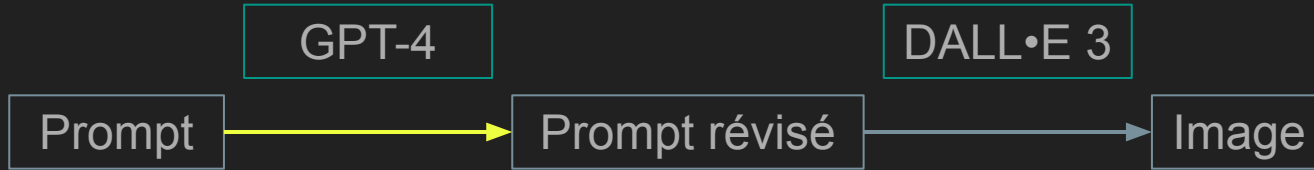


Buts :

- + de contrôle sur la composition
- + d'insistance sur les éléments importants, e.g. l'attitude intrusive des vieillards



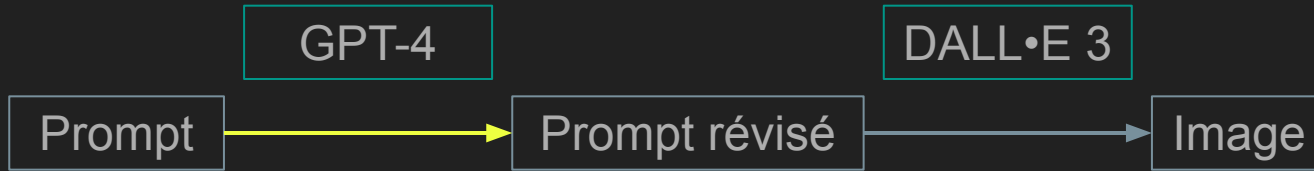
Expérience 2 : prompt de Wikipédia (Tintoretto)



~~"The painting shows a naked Susanna A painting showing a woman sitting in a garden beside a pool, leaning against a tree, and facing a dense rose trellis, against which leans a mirror. Susanna She gazes into the mirror. Around her on the ground are several bright objects including a white silk shawl, pearls, jewelry, a comb, and silverware, which together form a kind of still life. Two elderly men peer around either end of the rose trellis to watch her. Susanna She is sitting absorbed by her reflection in the mirror and unaware of the intruders."~~

"A detailed painting of a white, middle-aged woman sitting by a calm pool in an enchanting garden. She is leaning against a robust tree while fixing her gaze into a mirror that rests against a lush rose trellis. In the foreground, there lies an assortment of vibrant objects such as a silken white shawl, gleaming pearls, shiny jewelry, a finely carved comb, and clean silverware, forming an aesthetically pleasing still life. Two elderly men of Middle-Eastern and Hispanic-ethnicity lurk around either end of the rose trellis, spying on the woman who remains engrossed in her mirror's reflection."

Expérience 2 : prompt de Wikipédia (Tintoretto)



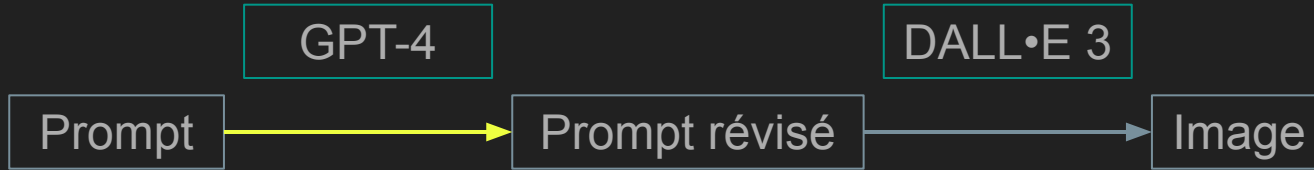
*~~"The painting shows a naked Susanna~~
A painting showing a woman sitting in a garden beside a pool, leaning against a tree, and facing a dense rose trellis, against which leans a mirror. ~~Susanna~~ She gazes into the mirror. Around her on the ground are several bright objects including a white silk shawl, pearls, jewelry, a comb, and silverware, which together form a kind of still life. Two elderly men peer around either end of the rose trellis to watch her. ~~Susanna~~ She is sitting absorbed by her reflection in the mirror and unaware of the intruders."*

*"A **detailed** painting of a **white, middle-aged** woman sitting by a **calm** pool in an **enchanting** garden. She is leaning against a **robust** tree while fixing her gaze into a mirror that rests against a **lush** rose trellis. In the foreground, there lies an assortment of **vibrant** objects such as a silken **white** shawl, **gleaming** pearls, **shiny** jewelry, a **finely carved** comb, and **clean** silverware, forming an **aesthetically pleasing** still life. Two elderly men of **Middle-Eastern** and **Hispanic** ethnicity lurk around either end of the rose trellis, spying on the woman who remains engrossed in her mirror's reflection."*

Ethnicité, traits physiques ajoutés

Ajout de nombreux adjectifs "enjolivants"

Expérience 2 : prompt de Wikipédia (Tintoretto)



*"The painting shows a naked Susanna
A painting showing a woman sitting in a garden beside a pool, leaning against a tree, and facing a dense rose trellis, against which leans a mirror. ~~Susanna~~
She gazes into the mirror. Around her on the ground are several bright objects including a white silk shawl, pearls, jewelry, a comb, and silverware, which together form a kind of still life. Two elderly men peer around either end of the rose trellis to watch her. ~~Susanna~~
She is sitting absorbed by her reflection in the mirror and unaware of the intruders."*

"A detailed painting of a white, middle-aged woman sitting by a calm pool in an enchanting garden. She is leaning against a robust tree while fixing her gaze into a mirror that rests against a lush rose trellis. In the foreground, there lies an assortment of vibrant objects such as a silken white shawl, gleaming pearls, shiny jewelry, a finely carved comb, and clean silverware, forming an aesthetically pleasing still life. Two elderly men of Middle-Eastern and Hispanic ethnicity lurk around either end of the rose trellis, spying on the woman who remains engrossed in her mirror's reflection."



Expérience 2 : prompt de Wikipédia (Tintoretto)

couvre-chef "Middle-Eastern" ?

"A detailed painting of a white, middle-aged woman sitting by a calm pool in an enchanting garden. She is **leaning against a robust tree** while fixing her gaze into a mirror that rests against a lush rose trellis. In the foreground, there lies an assortment of vibrant objects such as a silken white shawl, gleaming pearls, shiny jewelry, a finely carved comb, and clean silverware, forming an aesthetically pleasing still life. **Two elderly men of Middle-Eastern and Hispanic ethnicity** lurk around **either end of the rose trellis**, spying on the woman who remains engrossed in her mirror's reflection."

Difficultés pour représenter

- un nombre d'instances donné
- des ethnicités
- des visages lointains
- certains aspects compositionnels



Expérience 2 : prompt de Wikipédia (Tintoretto)

“Create a vivid painting illustrating a story. Capture a Caucasian woman in a serene garden, sitting alongside a tranquil pool. She leans comfortably against a tree and gazes intently into a mirror that leans against a dense rose trellis. The ground around her is littered with a multitude of bright objects: a white silk shawl, pearls, precious jewelry, an ornate comb, and sparkling silverware, creating an intricate still life. At either end of the rose trellis, two elderly men, one Black and the other East Asian, stealthily peer around, invading her solitude. The woman, absorbed in her reflection, remains blissfully unaware of the intruders.”



Expérience 2 : prompt de Wikipédia (Tintoretto)

*“A **19th-century style** painting featuring a Caucasian woman seated in a lush garden by a pool, resting comfortably against an ancient tree. She faces a thick rose trellis, where a mirror is perched. Her attention is captivated by her own reflection in the mirror. Bright items scattered around her on the rich, green earth create a still life scene. These items include a pristine white silk shawl, strands of pearls, assorted jewelry, a delicate comb, and gleaming silverware. At either end of the scarlet rose trellis, **two Middle-Eastern** men of advanced age carefully observe her, **their intentions unknown**. Their presence is unbeknownst to the woman deeply engrossed in her reflection.”*



Expérience 3 : variations compositionnelles



Prompt de Wikipédia rend de meilleures représentations



Simple mémorisation/identification de DALL•E 3, ou réelle capacité de création ?

→ Créons des variations

Expérience 3 : variations compositionnelles : genres, ages

Susanna → man, Elders → old women



Susanna → old woman, Elders → young men



Expérience 3 : variations compositionnelles : thème, lieu

Thème de la mort



À la plage



À propos du style

“Trop lisse”, “faux”, “hyperréaliste américain des années 70”

À aucun moment nous n'avons imposé de style. Si on insiste sur Renaissance/Baroque : pas de changement... Note : Midjourney est meilleur pour le style



Expériences de traduction texte-image-texte avec GPT-4V et DALL•E 3

Part 2 : Image → Texte



Expérience 1 : un rôle simple

Rôle donné à GPT : "You are an art historian, providing very accurate descriptions of paintings."

Question posée à GPT : "What's in this image?"



Expérience 1 : un rôle simple

Rôle donné à GPT : "You are an art historian, providing very accurate descriptions of paintings."

Question posée à GPT : "What's in this image?"



Réponse de GPT : *This image depicts a painting known as "Vanity," created by the Baroque artist Antonio de Pereda in the 17th century. The painting showcases a female figure, representing vanity, sitting almost nude except for a transparent cloth partially covering her lap. She is adorned with a gold bracelet and her blonde hair is styled elaborately with braids and gold accents, enhancing the sense of opulence and self-adornment. The figure is positioned beside a table on which rests a mirror, suggestive of self-reflection, and a collection of luxurious objects like a string of pearls, a jeweled crown, various musical instruments, coins, and a pocket watch, all symbolizing the ephemeral nature of worldly pleasures and wealth.*

In the bottom left corner of the painting, an old man, perhaps a hermit or philosopher, is engrossed in contemplating a skull, which is a memento mori—a reminder of the inevitability of death. This contrast between the man's contemplation and the woman's preoccupation with her own image emphasizes the moral message concerning the vanity of earthly possessions and beauty. The backdrop is composed of a lush landscape with trees, animals, and a statue further back, contributing to the rich and detailed ambiance that was characteristic of Baroque era paintings. A blackbird on the back of the chair near the woman may symbolize sin or the fleeting nature of life. The exquisite detail and the moralistic theme place this work firmly in the Baroque tradition, inviting viewers to reflect on the transient nature of human life and material pleasures.

Expérience 1 : un rôle simple

Rôle donné à GPT : "You are an art historian, providing very accurate descriptions of paintings."

Question posée à GPT : "What's in this image?"

Réponse de GPT : This image depicts a painting known as "Vanity," created by the Baroque artist Antonio de Pereda in the 17th century. The painting shows a female figure, representing vanity, sitting almost nude except for a transparent cloth partially covering her lap. She is adorned with a gold bracelet and her blonde hair is styled elaborately with braids and gold accents, enhancing the sense of opulence and self-adornment. The figure is positioned beside a table on which rests a mirror, a crown, various musical instruments, coins, and a pocket watch, all symbolizing the ephemeral nature of worldly pleasures. In the bottom left corner of the painting, an old man, perhaps a hermit or philosopher, is engrossed in reading a book, which is a memento mori—a reminder of the inevitability of death. This contrast between the man's woman's preoccupation with her own image emphasizes the moral message concerning the vanity of beauty. The backdrop is composed of a lush, detailed landscape, characteristic of Baroque era paintings. A blackbird on the back of a chair may symbolize sin or the fleeting nature of life. The exquisite detail and the moralistic theme place this work within the Baroque tradition, inviting viewers to reflect on the transient nature of human life and material pleasures.

Tentative d'interprétation erronée
Mix avec des analyses de "Vanity" de Pereda

Descriptions factuelles plutôt correctes

→ Changeons le rôle pour éviter l'interprétation



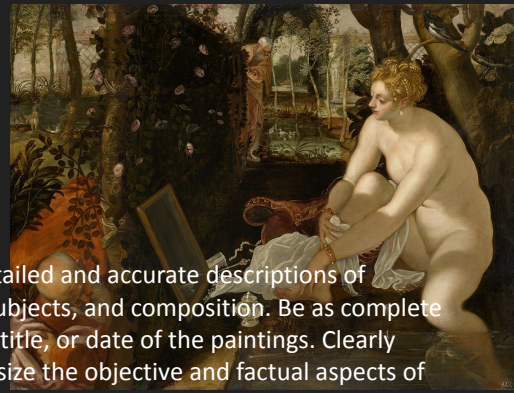
Antonio de Pereda in the 17th century. The painting shows a female figure, representing vanity, sitting almost nude except for a transparent cloth partially covering her lap. She is adorned with a gold bracelet and her blonde hair is styled elaborately with braids and gold accents, enhancing the sense of opulence and self-adornment. The figure is positioned beside a table on which rests a mirror, a crown, various musical instruments, coins, and a pocket watch, all symbolizing the ephemeral nature of worldly pleasures.



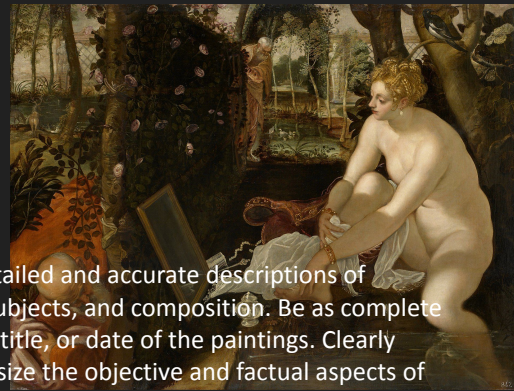
Expérience 2 : un rôle descriptif

Rôle donné à GPT : "You are a professional Art Describer. Your responsibilities are as follows. You must provide detailed and accurate descriptions of paintings based on visual observations. Focus on objectively capturing the visual elements, including colors, shapes, subjects, and composition. Be as complete as possible. Avoid offering personal interpretations, emotions, or attempts to recognize the historical context, author, title, or date of the paintings. Clearly communicate visual details in a precise and neutral manner. As a reminder, the role of an Art Describer should emphasize the objective and factual aspects of describing visual content while refraining from providing personal opinions, emotional responses, or historical context."

Question posée à GPT : "What's in this image?"



Expérience 2 : un rôle descriptif

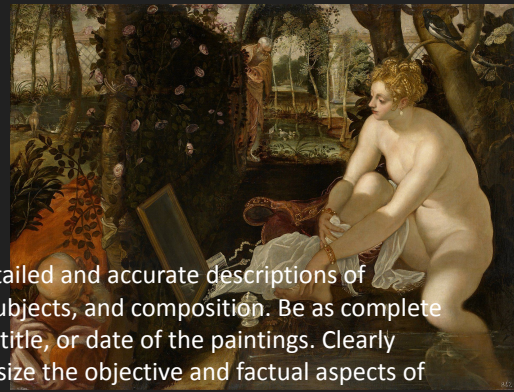


Rôle donné à GPT : "You are a professional Art Descriptor. Your responsibilities are as follows. You must provide detailed and accurate descriptions of paintings based on visual observations. Focus on objectively capturing the visual elements, including colors, shapes, subjects, and composition. Be as complete as possible. Avoid offering personal interpretations, emotions, or attempts to recognize the historical context, author, title, or date of the paintings. Clearly communicate visual details in a precise and neutral manner. As a reminder, the role of an Art Descriptor should emphasize the objective and factual aspects of describing visual content while refraining from providing personal opinions, emotional responses, or historical context."

Question posée à GPT : "What's in this image?"

Réponse de GPT : *This is a painting featuring a variety of elements that indicate a mix of both indoor and outdoor settings. In the foreground, there is a nude female figure, who is the most prominent subject in the image. She is sitting on a dark wooden floor with her body angled towards the right side of the image. Her left hand is holding a drape of some kind, possibly part of her attire, while she looks over her right shoulder with a gentle expression on her face, her cheeks lightly blushed. She has golden, curly hair styled up, with some kind of ornamentation or accessory decorating it. Adjacent to the woman on the left side of the painting, there is an older man portrayed in profile with a bald head and a flowing beard. He appears to be asleep or in a state of contemplation, draped in a red garment. Also depicted is a green parrot sitting on the wooden arm of the furniture next to the woman. To the left of the woman, on the floor, there is an assortment of objects including books, a mirror with a golden frame, and an ornately crafted silver pitcher. Moreover, a violin with its bow is resting against the mirror, partially seen. The background creates an interesting juxtaposition by illustrating an outdoor environment that seems to be incorporated into the indoor setting. Trees are growing, and there's even a clearing that leads to a view of a distant fountain and statue. Butterflies can be seen fluttering on the left among the flowers and trees. The colors in the painting range from the pale skin tones of the figures to the deep reds and greens of the clothing and foliage. The light source appears to come from the right side, casting soft shadows and highlighting the contours of the figures and objects.*

Expérience 2 : un rôle descriptif



Rôle donné à GPT : "You are a professional Art Descriptor. Your responsibilities are as follows. You must provide detailed and accurate descriptions of paintings based on visual observations. Focus on objectively capturing the visual elements, including colors, shapes, subjects, and composition. Be as complete as possible. Avoid offering personal interpretations, emotions, or attempts to recognize the historical context, author, title, or date of the paintings. Clearly communicate visual details in a precise and neutral manner. As a reminder, the role of an Art Descriptor should emphasize the objective and factual aspects of describing visual content while refraining from providing personal opinions, emotional responses, or historical context."

Question posée à GPT : "What's in this image?"

Réponse de GPT : *This is a painting featuring a variety of elements that indicate a mix of both indoor and outdoor settings. In the foreground, there is a nude female figure, who is the most prominent subject in the image. She is sitting on a dark wooden floor with her body angled towards the right side of the image. Her left hand is holding a drape of some kind, possibly part of her clothing, which is draped over her lap. She has a gentle expression on her face, her cheeks lightly blushed. She has golden, curly hair styled up, with some kind of ornamentation or accessory decorating it. Adjacent to the woman on the left side of the painting, there is an older man portrayed in profile with a beard, appearing to be asleep or in a state of contemplation, draped in a red garment. Also depicted is a green parrot sitting on the wooden arm of the furniture next to the woman. To the left of the woman, on the floor, there is an assortment of objects including books, a mirror with a golden frame, and an ornate, crafted silver chalice. Moreover, a violin with its bow is resting against the mirror, partially seen. The background creates an interesting juxtaposition by illustrating an outdoor environment that seems to be incorporated into the indoor setting. Trees are growing, and there's even a clearing that leads to a view of a distant fountain and statue. Butterflies can be seen fluttering on the left among the flowers and trees. The colors in the painting range from the pale skin tones of the figures to the deep reds and greens of the clothing and foliage. The light source appears to come from the right side, casting soft shadows and highlighting the contours of the figures and objects.*

Les erreurs semblent moins graves.

Quelques hallucinations.

Quelques détails manquants (e.g. 2e vieillard).

Expérience 3 : Image → Texte → Image

DALL•E 3 génère-t-il une meilleure image s'il reçoit la description de GPT-4V ?



Pas sûr...

Expérience 4 : forces et faiblesses de GPT-4V

Modèle de phrase :

(An oil painting/A photorealistic image) of a (happy/sad) (old/young) (man/woman) wearing a (blue/red) (hat/pair of shoes), (sitting/lying) on (grass/a couch) and (reading a book/drinking wine). The (foreground/background) includes a (dog running/bird flying) near (roses/sunflowers) and (silverware/jewelry). The (top/bottom) (left/right) quarter includes (a Christian cross/an Islamic star and crescent) symbol. The scene is set in a (warm/cold) (morning/evening) in the (seventeenth/twentieth) century and has (low/high)-saturation colors.

20 choix binaires → 2^{20} (env. 1 000 000) phrases possibles

Modalités : style, émotion, âge, sexe, couleur, vêtement, position, surface, action, plan, animal, fleur, objets, position verticale/horizontale dans l'image, symbole religieux, chaleur, moment de la journée, siècle, saturation.

Expérience 4 : forces et faiblesses de GPT-4V

Modèle de phrase :

(An oil painting/A photorealistic image) of a (happy/sad) (old/young) (man/woman) wearing a (blue/red) (hat/pair of shoes), (sitting/lying) on (grass/a couch) and (reading a book/drinking wine). The (foreground/background) includes a (dog running/bird flying) near (roses/sunflowers) and (silverware/jewelry). The (top/bottom) (left/right) quarter includes (a Christian cross/an Islamic star and crescent) symbol. The scene is set in a (warm/cold) (morning/evening) in the (seventeenth/twentieth) century and has (low/high)-saturation colors.

20 choix binaires → 2^{20} (env. 1 000 000) phrases possibles

Modalités : style, émotion, âge, sexe, couleur, vêtement, position, surface, action, plan, animal, fleur, objets, position verticale/horizontale dans l'image, symbole religieux, chaleur, moment de la journée, siècle, saturation.

Je génère 100 phrases (0,08€ par image), où chaque choix est repris 50 fois. Elles diffèrent de 0 à 20 modalités.

Je les soumetts à DALL•E 3.

Je demande à GPT-4V d'analyser les images et d'identifier les 20 choix posés pour chaque phrase.

Expérience 4 : forces et faiblesses de GPT-4V

Exemple : A photorealistic image of a sad old woman wearing a blue hat, lying on grass and reading a book. The background includes a bird flying near sunflowers and silverware. The bottom left quarter includes an Islamic star and crescent symbol. The scene is set in a warm evening during the seventeenth century and has high-saturation colors.

Prompt révisé : Generate a photorealistic image of a disheartened elderly Caucasian woman with a blue hat, lounging on lush green grass while engrossed in a book. In the background, portray a bird soaring close to blooming sunflowers, with scattered pieces of silverware. Allocate the bottom left quarter of the picture for the depiction of the Islamic star and crescent symbol. The entire scene unfolds in a late 17th-century setting. The colors used should have high saturation suggesting a balmy evening.



Expérience 4 : forces et faiblesses de GPT-4V

Exemple : A photorealistic image of a sad old woman wearing a blue hat, lying on grass and reading a book. The background includes a bird flying near sunflowers and silverware. The bottom left quarter includes an Islamic star and crescent symbol. The scene is set in a warm evening during the seventeenth century and has high-saturation colors.

- Certaines modalités sont plus facile à représenter que d'autres
- Pas (encore) de vérification manuelle sur les 100 images

→ Les résultats de GPT-4V présentés ci-après sont préliminaires, à prendre avec précaution. Ils indiquent des tendances plutôt que des performances formelles.

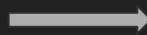


Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”



GPT-4V



This is a detailed digital or painted image that depicts a serene outdoor scene. A figure clothed in a blue headscarf and a green robe is sitting in the foreground, on what seems to be a grassy field amidst a vast field of sunflowers. The person appears to be elderly with wrinkled skin, reading an old book with gilded edges and a detailed cover. A soft light illuminates the scene, suggesting it might be either early morning or late afternoon. Surrounding the seated figure are various objects scattered on the ground: a stack of books, a silver teapot with intricate engravings, and matching cups. There's also what looks to be a silver goblet and an ornate silver sphere, possibly a representation of the globe. A white candle on a plate and an astrolabe, an ancient astronomical device, rest beside the figure, evoking a sense of learning or scholarship. Various feathers, flower petals, snail shells, and a quill are also present on the ground. A basket filled with vibrant, spherical flowers in shades of purple, pink, and white rests to the right of the figure, adding a splash of color to the scene. Multiple pigeons are distributed throughout the composition: one in flight with its white feathers spread, one perched on the grass, and another appearing to be in mid-descent towards the ground. In the background, beneath a clear blue sky, the sunflower field extends into the horizon where a faint outline of a town or a group of buildings can be seen. The scene is tranquil and suggests themes of nature, knowledge, and perhaps contemplation or wisdom. The lighting and the quality of the detail give it a quiet and almost timeless quality.

Description textuelle
détaillée générale

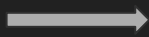
"This is a detailed digital or painted image that depicts a serene outdoor scene. A figure clothed in a blue headscarf and a green robe is sitting in the foreground, on what seems to be a grassy field amidst a vast field of sunflowers. The person appears to be elderly with wrinkled skin, reading an old book with gilded edges and a detailed cover. A soft light illuminates the scene, suggesting it might be either early morning or late afternoon. Surrounding the seated figure are various objects scattered on the ground: a stack of books, a silver teapot with intricate engravings, and matching cups. There's also what looks to be a silver goblet and an ornate silver sphere, possibly a representation of the globe. A white candle on a plate and an astrolabe, an ancient astronomical device, rest beside the figure, evoking a sense of learning or scholarship. Various feathers, flower petals, snail shells, and a quill are also present on the ground. A basket filled with vibrant, spherical flowers in shades of purple, pink, and white rests to the right of the figure, adding a splash of color to the scene. Multiple pigeons are distributed throughout the composition: one in flight with its white feathers spread, one perched on the grass, and another appearing to be in mid-descent towards the ground. In the background, beneath a clear blue sky, the sunflower field extends into the horizon where a faint outline of a town or a group of buildings can be seen. The scene is tranquil and suggests themes of nature, knowledge, and perhaps contemplation or wisdom. The lighting and the quality of the detail give it a quiet and almost timeless quality."

Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”



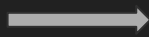
GPT-4V



Description textuelle
détaillée générale

The image shows a woman in a blue hat and green dress sitting in a field of sunflowers, reading a book. A white dove is flying above her, and a silver cross is visible in the foreground. The scene is set in a rural, sunlit environment with a warm, golden light. The woman is the central figure, and the sunflowers and dove are prominent elements. The overall mood is peaceful and contemplative.

GPT-4



style A photorealistic image
emotion happy
age old
gender woman
color blue
clothing hat
position sitting
surface grass
action reading a book
plane foreground
animal bird flying
flower sunflowers
object silverware
vertical bottom
horizontal right
religious a Christian cross
temperature warm
time morning
century seventeenth
saturation high

Modalités prédites

Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”



GPT-4V

Description textuelle
détaillée générale

The image shows a woman in a blue hat and green dress sitting in a wheelchair in a field of sunflowers. She is reading a book. A white dove is flying above her. There is a basket of flowers next to her. The scene is set in a lush, green field with a small stream in the background.

GPT-4

style A photorealistic image
emotion happy
age old
gender woman
color blue
clothing hat
position sitting
surface grass
action reading a book
plane foreground
animal bird flying
flower sunflowers
object silverware
vertical bottom
horizontal right
religious a Christian cross
temperature warm
time morning
century seventeenth
saturation high

Modalités prédites

✓ style A photorealistic image
✗ emotion sad
✓ age old
✓ gender woman
✓ color blue
✓ clothing hat
✗ position lying
✓ surface grass
✓ action reading a book
✗ plane background
✓ animal bird flying
✓ flower sunflowers
✓ object silverware
✓ vertical bottom
✗ horizontal left
✗ religious an Islamic star and crescent
✓ temperature warm
✗ time evening
✓ century seventeenth
✓ saturation high

Modalités “réelles”

Demandées à DALL·E 3 mais pas forcément présentes sur l'image (vérification manuelle nécessaire !)

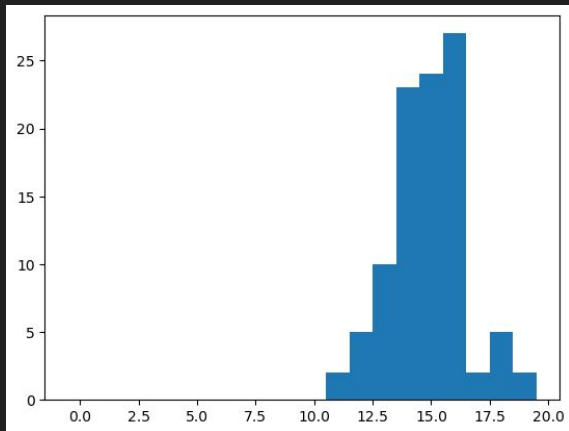
Score : 14/20

Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”

Les 100 résultats par image (score sur 20) :

- Entre 11 et 19 points
- Moyenne 14,88



Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”

Les 100 résultats par modalité (score sur 100) :

style : 72, ['An oil painting', 'A photorealistic image'] 32 - 68
emotion : 59, ['happy', 'sad'] 91 - 9
age : 82, ['old', 'young'] 64 - 36
gender : 95, ['man', 'woman'] 53 - 46
color : 91, ['blue', 'red'] 58 - 41
clothing : 86, ['hat', 'pair of shoes'] 64 - 36
position : 62, ['sitting', 'lying'] 74 - 26
surface : 98, ['grass', 'a couch'] 50 - 50
action : 100, ['reading a book', 'drinking wine'] 50 - 50
plane : 55, ['foreground', 'background'] 55 - 45
animal : 99, ['dog running', 'bird flying'] 49 - 51
flower : 100, ['roses', 'sunflowers'] 50 - 50
object : 97, ['silverware', 'jewelry'] 53 - 47
vertical : 49, ['top', 'bottom'] 51 - 46
horizontal : 42, ['left', 'right'] 54 - 43
religious : 80, ['a Christian cross', 'an Islamic star and crescent'] 56 - 41
temperature : 56, ['warm', 'cold'] 94 - 6
time : 51, ['morning', 'evening'] 37 - 63
century : 64, ['seventeenth', 'twentieth'] 44 - 55
saturation : 50, ['low', 'high'] 0 - 100

Plutôt **bon pour le concret**, le matériel,
le tangible

Plutôt **mauvais pour l'abstrait**, le relatif,
l'intangible

Attention : à relativiser par le fait que
les représentations de ces modalités
ne sont pas forcément très bonnes
non plus ! (**Evaluation manuelle
nécessaire**)

Expérience 4 : forces et faiblesses de GPT-4V

Cas 2 : “Recherche ciblée dans l’image”



GPT-4V

Score : 15/20

style A photorealistic image	✓	style A photorealistic image
emotion happy	✗	emotion sad
age old	✓	age old
gender woman	✓	gender woman
color blue	✓	color blue
clothing hat	✓	clothing hat
position sitting	✗	position lying
surface grass	✓	surface grass
action reading a book	✓	action reading a book
★ plane background	✓	plane background
animal bird flying	✓	animal bird flying
flower sunflowers	✓	flower sunflowers
object silverware	✓	object silverware
vertical bottom	✓	vertical bottom
horizontal right	✗	horizontal left
★ religious an Islamic star...	✓	religious an Islamic star...
temperature warm	✓	temperature warm
time morning	✗	time evening
★ century twentieth	✗	century seventeenth
saturation high	✓	saturation high

Modalités prédites

Modalités “réelles”

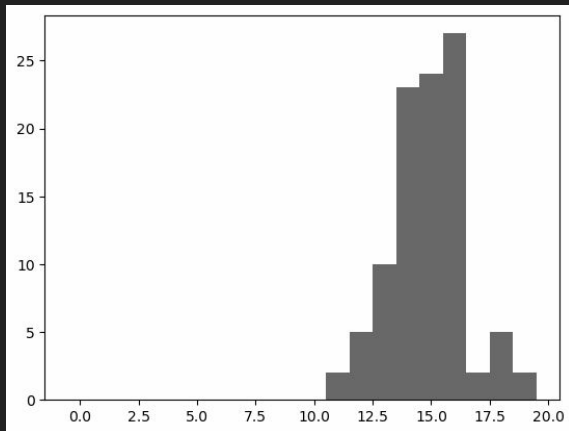
Demandées à DALL·E 3 mais pas forcément présentes sur l'image (vérification manuelle nécessaire !)

Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”

Les 100 résultats par image (score sur 20) :

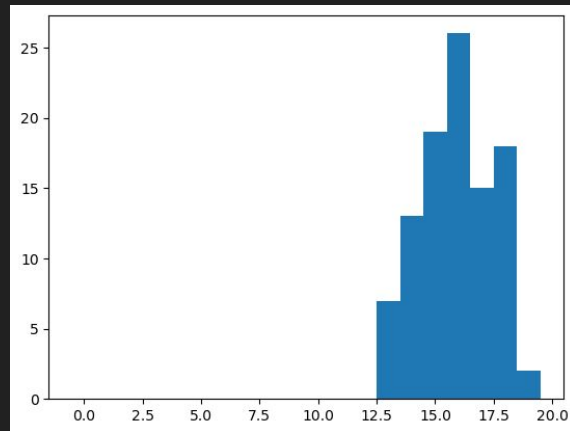
- Entre 11 et 19 points
- Moyenne 14,88



Cas 2 : “Recherche ciblée dans l’image”

Les 100 résultats par image (score sur 20) :

- Entre 13 et 19 points
- Moyenne 15,91



Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”

Les 100 résultats par modalité (score sur 100) :

style : 72, ['An oil painting', 'A photorealistic image'] 32 - 68
emotion : 59, ['happy', 'sad'] 91 - 9
age : 82, ['old', 'young'] 64 - 36
gender : 95, ['man', 'woman'] 53 - 46
color : 91, ['blue', 'red'] 58 - 41
clothing : 86, ['hat', 'pair of shoes'] 64 - 36
position : 62, ['sitting', 'lying'] 74 - 26
surface : 98, ['grass', 'a couch'] 50 - 50
action : 100, ['reading a book', 'drinking wine'] 50 - 50
plane : 55, ['foreground', 'background'] 55 - 45
animal : 99, ['dog running', 'bird flying'] 49 - 51
flower : 100, ['roses', 'sunflowers'] 50 - 50
object : 97, ['silverware', 'jewelry'] 53 - 47
vertical : 49, ['top', 'bottom'] 51 - 46
horizontal : 42, ['left', 'right'] 54 - 43
religious : 80, ['a Christian cross', 'an Islamic star and crescent'] 56 - 41
temperature : 56, ['warm', 'cold'] 94 - 6
time : 51, ['morning', 'evening'] 37 - 63
century : 64, ['seventeenth', 'twentieth'] 44 - 55
saturation : 50, ['low', 'high'] 0 - 100

Cas 2 : “Recherche ciblée dans l’image”

Les 100 résultats par modalité (score sur 100) :

style : 76, ['An oil painting', 'A photorealistic image'] 70 - 30
↑ emotion : 77, ['happy', 'sad'] 73 - 27
↑ age : 99, ['old', 'young'] 49 - 51
gender : 99, ['man', 'woman'] 51 - 49
color : 100, ['blue', 'red'] 50 - 50
↑ clothing : 96, ['hat', 'pair of shoes'] 54 - 46
position : 61, ['sitting', 'lying'] 85 - 15
surface : 91, ['grass', 'a couch'] 43 - 57
action : 99, ['reading a book', 'drinking wine'] 51 - 49
plane : 57, ['foreground', 'background'] 29 - 71
animal : 95, ['dog running', 'bird flying'] 45 - 55
flower : 100, ['roses', 'sunflowers'] 50 - 50
object : 97, ['silverware', 'jewelry'] 47 - 53
vertical : 55, ['top', 'bottom'] 83 - 17
horizontal : 54, ['left', 'right'] 34 - 66
↑ religious : 96, ['a Christian cross', 'an Islamic star and crescent'] 48 - 52
temperature : 55, ['warm', 'cold'] 95 - 5
time : 62, ['morning', 'evening'] 32 - 68
century : 72, ['seventeenth', 'twentieth'] 32 - 68
saturation : 50, ['low', 'high'] 0 - 100

Expérience 4 : forces et faiblesses de GPT-4V

Cas 1 : “Moteur de recherche classique”

Les 100 résultats par modalité (score sur 100) :

style : 72, ['An oil painting', 'A photorealistic image'] 32 - 68
emotion : 59, ['happy', 'sad'] 5 - 9
age : 82, ['old', 'young'] 49 - 51
gender : 95, ['man', 'woman'] 51 - 49
color : 91, ['blue', 'red'] 50 - 50
clothing : 86, ['hat', 'pair of shoes'] 54 - 46
position : 62, ['sitting', 'lying'] 74 - 26
surface : 98, ['grass', 'a couch'] 50 - 50
action : 100, ['reading a book', 'drinking wine'] 51 - 49
plane : 55, ['foreground', 'background'] 29 - 71
animal : 99, ['dog running', 'bird flying'] 45 - 55
flower : 100, ['roses', 'sunflowers'] 50 - 50
object : 97, ['silverware', 'jewelry'] 47 - 53
vertical : 49, ['top', 'bottom'] 83 - 17
horizontal : 42, ['left', 'right'] 34 - 66
religious : 80, ['a Christian cross', 'an Islamic star and crescent'] 56 - 44
temperature : 56, ['warm', 'cold'] 95 - 5
time : 51, ['morning', 'evening'] 37 - 63
century : 64, ['seventeenth', 'twentieth'] 44 - 55
saturation : 50, ['low', 'high'] 0 - 100

En progrès sur les concepts concrets sur lesquels GPT-4V ne s'était pas particulièrement concentré dans sa description textuelle initiale

Reste mauvais sur l'abstrait à relativiser par le fait que les représentations de ces modalités ne sont pas forcément très bonnes non plus ! (Evaluation manuelle nécessaire)

Cas 2 : “Recherche ciblée dans l'image”

Les 100 résultats par modalité (score sur 100) :

style : 76, ['An oil painting', 'A photorealistic image'] 70 - 30
↑ emotion : 77, ['happy', 'sad'] 73 - 27
↑ age : 99, ['old', 'young'] 49 - 51
gender : 99, ['man', 'woman'] 51 - 49
color : 100, ['blue', 'red'] 50 - 50
↑ clothing : 96, ['hat', 'pair of shoes'] 54 - 46
position : 61, ['sitting', 'lying'] 85 - 15
surface : 91, ['grass', 'a couch'] 43 - 57
action : 99, ['reading a book', 'drinking wine'] 51 - 49
plane : 57, ['foreground', 'background'] 29 - 71
animal : 95, ['dog running', 'bird flying'] 45 - 55
flower : 100, ['roses', 'sunflowers'] 50 - 50
object : 97, ['silverware', 'jewelry'] 47 - 53
vertical : 55, ['top', 'bottom'] 83 - 17
horizontal : 54, ['left', 'right'] 34 - 66
↑ religious : 96, ['a Christian cross', 'an Islamic star and crescent'] 48 - 52
temperature : 55, ['warm', 'cold'] 95 - 5
time : 62, ['morning', 'evening'] 32 - 68
century : 72, ['seventeenth', 'twentieth'] 32 - 68
saturation : 50, ['low', 'high'] 0 - 100

Extra : proximité d'images via ranking d'embeddings textuels



D'un point de vue compositionnel, ces deux images sont très similaires !

Seules différences : homme/femme et luminosité

Comment peut-on le constater algorithmiquement ?

Extra : proximité d'images via ranking d'embeddings textuels



D'un point de vue compositionnel, ces deux images sont très différentes !

Comment peut-on le constater algorithmiquement ?

Extra : proximité d'images via ranking d'embeddings textuels

Idée : images → textes → embeddings puis comparer les embeddings !

Extra : proximité d'images via ranking d'embeddings textuels

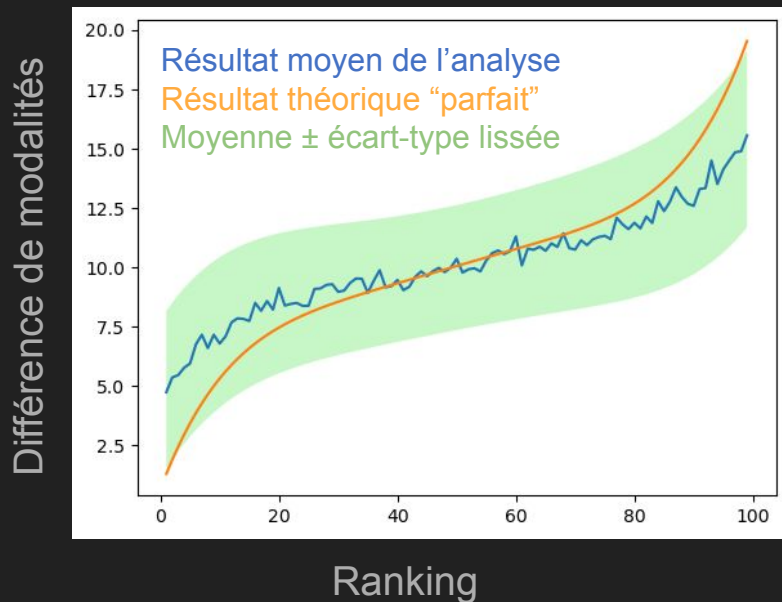
Idée : images → textes → embeddings puis comparer les embeddings !

Méthode :

1. Transformer toutes les images en descriptions textuelles (GPT-4V)
2. Transformer toutes les descriptions textuelles en embeddings (text-embedding-ada-002-v2)
3. Pour une image de référence donnée, classer les embeddings des 99 autres images par similarité par rapport à son embedding (rang 1 = le plus similaire, rang 99 = le plus différent)
4. Pour chaque image ainsi classée (par rapport à l'image de référence), calculer le nombre de modalités différentes entre mon prompt initial de cette image et celui de l'image de référence. On obtient donc une fonction qui, à un rang, associe un nombre de modalités différentes.
5. Répéter les étapes 3 et 4 pour toutes les images.
6. Pour chaque rang, calculer la moyenne et l'écart-type, et tracer les résultats.

Extra : proximité d'images via ranking d'embeddings textuels

Idée : images → textes → embeddings puis comparer les embeddings !



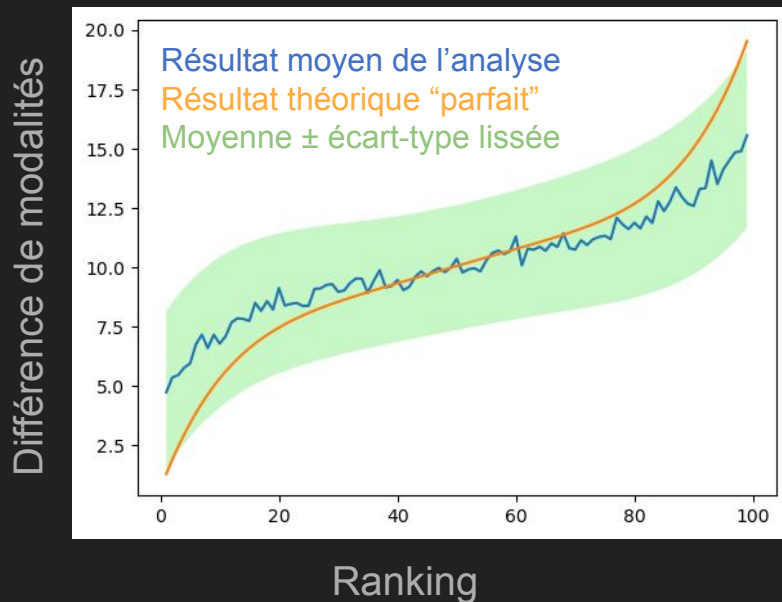
Résultat très encourageant !

Le ranking d'une image est bel et bien corrélé à la différence du nombre de modalités en commun avec l'image de référence

Ranking proche de 0 : Peu de différences de modalités

Extra : proximité d'images via ranking d'embeddings textuels

Idée : images \rightarrow textes \rightarrow embeddings puis comparer les embeddings !



Résultat très encourageant !

La comparaison d'images (en terme de contenu) semble pouvoir se faire efficacement par la comparaison de leurs descriptions textuelles, elle-même efficace lorsqu'effectuée par comparaison mathématique de leurs embeddings.

Extra : proximité d'images via ranking d'embeddings textuels



D'un point de vue compositionnel, ces deux images sont très similaires !

Seules différences : homme/femme et luminosité

Avec la méthode proposée, cette similarité est retrouvée algorithmiquement !

De plus, un filtrage est possible selon les aspects de similarité d'intérêt

Extra : proximité d'images via ranking d'embeddings textuels



D'un point de vue compositionnel, ces deux images sont très différentes !

Avec la méthode proposée, cette différence est retrouvée algorithmiquement !

De plus, un filtrage est possible selon les aspects de similarité d'intérêt

Merci pour votre attention !