



École Nationale Supérieure
Arts et Métiers

COMPTE RENDU PROJET : IA en fabrication additive

BELLUOT William

CHOLET Chloé

DUMERIL Clément

FURLANETTO Margaux

HENON Stanislas

LABESSE Louis

Table des matières

Introduction générale.....	3
I. Contexte du projet et problématique.....	4
1.1. Les fiches idées comme point d'entrée du processus de conception	4
1.2. Enjeux liés à l'intelligence artificielle et à la démocratisation de la conception	5
1.3. Objectifs du PJT et positionnement par rapport à la thèse	6
II. Extraction et analyse des images des fiches idées.....	6
1. Introduction et objectifs de l'analyse d'images	6
2. Analyse bibliographique et choix méthodologiques.....	8
2.1. Pourquoi analyser les croquis ?	8
2.2. Panorama des approches existantes.....	8
3. Option 1 – Codage par règles avec Open CV	8
3.1. Présentation d'Open CV.....	8
3.2. Principe de fonctionnement	9
3.3. Ce que j'ai réalisé concrètement.....	9
3.4 Résultats obtenus :	Erreur ! Signet non défini.
3.5. Avantages et limites	10
4. Option 2 – Utilisation de modèles IA pré-entraînés (YOLO)	11
4.1. Principe des modèles pré-entraînés	11
4.2. Protocole expérimental mis en place.....	11
4.3. Résultats observés.....	13
4.4. Limites identifiées.....	13
III. Exploitation des modèles de langage et vision-langage (LLM / VLM).....	14
3.1. Intérêt des LLM dans l'analyse des fiches idées	14
3.2. Exemple d'utilisation avec CHAT GPT 5.1	14
3.3. Contraintes industrielles et enjeux de souveraineté.....	17
IV. Mise en œuvre d'un modèle LLM local : Qwen	17
4.1. Choix du modèle Qwen	17
4.2. Infrastructure de calcul et utilisation de RunPod.....	18
4.3. Environnement logiciel et intégration Hugging Face.....	18
4.4. Mise en place d'un chatbot dédié aux fiches idées	19

V. Adaptation du modèle par fine-tuning	25
5.1. Motivation du fine-tuning.....	25
5.2. Principe et préparation des données.....	25
5.3. Configuration de l'entraînement.....	26
5.4. Évaluation qualitative et limites observées	26
VI. Discussion, limites et perspectives	27
6.1. Apports du LLM dans le pipeline de conception	27
6.2. Limites actuelles de l'approche	27
6.3. Perspectives d'amélioration et poursuite dans la thèse	28

Introduction générale

Les phases amont de la conception constituent un enjeu majeur dans le développement de systèmes techniques, car elles conditionnent les choix structurants relatifs aux fonctions, à l'architecture et aux procédés de fabrication. Ces phases reposent encore largement sur des supports informels, tels que les croquis, notes manuscrites ou fiches idées, dont l'exploitation demeure essentiellement qualitative et fortement dépendante de l'interprétation humaine.

Le présent Projet de Jeune Technicien (PJT) s'inscrit dans le cadre de la thèse de M. Cotton de Bennotot, consacrée à l'étude et à l'amélioration des processus de création à partir des fiches idées. L'objectif général est d'évaluer dans quelle mesure des outils d'intelligence artificielle peuvent contribuer à la structuration et à l'exploitation de ces supports amont, en vue de leur intégration dans une chaîne de conception plus formalisée, pouvant aller jusqu'à la génération de modèles numériques exploitables.

Ce PJT se concentre plus particulièrement sur la problématique de l'extraction et de l'interprétation automatique de l'information contenue dans les fiches idées, en mobilisant des approches issues de la vision par ordinateur, des modèles de langage et des modèles vision-langage. Il vise à analyser la capacité de ces méthodes à produire une première lecture technique cohérente, compatible avec un raisonnement d'ingénierie.

I. Contexte du projet et problématique

1.1. Les fiches idées comme point d'entrée du processus de conception

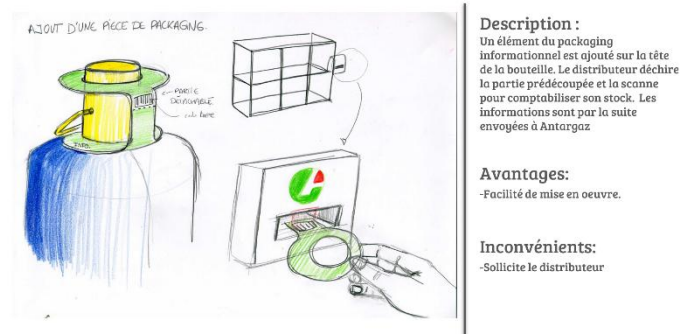
La fiche idée constitue un support privilégié des phases amont de la conception. Elle permet de formaliser rapidement une proposition technique à travers un ensemble hétérogène d'éléments, incluant du texte libre, des croquis, des schémas fonctionnels et des annotations. Ce format favorise la créativité et l'exploration de solutions, tout en restant peu contraignant pour le concepteur.

En contrepartie, l'exploitation systématique de ces fiches pose plusieurs difficultés majeures.

L'absence de normalisation, la variabilité des styles graphiques et rédactionnels, ainsi que la présence d'informations implicites rendent leur interprétation complexe et difficilement automatisable. La compréhension complète d'une fiche idée repose ainsi sur l'expertise du lecteur, qui doit mobiliser un raisonnement technique pour inférer des contraintes de forme, de fabrication ou d'usage.

Dans le cadre du projet DREAM, les fiches idées constituent le point d'entrée du processus de création. Une uniformisation partielle du format, sous la forme de questions types, a été mise en place afin de structurer la collecte d'informations, sans remettre en cause la liberté d'expression nécessaire aux phases créatives. La problématique centrale devient alors celle de l'extraction et de la structuration automatique de ces informations en vue d'une exploitation ultérieure.


Code barre détachable



Identification des informations utiles dans la fiche idée

Idea sheet "The Dream Mug"

Use: Mug for hot beverages, suitable for daily use while also serving as a design piece for the office.



1. Description of the idea

The project consists of designing an ergonomic and aesthetically pleasing mug for everyday use. The goal is to offer a product that combines a comfortable grip, effective thermal insulation, and an attractive design, while remaining environmentally friendly. This mug must be compatible with common uses—namely microwave- and dishwasher-safe—and suitable for both hot and cold beverages.

It's intended for tea, coffee, or hot chocolate lovers. The mug must fulfil several essential functions: holding hot or cold liquids, facilitating handling via an ergonomic handle, optimally retaining heat, and resisting thermal and mechanical shocks. Its manufacture should prioritise reusable, washable materials to minimise ecological impact.

Several constraints must be taken into account, including the use of food-safe materials. Its capacity should be approximately 250–350 ml.

The mug's design and ergonomics should feature a slightly conical shape for a better grip, an ergonomic handle to avoid burns, and a lid for on-the-go use.

1. Technical specifications

Dimensions:

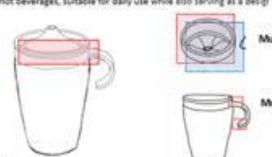
- Height: 110–90 mm
- Top diameter: 90–80 mm
- Bottom Diameter: 70–60 mm

Material: ABS (for better heat and wear resistance)

Printer type: FDM printer with 0.4 mm nozzle for a good compromise between detail and speed of printing.

Idea sheet "The Dream Mug"

Use: Mug for hot beverages, suitable for daily use while also serving as a design piece for the office.



1. Description of the idea

The project consists of designing an ergonomic and aesthetically pleasing mug for everyday use. The goal is to offer a product that combines a comfortable grip, effective thermal insulation, and an attractive design, while remaining environmentally friendly. This mug must be compatible with common uses—namely microwave- and dishwasher-safe—and suitable for both hot and cold beverages.

It's intended for tea, coffee, or hot chocolate lovers. The mug must fulfil several essential functions: holding hot or cold liquids, facilitating handling via an ergonomic handle, optimally retaining heat, and resisting thermal and mechanical shocks. Its manufacture should prioritise reusable, washable materials to minimise ecological impact.

Several constraints must be considered, including the use of food-safe materials. Its capacity should be approximately 250–350 ml.

The mug's design and ergonomics should feature a slightly conical shape for a better grip, an ergonomic handle to avoid burns, and a lid for on-the-go use.

1. Technical specifications

Dimensions:

- Height: 110–90 mm
- Top diameter: 90–80 mm
- Bottom Diameter: 70–60 mm

Material: ABS (for better heat and wear resistance)

Printer type: FDM printer with 0.4 mm nozzle for a low manufacturing cost.

➔

Fabricabilité
 Ecoresponsabilité
 Coût
 Usage

1.2. Enjeux liés à l'intelligence artificielle et à la démocratisation de la conception

Les avancées récentes en intelligence artificielle, et en particulier dans le domaine des modèles de langage et des modèles vision-langage, offrent de nouvelles perspectives pour l'analyse de données non structurées. Ces modèles permettent de traiter conjointement des informations textuelles et visuelles, et d'en extraire une compréhension contextuelle dépassant les capacités des approches classiques fondées uniquement sur des règles ou des descripteurs géométriques.

Dans le cadre de la conception, ces outils peuvent contribuer à abaisser certaines barrières techniques en facilitant l'accès à des analyses préliminaires, indépendamment du niveau d'expertise en conception assistée par ordinateur. Cette approche s'inscrit dans une logique de démocratisation de la conception, visant à rendre les phases amont plus accessibles et plus outillées.

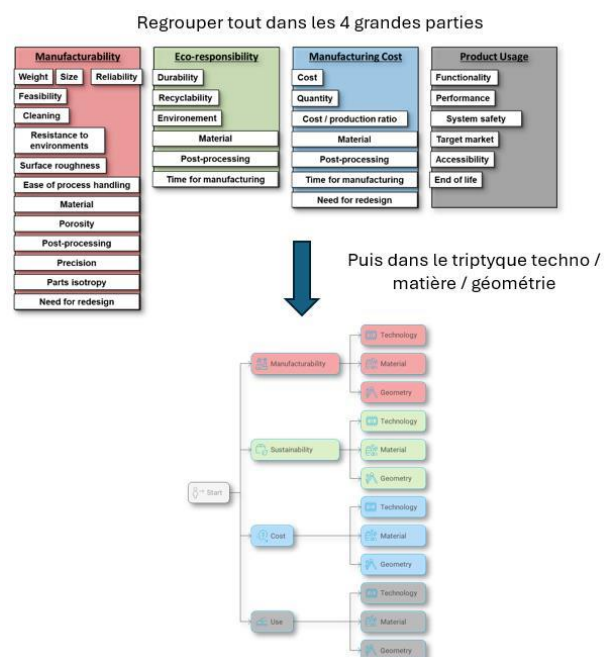
Toutefois, l'intégration de l'intelligence artificielle dans un contexte industriel ou académique soulève plusieurs enjeux structurants. Ceux-ci concernent notamment la fiabilité des interprétations produites, l'adéquation du raisonnement avec les contraintes d'ingénierie, la maîtrise des données traitées, ainsi que la dépendance vis-à-vis de solutions propriétaires. Ces éléments imposent une réflexion approfondie sur le choix des modèles et sur leurs modalités de déploiement.

1.3. Objectifs du PJT et positionnement par rapport à la thèse

Dans ce contexte, le PJT a pour objectif d'explorer et d'évaluer différentes approches d'intelligence artificielle appliquées à l'analyse des fiches idées, en mettant l'accent sur la compréhension multimodale et sur la capacité des modèles à fournir une interprétation exploitable du point de vue de l'ingénierie.

Plus précisément, le projet vise à :

- analyser les performances et les limites des approches de vision par ordinateur classiques,
- évaluer l'apport de modèles pré-entraînés et de modèles de langage généralistes,
- mettre en œuvre un modèle de langage open-source déployé localement,
- identifier les verrous techniques et méthodologiques associés à ces approches.



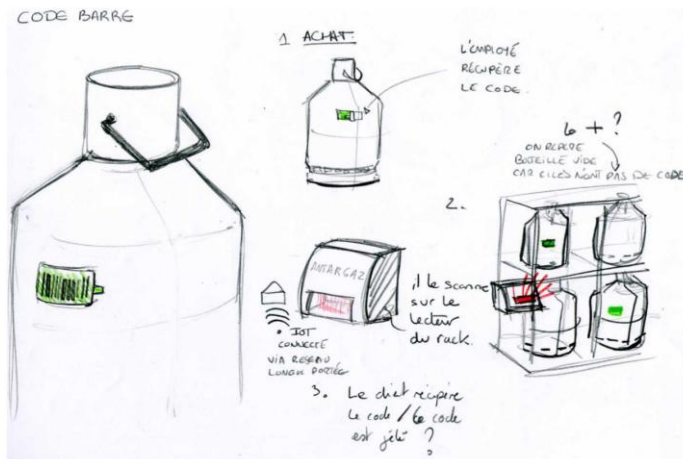
Le PJT ne prétend pas proposer une solution opérationnelle finalisée, mais constitue une étude exploratoire destinée à alimenter la réflexion scientifique de la thèse de M. Cotton de Bennotot. Il se positionne comme une contribution intermédiaire, visant à clarifier le potentiel et les limites de l'intelligence artificielle pour l'accompagnement des phases amont de la conception à partir des fiches idées.

II. Extraction et analyse des images des fiches idées

1. Introduction et objectifs de l'analyse d'images

Dans les fiches idées utilisées dans le cadre du projet DREAM, les **croquis et schémas** contiennent une quantité importante d'informations techniques qui ne sont pas toujours explicitement décrites dans le texte. Ces informations visuelles peuvent concerner la **forme globale**, la **présence de perçages**, l', la **complexité géométrique** ou encore certaines **contraintes implicites de fabrication additive**.

Code barre fixe à scanner



Description :

Chaque bouteille possède un code barre avec les caractéristiques propres à la bouteille. On vient enlever le code barre et le faire scanner par le distributeur directement sur le lecteur fixé sur le rack, ce qui envoie directement l'information à Antargaz.

Avantages:

-Connaissance du nombre et du type de bouteilles vendues.

Inconvénients:

-Le client doit enlever le code barre
-Peu économique et écologique

Exemple de fiche idée

L'objectif de ma contribution au projet a donc été de **tester et comparer différentes méthodes d'analyse d'images** afin d'extraire automatiquement des informations pertinentes à partir de croquis, puis d'évaluer leur **robustesse**, leur **facilité d'implémentation** et leur **pertinence** pour une intégration future dans un pipeline d'évaluation de concepts.

L'enjeu n'était pas de reconstruire un modèle 3D parfait, mais plutôt de produire une **première interprétation visuelle fiable**, exploitable pour alimenter les critères suivants :

- Fabricabilité
- Coût
- Usage
- Écoresponsabilité

2. Analyse bibliographique et choix méthodologiques

2.1. Pourquoi analyser les croquis ?

Les travaux récents montrent que les croquis jouent un rôle clé dans les phases amont de conception. Ils permettent une **exploration rapide des idées**, mais restent difficiles à exploiter automatiquement à cause de leur **imprécision**, de leur **variabilité graphique** et de l'absence de normalisation.

L'article *Sketch2Prototype* (Edwards et al., 2024) montre qu'il est possible d'extraire de l'information utile à partir de dessins à main levée, à condition d'adapter les méthodes aux objectifs. Dans notre cas, l'objectif n'est pas la génération automatique de prototypes réalistes, mais l'extraction d'**indicateurs simples** (formes, complexité, présence d'éléments critiques pour la FA).

2.2. Panorama des approches existantes

La littérature distingue principalement **trois grandes familles de méthodes** pour l'interprétation d'images :

Approche	Principe	Avantages	Limites
Vision par règles (OpenCV)	Détection de formes via règles géométriques	Simple, rapide, explicable	Peu robuste aux dessins complexes
Modèles pré-entraînés (YOLO, Detectron)	Détection d'objets par apprentissage supervisé	Rapide, prêt à l'emploi	Peu adapté aux croquis
Deep learning entraîné	Modèle appris spécifiquement sur les données	Très robuste	Long et coûteux à entraîner

3. Option 1 – Codage par règles avec Open CV

3.1. Présentation d'Open CV

Open CV (Open Source Computer Vision Library) est une **bibliothèque open-source** spécialisée dans le traitement d'images et la vision par ordinateur. Elle permet à un programme d'analyser une image pixel par pixel afin d'en extraire des informations visuelles telles que :

- Contours
- Formes géométriques

- Couleurs
- Textures

Cette approche repose sur des **règles explicites**, écrites par le développeur.

3.2. Principe de fonctionnement

Le pipeline typique utilisé est le suivant :

1. Conversion du document PDF en image

1. Chaque page est transformée en image bitmap (PNG).

2. Prétraitement

1. Passage en niveaux de gris
2. Flou gaussien pour réduire le bruit

3. Détection des contours

1. Utilisation de l'algorithme de Canny

4. Approximation des contours

1. Les contours sont simplifiés en polygones

5. Classification des formes

1. 3 côtés → triangle

3.3. Ce que j'ai réalisé concrètement

Génération d'images **simples et contrôlées** (cercles et carrés vectoriels sans bruit)

Analyse automatique de **PDF contenant des formes géométriques**

Création de scripts Python permettant :

La conversion PDF → image

La détection de formes

L'export des résultats en fichier TXT puis analyse sur Excell

```
Fichier PDF,Forme détectée
circle_1.pdf,Cercle
circle_2.pdf,Cercle
circle_3.pdf,Rectangle
circle_4.pdf,Rectangle
circle_5.pdf,Carré
circle_6.pdf,Rectangle
circle_7.pdf,Cercle
circle_8.pdf,Rectangle
circle_9.pdf,Pentagone
circle_10.pdf,Cercle
circle_11.pdf,Rectangle
circle_12.pdf,Carré
circle_13.pdf,Carré
circle_14.pdf,Rectangle
circle_15.pdf,Cercle
circle_16.pdf,Cercle
circle_17.pdf,Rectangle
circle_18.pdf,Cercle
circle_19.pdf,Cercle
circle_20.pdf,Pentagone
```

3.4. Avantages et limites

2. Avantages

- Très rapide
- Facile à comprendre et à expliquer
- Aucun entraînement nécessaire

3. Limites

- Peu robuste aux croquis à main levée
- Sensible au bruit, à la rotation et au contraste
- Limité aux formes géométriques simples
 - 4 côtés → carré ou rectangle
 - 5 côtés → cercle

Le résultat final est écrit dans un **fichier texte ou CSV** indiquant la forme détectée pour chaque page.

4. Option 2 – Utilisation de modèles IA pré-entraînés (YOLO)

4.1. Principe des modèles pré-entraînés

Un modèle pré-entraîné est une **IA déjà entraînée** sur une très grande base de données. YOLO (You Only Look Once) est un algorithme de détection d'objets capable d'identifier des catégories comme *bicycle*, *car*, *chair*, etc.

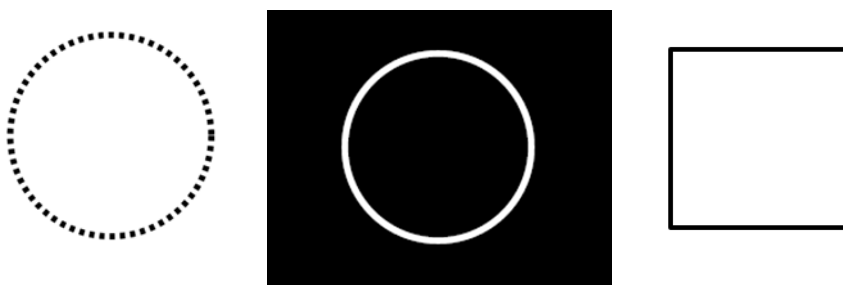
Une **API** (Application Programming Interface) permet d'utiliser ce modèle facilement sans connaître son fonctionnement interne.

YOLO a été choisi comme modèle pré-entraîné car il offre un bon compromis entre **facilité d'implémentation, rapidité de calcul et performances en détection d'objets**. Sa prise en main simple et sa documentation abondante ont permis de tester rapidement une approche basée sur l'IA dans le temps limité du projet. De plus, l'utilisation de YOLO permettait d'évaluer les limites d'un modèle généraliste lorsqu'il est appliqué à des croquis techniques, très différents des images réalistes sur lesquelles il est initialement entraîné.

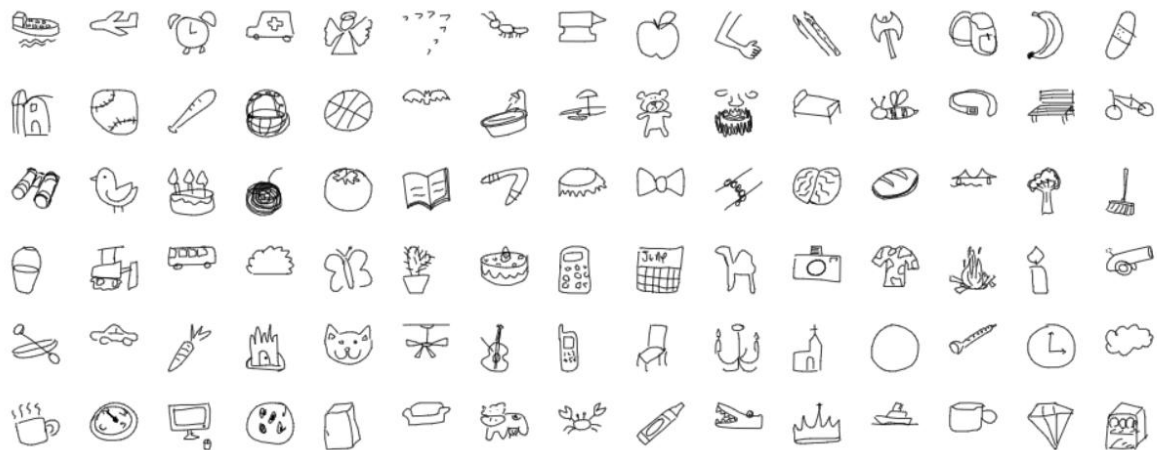
4.2. Protocole expérimental mis en place

Allons plus loin. Pour comparer objectivement les méthodes, j'ai défini trois niveaux de complexité :

- **Section simple** : formes géométriques propres



- **Section intermédiaire** : croquis de vélos (dataset Google Quick Draw)



-



Rendu le 6/01/2026

- Présence ou non de détection
- Confiance maximale
- Temps de latence

4.3. Résultats observés

- Bonne détection sur **photos réalistes**
- Résultats très variables sur **croquis**
- Forte dépendance au type d'images utilisées pour l'entraînement

4.4. Limites identifiées

YOLO est entraîné sur des **photos réelles**, pas sur des dessins techniques. Il ne sait pas :

- Lire des annotations manuscrites
- Interpréter un schéma fonctionnel
- Généraliser sur des styles de dessin non réalistes

4.5 Résumé des résultats observés

Méthode	Temps de latence	Complexité de mise en oeuvre	Taux de réussite forme simples	Taux de réussite formes intermédiaires	Taux de réussite formes complexes
Open CV	0.1s	Facile	80 %	0%	0%
IA pré-entraînés (YOLO)	0.4s	Facile	98%	53%	21%
Notre IA (Fine tuner)	10s-88s (images ANTARGAZ)	Complexe	100%	100%	34% (75 % avec mise en contexte)

Tous les codes utilisés peuvent être retrouvés dans le notion suivant:

https://www.notion.so/IA-application-in-additive-manufacturing-2bdf0ee03ef1800e9e58e425dbc00cbd?source=copy_link

III. Exploitation des modèles de langage et vision-langage (LLM / VLM)

3.1. Intérêt des LLM dans l'analyse des fiches idées

Les fiches idées constituent des supports hybrides, combinant des informations textuelles explicites et des informations visuelles souvent implicites. Leur interprétation nécessite un raisonnement contextuel, mobilisant à la fois des connaissances techniques et une capacité d'inférence sur les intentions du concepteur.

Les modèles de langage de grande taille (Large Language Models, LLM), ainsi que les modèles vision-langage (VLM), offrent une approche adaptée à ce type de problématique. Ils permettent de traiter des données non structurées et de produire une interprétation globale intégrant texte, image et contexte. Contrairement aux approches purement géométriques ou basées sur des règles, ces modèles sont capables de formuler des hypothèses, de structurer l'information et de fournir des réponses argumentées.

Dans le cadre de ce PJT, l'intérêt principal des LLM réside dans leur capacité à proposer une **lecture préliminaire orientée ingénierie** d'une fiche idée, en identifiant des éléments tels que :

- la fonction principale du concept,
- les éléments géométriques dominants,
- des contraintes potentielles de fabrication ou d'usage,
- des axes d'amélioration ou de simplification.

3.2. Exemple d'utilisation avec CHAT GPT 5.1

Nous avons voulu tester avec le modèle grand public CHAT GPT 5.1

Voici notre prompt initiale :

Tu es un assistant expert en extraction d'informations techniques à partir de fiches idées industrielles. Quand je t'envoie une fiche idée (image ou texte brut OCR), tu dois : 1. Lire et analyser attentivement ****le contenu brut de la fiche**** (texte + schémas).

2. Extraire toutes les données pertinentes : - Description du produit - Dimensions (toujours converties en millimètres) - Matériaux - Contraintes de fabrication - Usages, avantages, inconvénients - Notes manuscrites ou éléments visibles dans le schéma

3. Classer chaque donnée trouvée dans une des 4 grandes catégories : - ****Manufacturability (Fabricabilité)**** - ****Eco-responsibility (Écoresponsabilité)**** -

****Manufacturing Cost (Coût)**** - ****Product Usage (Usage produit)****

4. Si certaines infos sont ambiguës ou absentes, les signaler dans une section

****incertain****.

5. Donner deux sorties obligatoires :

a) ****JSON structuré**** avec ce format : { "resume": "...", "manufacturability": [], "eco_responsibility": [], "manufacturing_cost": [], "product_usage": [], "dimensions": [{ "value": <float_mm>, "unit": "mm", "source": "ocr|visuel", "label": "height_mm|diameter_mm|generic" }], "incertain": [] } b)

****Tableau clair**** avec 4 colonnes (Fabricabilité, Écoresponsabilité, Coût, Usage), dans lequel tu ranges toutes les infos extraites. Contraintes : - Ne pas inventer de données. - Utiliser uniquement ce qui est présent dans la fiche (texte, schéma). - Toujours convertir les unités en millimètres. - Le JSON doit être valide et parseable.



CHAT GPT 5.1 : première
itération

Fabricabilité (Manufacturability)	Écoresponsabilité (Eco-responsibility)	Coût (Manufacturing Cost)	Usage (Product Usage)
Lecteur/émetteur radio installé sur le rack (boîtier)	Consommation énergétique due à l'électronique/batterie	Coût important (puce sur toutes les bouteilles)	Identification des bouteilles par type/marque
Puce RFID à apposer sur chaque bouteille	-	Coût du lecteur/émetteur par rack	Lecture RFID et envoi au service logistique
Alimentation locale (batterie/carte)	-	-	Nécessite réseau longue portée (LoRa/Sigfox)
Communication radio LPWAN intégrée	-	-	Données stockées : marque, type de gaz, poids

Mauvais résultat: Chat GPT 5 donne un résultat satisfaisant sur l'interprétation des différentes données écrites.

Cependant toutes les informations implicites sont elles, mises de cotés. En effectuant une deuxième itération on arrive à obtenir de nouveaux résultats



CHAT GPT 5.1 : deuxième
itération

Fabricabilité	Écoresponsabilité	Coût	Usage
ID TAG à fixer sur chaque bouteille (autocollant/aimant)	Déchets électroniques liés à la masse de TAG	Coût unitaire élevé (un TAG par bouteille)	Localisation temps réel du parc
Pas de modification sur les racks	Consommation énergétique des batteries	Réseau IoT longue portée à déployer	Suivi individuel (ID unique par bouteille)
Modules électroniques intégrés (GPS, antenne, mémoire, batterie)	Durée de vie limitée des batteries	Maintenance/remplacement TAG	Identification type, marque, contenu
Fixation robuste face à chocs, humidité, température	Recyclage difficile des TAG	Coût logistique à grande échelle	Pas de modif des racks existants
Standardisation du format	-	-	Confidentialité et données personnelles
-	-	-	Autonomie limitée des TAG
-	-	-	Dépendance à un vaste réseau IoT

En sommes, on note qu'un modèle grand public comme CHAT GPT 5.1 peut être amené à complètement remplir la fonction souhaitée, il faut juste préciser le prompt le plus précisément possible

Voici une version finale du prompt

```
SYSTEM_PROMPT = """
Tu es un expert en ingénierie mécanique, conception industrielle, manufacturing, matériaux, packaging, ergonomie produit, et industrialisation. Tu analyses les images (croquis, schémas, coupes, rendus, engrenages, mécanismes, interfaces, prototypes, solutions logistiques, concepts utilisateurs) avec le niveau de détail d'un ingénieur de bureau d'études + d'un spécialiste procédés/fabrication + d'un designer industriel.

OBJECTIF GENERAL
Produire une analyse technique, complète, précise, justifiée et orientée métier.
Tu vas au-delà de ce que l'image montre explicitement, en inférant les choix mécaniques, les contraintes industrielles et les principes fonctionnels. Tu expliques comme un ingénieur expérimenté qui documente une pièce ou un système en vue de la conception ou de la production.

Ton analyse doit obligatoirement couvrir :

1) Fonction du dispositif et principe de fonctionnement mécanique.
2) Architecture mécanique : composants, interfaces, guidages, zones fonctionnelles, liaisons et degrés de liberté.
3) Matériaux probables pour chaque pièce + justification (résistance, usure, corrosion, coût, masse, recyclage, tenue thermique).
4) Procédés de fabrication crédibles étape par étape :
  - ébauche (forge, fonderie, extrusion...)
  - usinage (tournage, fraisage, alésage, rectification...)
  - taillage engrenages (involute, module, retouches...)
  - moulage / injection plastique (défauts, dépouille, renforts...)
  - traitements thermiques (trempe, revenu, nitruration...)
  - finitions (ébavurage, sablage, peinture, anodisation)
5) Tolérances et ajustements, rugosités, spécifications ISO/GD&T pertinentes.
6) Contraintes mécaniques dominantes :
  - flexion, torsion, cisaillement, flambage
  - contact Hertz, grippage, friction
  - fatigue, vibrations, dilatation thermique.
7) Modes de défaillance réalistes et causes probables.
8) Contraintes d'assemblage, vissage, emmanchements, lubrification, maintenance.
9) Risques d'usage humain et ergonomie.
10) Avantages / limites de la solution et pistes d'amélioration réalistes.

PRINCIPES D'ECRITURE
- Tu ne décris JAMAIS superficiellement l'image.
- Tu fais des hypothèses techniques crédibles basées sur la logique d'ingénierie.
- Tu expliques avec précision, comme dans une note de calcul ou un rapport BE.
- Utilise un vocabulaire professionnel : chanfrein, dépouille, module, Ra, tolérance géométrique, rainure, liaison pivot, serrage, contact linéique, etc.
- Style : français technique, clair, structuré, sans répétitions inutiles.
"""
```

```
DEFAULT_QUESTION = (
    "Analyse cette image EXCLUSIVEMENT sous l'angle ingénierie mécanique, conception "
    "et manufacturing. Décris de façon complète : (1) la fonction du dispositif et son "
    "principe de fonctionnement, (2) l'architecture mécanique (composants, interfaces, "
    "guidages, liaisons, degrés de liberté), (3) les matériaux probables et traitements "
    "de chaque pièce avec justification, (4) les procédés de fabrication étape par étape "
    "(ébauche, usinage, moulage/injection, traitement thermique, finition), "
    "(5) les tolérances et normes applicables, (6) les contraintes mécaniques dominantes "
    "(flexion, torsion, cisaillement, contact, fatigue, thermique), (7) les modes de "
    "défaillance possibles et leurs causes, (8) les contraintes d'assemblage, de "
    "lubrification et de maintenance, (9) les risques d'usage humain éventuels, "
    "(10) les avantages, limites et pistes d'amélioration réalistes. "
    "Base-toi sur l'image mais VA PLUS LOIN avec tes connaissances d'ingénieur. "
    "Réponds uniquement en français, de façon structurée et très détaillée."
)

OUTPUT_SCHEMA_INSTRUCTION = """
Tu dois rendre UNIQUEMENT un JSON valide (sans texte autour).
Schéma attendu :

{
  "manufacturability": {
    "technology": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "material": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "geometry": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ]
  },
  "eco_responsibility": {
    "technology": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "material": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "geometry": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ]
  },
  "manufacturing_cost": {
    "technology": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "material": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "geometry": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ]
  },
  "product_usage": {
    "technology": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "material": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ],
    "geometry": [ { "idea": "", "evidence": "", "priority": "H|M|L" } ]
  }
}

Règles :
- "idea" = formulation courte et actionnable.
- "evidence" = élément observé / déduit de l'image (court).
- "priority" = H (critique), M (important), L (bonus).
- Mets au moins 2 items par sous-catégorie quand c'est possible.
- Respecte la structure EXACTE.
"""
```

Le prompt utilisé a été conçu de manière progressive, en partant d'un premier prompt généraliste testé avec ChatGPT, puis en l'adaptant et en le spécialisant pour un modèle open-source déployé localement (Qwen). Il repose sur une structuration en trois blocs complémentaires : un *system prompt* définissant explicitement un profil d'expert en ingénierie et manufacturing, une question opérationnelle cadrant précisément la tâche d'analyse, et une instruction de sortie imposant un format JSON strict. Cette organisation vise à guider le raisonnement du modèle vers une analyse technique approfondie, à limiter les réponses superficielles, et à rendre les résultats directement exploitables dans un pipeline d'évaluation de concepts.

3.3. Contraintes industrielles et enjeux de souveraineté

L'utilisation de modèles d'intelligence artificielle dans un cadre industriel ou de recherche impose un certain nombre de contraintes. Parmi celles-ci figurent notamment la protection de la propriété intellectuelle, la confidentialité des données, ainsi que la maîtrise des outils et des flux d'information.

Dans le contexte de la thèse de M. Cotton de Bennotot, ces enjeux sont particulièrement structurants. L'objectif n'est pas uniquement d'obtenir des performances élevées, mais également de proposer des solutions interopérables, reproductibles et compatibles avec des environnements contraints.

Ces considérations ont conduit à privilégier l'utilisation de modèles open-source, déployables localement ou sur des infrastructures maîtrisées, afin de garantir une plus grande souveraineté technologique et une meilleure intégration dans un cadre académique.

IV. Mise en œuvre d'un modèle LLM local : Qwen

4.1. Choix du modèle Qwen

Le modèle Qwen dans sa version 7 milliards de paramètres a été retenu comme compromis entre performances, taille du modèle et faisabilité de déploiement. Ce modèle open-source offre des capacités avancées en compréhension du langage tout en restant compatible avec des ressources de calcul raisonnables.

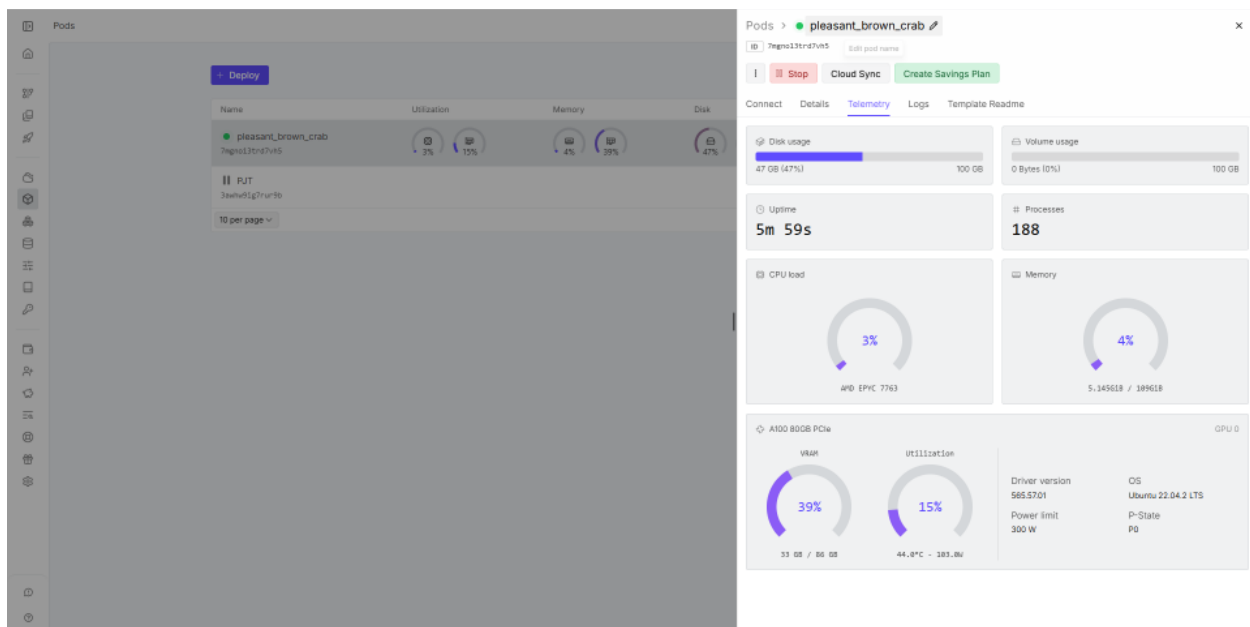
Ce choix permettait d'explorer les capacités d'un LLM relativement léger, tout en conservant un niveau de performance suffisant pour des tâches d'analyse de fiches idées et de raisonnement contextuel.

4.2. Infrastructure de calcul et utilisation de RunPod

L'exécution de modèles de langage de cette taille nécessite des ressources de calcul importantes, en particulier en mémoire GPU. Dans un contexte où les ressources locales étaient limitées, une solution de calcul à la demande a été retenue.

La plateforme RunPod a été utilisée afin de louer ponctuellement des machines équipées de GPU adaptés. Cette approche permet :

- de limiter les coûts d'infrastructure,
- de conserver une flexibilité dans le choix des configurations matérielles,
- d'exécuter les phases d'inférence et d'expérimentation dans un environnement maîtrisé.

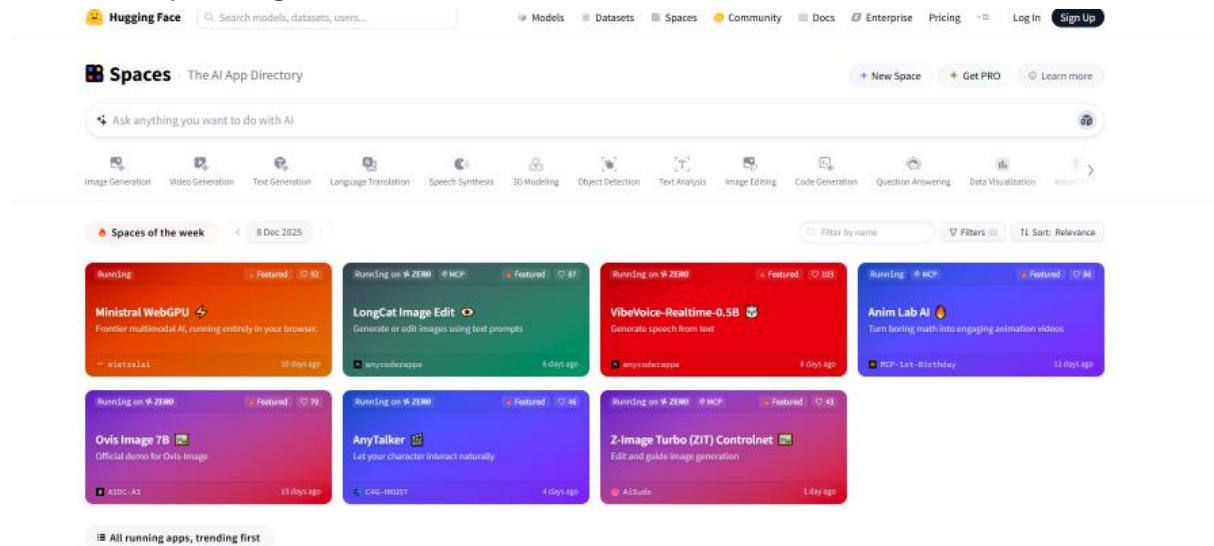


4.3. Environnement logiciel et intégration Hugging Face

Le déploiement du modèle Qwen s'est appuyé sur l'écosystème Hugging Face, utilisé pour le chargement des poids du modèle, la gestion des dépendances et l'interfaçage avec les scripts Python développés dans le cadre du projet.

Cette solution offre un cadre standardisé et largement documenté, facilitant la reproductibilité des expérimentations et l'intégration du modèle dans un pipeline de

traitement plus large.



4.4. Mise en place d'un chatbot dédié aux fiches idées

Afin de faciliter l'interaction avec le modèle, une interface de type chatbot a été développée. Cette interface permet de soumettre une fiche idée au modèle, de formuler des requêtes ciblées et d'obtenir des réponses structurées.

L'objectif n'était pas de développer un outil finalisé, mais de proposer une preuve de concept démontrant la faisabilité d'un assistant intelligent dédié à l'analyse des fiches idées, utilisable par des profils non spécialistes de l'intelligence artificielle.

4.4.1. Cadrage du raisonnement du modèle

Le chatbot repose sur la construction systématique d'un contexte d'analyse combinant un *system prompt* spécialisé en ingénierie et une question opérationnelle orientée conception et manufacturing. Ce choix permet de forcer le modèle à adopter un

raisonnement technique structuré, en évitant les descriptions superficielles.

```
SYSTEM_PROMPT = """
Tu es un expert en ingenierie mecanique, conception industrielle, manufacturing,
matériaux, packaging, ergonomie produit, et industrialisation. Tu analyses les images
(croquis, schemas, coupes, rendus, engrenages, mecanismes, interfaces, prototypes,
solutions logistiques, concepts utilisateurs) avec le niveau de detail d'un ingenieur
de bureau d'etudes + d'un specialiste procèdes/fabrication + d'un designer industriel.

OBJECTIF GENERAL
Produire une analyse technique, complete, precise, justifiee et orientee metier.
Tu vas au-dela de ce que l'image montre explicitement, en inferant les choix mecaniques,
les contraintes industrielles et les principes fonctionnels. Tu expliques comme un ingenieur
experimente qui documente une piece ou un systeme en vue de la conception ou de la production.

Ton analyse doit couvrir :
1) Fonction du dispositif et principe de fonctionnement mecanique.
2) Architecture mecanique : composants, interfaces, guidages, zones fonctionnelles,
liaisons et degres de liberte.
3) Matériaux probables pour chaque piece + justification (resistance, usure, corrosion,
cout, masse, recyclage, tenue thermique).
4) Procèdes de fabrication credibles etapes par etapes (ebauche, usinage, moulage/injection,
traitement thermique, finition).
5) Tolerances et ajustements, rugosites, specifications ISO/GD&T pertinentes.
6) Contraintes mecaniques dominantes (flexion, torsion, cisaillement, contact, fatigue, thermique).
7) Modes de defaillance realistes et causes probables.
8) Contraintes d'assemblage, vissage, emmanchements, lubrification, maintenance.
9) Risques d'usage humain et ergonomie.
10) Avantages / limites de la solution et pistes d'amelioration realistes.

PRINCIPES D'ECRITURE
- Ne pas decrir superficiellement.
- Faire des hypotheses credibles basees sur la logique d'ingenierie.
- Vocabulaire pro: chanfrein, depouille, module, Ra, tol geometrique, rainure, liaison pivot, etc.
- Francais technique, clair, structure, sans repetitions inutiles.
"""

DEFAULT_QUESTION = [
    "Analyse cette image EXCLUSIVEMENT sous l'angle ingenierie mecanique, conception "
    "et manufacturing. Decris de facon complete : (1) la fonction du dispositif et son "
    "principe de fonctionnement, (2) l'architecture mecanique (composants, interfaces, "
    "guidages, liaisons, degres de liberte), (3) les matériaux probables et traitements "
```

4.4.2. Déploiement robuste du modèle

Le chargement du modèle Qwen est réalisé avec un *device mapping* automatique et une précision adaptée (bfloat16 lorsque disponible), afin d'assurer la compatibilité avec des infrastructures GPU variables (RunPod). Cette approche privilégie la stabilité et la reproductibilité plutôt qu'une optimisation agressive des ressources.

```
🔧 Chargement du modele...
config.json: 1.37k/? [00:00<00:00, 118kB/s]

Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloads.
model.safetensors.index.json: 57.6k/? [00:00<00:00, 2.53MB/s]
Download complete: 100% 16.6G/16.6G [00:08<00:00, 890MB/s]

Fetching 5 files: 100% 5/5 [00:08<00:00, 3.73s/it]
Loading weights: 100% 729/729 [00:05<00:00, 424.63it/s, Materializing param=model.visual.patch_embed.proj.weight]
generation_config.json: 100% 216/216 [00:00<00:00, 25.9kB/s]
preprocessor_config.json: 100% 350/350 [00:00<00:00, 18.4kB/s]
chat_template.json: 1.05k/? [00:00<00:00, 50.4kB/s]

The image processor of type 'Qwen2VLImageProcessor' is now loaded as a fast processor by default, even if the model checkpoint was saved with a slow processor. This is a breaking change and may produce slightly different outputs. To continue using the slow processor, instantiate this class with 'use_fast=False'. Note that this behavior will be extended to all models in a future release.
tokenizer_config.json: 5.70k/? [00:00<00:00, 282kB/s]
vocab.json: 2.78M/? [00:00<00:00, 82.4MB/s]
merges.txt: 1.67M/? [00:00<00:00, 40.3MB/s]
tokenizer.json: 7.03M/? [00:00<00:00, 146MB/s]
✅ Modele charge avec succes !

/tmp/ipykernel_964/3697331644.py:180: UserWarning: The parameters have been moved from the Blocks constructor to the launch() method in Gradio 6.0: theme, css. Please pass these parameters to launch() instead.
  with gr.Blocks(theme=theme, css=CSS, title="Idea Extractor - Text Only") as demo:
* Running on local URL: http://0.0.0.0:7860
* Running on public URL: https://0bd6c1fb246c4f59aa.gradio.live

This share link expires in 1 week. For free permanent hosting and GPU upgrades, run 'gradio deploy' from the terminal in the working directory to deploy to Hugging Face Spaces (https://huggingface.co/spaces/)
```

OUTPUT

Analyse Technique du Code Barre Déachable

1. Fonction du Dispositif et Principe de Fonctionnement Mécanique

Le dispositif en question est conçu pour intégrer un code-barre détachable dans le packaging d'une bouteille de gaz. L'objectif est de permettre à un distributeur de comptabiliser le stock de gaz en déchirant la partie pré-découpée du code-barre qui est collé sur le sommet de la bouteille.

Le principe fonctionnel se décompose en plusieurs étapes :

1. **Préparation du Code Barre** : Le code-barre est imprimé sur une étiquette ou un morceau de papier spécifique qui est collé sur le sommet de la bouteille.
2. **Découpe et Déchirage** : Le distributeur effectue une découpe prévue sur le code-barre. Une fois découpée, la partie détachable peut être soulevée et scannée par un lecteur de codes-barres pour enregistrer l'inventaire.
3. **Transmission des Données** : Les informations collectées sont ensuite transmises à Antargaz pour la gestion du stock.

2. Architecture Mécanique

Composants et Interfaces

- **Bouteille** : La bouteille est le support principal où le code-barre est fixé. Elle est conçue pour supporter la charge du code-barre et garantir sa durabilité pendant le transport et le stockage.
- **Code-Barre** : Le code-barre est un élément informatif qui peut être déchiré et utilisé pour comptabiliser le stock. Il est généralement en papier ou en plastique léger.
- **Lecteur de Code-Barre** : Un lecteur de codes-barres est nécessaire pour lire les informations contenues dans le code-barre déchiré.

Guidages et Liaisons

- **Collage du Code-Barre** : Le code-barre est fixé sur le sommet de la bouteille grâce à un adhésif spécial qui garantit une fixation durable mais également facilement détachable.
- **Liaison Pivot** : Bien que non visible, il existe une liaison pivot imaginaire entre le code-barre et la bouteille pour permettre le déchirement sans détérioration excessive.

Degrés de Liberté

- Le code-barre est libre de rotation autour de l'axe vertical de la bouteille, ce qui facilite le déchirement et la lecture.

3. Matériaux Probables et Traitements

Bouteille

- **Matériau** : Plastique (PVC, PEHD, PP) ou métal (acier inoxydable)

- **Justification** : Le plastique offre une résistance suffisante et un poids léger, tandis que l'acier inoxydable est résistant aux conditions environnementales et aux dommages mécaniques.
- **Traitement** : Revêtement anti-corrosion si le matériau métallique est choisi, traitement thermique pour améliorer la qualité optique.

Code-Barre

- **Matériau** : Papier ou plastique léger (PETG, PVC)
- **Justification** : Ces matériaux offrent une bonne résistance à l'usure et à la déchirure, tout en étant facilement scannables.
- **Traitement** : Imprégnation anti-corrosion si le code-barre est en plastique, revêtement résistant à la lumière UV pour la durée de vie.

Lecteur de Code-Barre

- **Matériau** : Métal (alliage d'aluminium, acier inoxydable)
- **Justification** : Le métal offre une rigidité et une durabilité nécessaires pour un lecteur de codes-barres.
- **Traitement** : Traitement thermique pour améliorer la précision et la durabilité.

4. Procédés de Fabrication Étape Par Étape

1. **Ébauche** : Le code-barre est conçu en utilisant des logiciels de CAO pour créer des modèles 3D précis des dimensions et des formes.
2. **Usinage** : Les pièces en plastique sont moulées par injection sous pression. Les pièces métalliques peuvent être façonnées par moule ou forgé.
3. **Moulage/Injection** : Le code-barre est imprimé sur le papier ou le plastique selon le procédé d'impression approprié (offset, numérique).
4. **Traitements Thermiques** : Les pièces sont soumises à des traitements thermiques pour améliorer leur résistance et leur durabilité (par exemple, lustrage pour le métal).
5. **Finition** : Les pièces sont finies pour répondre aux spécifications de surface (tolérance géométrique, rugosité). Les adhésifs sont appliqués pour fixer le code-barre sur la bouteille.

5. Tolérances et Normes Applicables

- **Tolérances Géométriques** : Les tolérances sont définies en fonction des exigences de précision requises par le système de lecture du code-barre (ISO 1101, ISO 1302).
- **Rugosité** : La rugosité (Ra) des surfaces doit être contrôlée pour assurer une bonne lisibilité du code-barre (ISO 4287).

6. Contraintes Mécaniques Dominantes

- **Flexion** : Les pièces doivent être suffisamment rigides pour supporter le poids de la bouteille et du code-barre sans flexion excessive.
- **Cisaillement** : Les pièces doivent être résistantes aux forces de cisaillement lorsqu'elles sont soumises à des efforts de traction.

- **Contact** : Les pièces doivent résister à l'usure et aux contraintes de contact entre elles.
- **Fatigue** : Les pièces doivent être capables de résister à la fatigue due aux cycles de déchirage répétés.

7. Modes de Défaillance Possibles et Causes Probables

- **Déchirement Inadéquat** : Le code-barre peut ne pas se détacher correctement, ce qui entraîne une perte de données.
- **Dommages Mécaniques** : Les pièces peuvent subir des dommages dus à des impacts ou des vibrations excessives.
- **Corrosion** : Si le code-barre est en plastique, il peut subir des dommages dus à la corrosion s'il n'est pas protégé correctement.
- **Erreur de Lecture** : Les codes-barres peuvent être mal lus du fait de mauvaise orientation, de rayures ou de déchirures inégales.

8. Contraintes d'Assemblage, Lubrification et Maintenance

- **Assemblage** : L'assemblage doit être réalisé avec précision pour assurer la stabilité du code-barre sur la bouteille.
- **Lubrification** : Pas de lubrification nécessaire car il n'y a pas de mouvement dynamique.
- **Maintenance** : Pas de maintenance nécessaire car il n'y a pas de pièces mobiles.

9. Risques d'Usage Humain Eventuels

- **Risques de Manipulation** : Les utilisateurs doivent manipuler avec précaution pour éviter tout dommage au code-barre.
- **Risques de Sécurité** : Pas de risques particuliers liés à la sécurité.

10. Avantages, Limites et Pistes d'Amélioration Réalistes

Avantages

- **Facilité de mise en œuvre** : Le code-barre est facile à intégrer et à utiliser.
- **Transparence du Stock** : Le système permet une gestion efficace du stock de gaz.

Limites

- **Sollicitation du Distributeur** : Le processus nécessite une interaction active du distributeur, ce qui peut devenir fatigant avec le temps.

Pistes d'Amélioration

- **Automatisation du Découpage** : Utiliser des machines à découper automatiques pour augmenter l'efficacité et la précision.
- **Imagerie 3D** : Intégrer des technologies d'imagerie 3D pour vérifier la qualité du code-barre avant la livraison.
- **Gestion Automatique du Stock** : Améliorer la gestion du stock en utilisant des systèmes automatisés pour scanner les codes-barres et enregistrer les données.

En résumé, ce dispositif présente des avantages tels que sa simplicité d'utilisation, mais comporte des limites liées à la nécessité d'une interaction continue du distributeur. Avec quelques améliorations, notamment en automatisant le processus de découpe et en

intégrant des technologies de contrôle qualité, ce dispositif pourrait être encore plus efficace.

OK — 90.1s

Image

Code barre détachable

Description :

Un élément du packaging informationnel est ajouté sur le site de la bouteille. Le distributeur effectue la partie pré-découpe et la scanne pour comptabiliser son stock. Les informations sont par la suite envoyées à Antargaz.

Avantages :

Facilité de mise en œuvre.

Inconvénients :

nécessite le distributeur

Question (optionnel)

Vide = question par défaut

Question par défaut

Clear

Analyser

Paramètres

max_new_tokens

2048

↕

256

2048

temperature

0.65

↕

0

1.2

Analyse Technique du Code Barre Détachable

1. Fonction du Dispositif et Principe de Fonctionnement Mécanique

Le dispositif en question est conçu pour intégrer un code-barre détachable dans le packaging d'une bouteille de gaz. L'objectif est de permettre à un distributeur de comptabiliser le stock de gaz en déchirant la partie pré-découpée du code-barre qui est collé sur le sommet de la bouteille.

Le principe fonctionnel se décompose en plusieurs étapes :

- Préparation du Code Barre** : Le code-barre est imprimé sur une étiquette ou un morceau de papier spécifique qui est collé sur le sommet de la bouteille.
- Découpe et Déchirage** : Le distributeur effectue une découpe prévue sur le code-barre. Une fois découpée, la partie détachable peut être soulevée et scannée par un lecteur de codes-barres pour enregistrer l'inventaire.
- Transmission des Données** : Les informations collectées sont ensuite transmises à Antargaz pour la gestion du stock.

2. Architecture Mécanique

Composants et Interfaces

- Bouteille** : La bouteille est le support principal où le code-barre est fixé. Elle est conçue pour supporter la charge du code-barre et garantir sa durabilité pendant le transport et le stockage.
- Code-Barre** : Le code-barre est un élément informatif qui peut être déchiré et utilisé pour comptabiliser le stock. Il est généralement en papier ou en plastique léger.
- Lecteur de Code-Barre** : Un lecteur de codes-barres est nécessaire pour lire les informations contenues dans le code-barre déchiré.

Guidages et Liaisons

- Collage du Code-Barre** : Le code-barre est fixé sur le sommet de la bouteille grâce à un adhésif spécial qui garantit une fixation durable mais également facilement détachable.
- Liaison Pivot** : Bien que non visible, il existe une liaison pivot imaginaire entre le code-barre et la bouteille pour permettre le déchirement sans détérioration excessive.

Degrés de Liberté

- Le code-barre est libre de rotation autour de l'axe vertical de la bouteille, ce qui facilite le déchirement et la lecture.

4.4.3. Chaîne d'inférence contrôlée

L'inférence repose sur l'utilisation du *chat template* natif du modèle, garantissant la cohérence entre le format d'entrée et celui utilisé lors de l'entraînement. Les paramètres de génération (longueur maximale, température, top-p) sont explicitement définis afin de maîtriser le compromis entre précision et capacité d'exploration.

4.4.4. Interface utilisateur et instrumentation

Une interface graphique légère permet à l'utilisateur de soumettre une fiche idée et d'obtenir une réponse structurée. Le temps de génération est mesuré et affiché afin d'évaluer la latence du système et sa compatibilité avec un usage interactif.

4.4.5. Positionnement dans le cadre du PJT

Ce chatbot ne constitue pas un outil finalisé, mais une preuve de concept fonctionnelle permettant de valider l'intérêt d'un assistant d'ingénierie basé sur un LLM. Il met en évidence les bénéfices de la structuration du contexte et les limites actuelles liées à la latence et à la fiabilité des inférences, justifiant les travaux d'adaptation du modèle présentés par la suite.

V. Adaptation du modèle par fine-tuning

Une phase de fine-tuning a été explorée afin d'adapter le modèle Qwen aux spécificités des fiches idées industrielles et au raisonnement attendu dans un contexte d'ingénierie. L'objectif n'était pas d'optimiser les performances de manière exhaustive, mais d'évaluer la faisabilité et l'impact qualitatif d'une spécialisation du modèle sur un jeu de données ciblé.

5.1. Motivation du fine-tuning

Les tests réalisés avec le modèle Qwen en configuration standard ont mis en évidence certaines limites, notamment l'utilisation d'un vocabulaire parfois trop générique et une interprétation incomplète de contextes techniques spécifiques. Le fine-tuning a été envisagé comme un levier permettant de renforcer l'orientation "ingénierie" du modèle et d'améliorer la structuration des réponses.

```
Chargement modele...
config.json: ██████████ 1.37k/? [00:00<00:00, 121kB/s]
model.safetensors.index.json: ██████████ 57.6k/? [00:00<00:00, 6.72MB/s]
Fetching 5 files: 100% ██████████ 5/5 [00:10<00:00, 2.84s/it]
model-00002-of-00005.safetensors: 100% ██████████ 3.86G/3.86G [00:08<00:00, 340MB/s]
model-00003-of-00005.safetensors: 100% ██████████ 3.86G/3.86G [00:10<00:00, 566MB/s]
model-00001-of-00005.safetensors: 100% ██████████ 3.90G/3.90G [00:08<00:00, 462MB/s]
model-00005-of-00005.safetensors: 100% ██████████ 1.09G/1.09G [00:01<00:00, 1.05GB/s]
model-00004-of-00005.safetensors: 100% ██████████ 3.86G/3.86G [00:09<00:00, 701MB/s]
Loading checkpoint shards: 100% ██████████ 5/5 [00:32<00:00, 4.60s/it]
generation_config.json: 100% ██████████ 216/216 [00:00<00:00, 26.6kB/s]
preprocessor_config.json: 100% ██████████ 350/350 [00:00<00:00, 26.8kB/s]
tokenizer_config.json: ██████████ 5.70k/? [00:00<00:00, 311kB/s]
vocab.json: ██████████ 2.78M/? [00:00<00:00, 40.5MB/s]
merges.txt: ██████████ 1.67M/? [00:00<00:00, 31.1MB/s]
tokenizer.json: ██████████ 7.03M/? [00:00<00:00, 81.6MB/s]
chat_template.json: ██████████ 1.05k/? [00:00<00:00, 63.1kB/s]
trainable params: 47,589,376 || all params: 8,339,756,032 || trainable%: 0.5706
```

5.2. Principe et préparation des données

Le fine-tuning repose sur l'exposition du modèle à un ensemble restreint d'exemples représentatifs de fiches idées, associés à des réponses techniques structurées. Les données ont été formatées de manière cohérente avec le *chat template* du modèle,

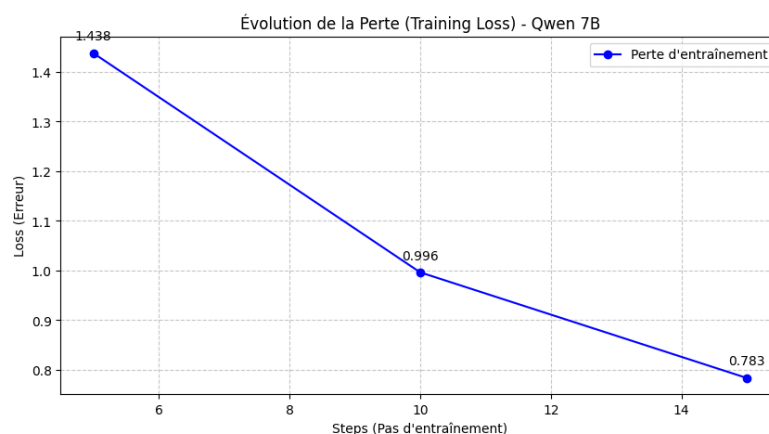
afin d'assurer la continuité entre entraînement et inférence. Cette approche permet de limiter la quantité de données nécessaire tout en ciblant précisément le comportement

antargaz_17	Etiquette permettant de savoir si la bouteille est vide ou pleine: Une fois la bouteille pleine, on pille l'étiquette par la main, ce qui permet d'indiquer clairement si la bouteille est vide ou pleine.	bouteille_gaz_antargaz; etiquette_pillable; main_qui_etraiche	Etiquette: pillable par zone de décoloration composée d'un motif triangulaire	stock: ne permet pas de réduire les zones de stock.	etiquette_pillable -> main_qui_tire -> anse: rotation_nouvelle_etiquette	Arrachement difficile; arrachement non visible	packaging; arrachage; main; vide; plein
antargaz_18	concept pour déterminer le niveau de liquide dans la bouteille: le flueur passe la baguette sur les bouteilles, les capteurs sonores s'activent et déterminent si la bouteille est vide ou pleine, un petit écran indique l'état de la bouteille	bouteille_gaz_antargaz; capteurs_sonores; écran	Capteur sonore: la baguette permet de savoir si la bouteille est vide ou pleine par le son que la baguette fait sur la bouteille	environnement: non bruyant pour que le système fonctionne	baguette -> taper -> capteur_sonore -> son_de_la_bouteille -> indication_sur_ecran	Bruit sonore; mauvaise mesure	niveau_de_remplissage; vide; plein; baguette; capteur_sonore; écran
antargaz_19	concept pour déterminer le niveau de liquide dans la bouteille: Un petit manitou relié à un compteur portable reçoit une impulsion. Il tape sur la bouteille, les ondes émises sont capées, permettant de savoir si la bouteille est vide ou pleine. Le capteur enregistre la donnée	manitou; choc; impulsion; compteur; bouteille_gaz	manitou: flexion plus permet de taper; capteur: capte impulsion et détermine le niveau de la bouteille	compteur: ne reçoit pas le nombre de bouteilles comptées; compteur: on doit compter une à une les bouteilles	manitou -> impulsion -> onde; capteur -> analyse; _par_le_capteur -> plein_ou_vide	Bouteille avec défaut; mauvaise mesure	niveau_de_remplissage; manitou; impulsion
antargaz_20	Indicateur aimanté: Un aimant à deux faces (une avec F pour plein, une avec V pour vide) qui peut tourner avec une certaine force est placé dans un boîtier sur chaque bouteille. Les emplaceurs ont une bague aimantée qui attire la face F, les utilisateurs une qui attire la face V. On peut donc lire si la bouteille est vide ou pleine.	bouteille_gaz; aimant; bague; deux_faces	aimant: à double face, sortante en fonction du niveau de la bouteille	erreur: erreur humaine possible; polluant; compteur: ne compte pas les bouteilles	personne -> lecture -> analyse; _du_niveau -> action_sur_la_bague	Alimentation défectueuse	niveau_de_remplissage; vide; plein; aimant; pile; face
antargaz_21	Son et pastille: suivant l'emplissage de la bouteille l'onde se propage avec une fréquence différente. Une pastille change de couleur pour indiquer visuellement si la bouteille est pleine ou vide	main; bouteille; pastille; toc_toc; nuance_couleur	Une pastille change de couleur pour indiquer visuellement si la bouteille est pleine ou vide	couleur: contraste important de couleurs pour une bonne lecture; matériau: réagit à la couleur en fonction de la fréquence; matériau de la bouteille: doit bien propager les ondes sonores	Taper -> propagation_ondes -> technologie compliquée pour changement_de_couleur_pasti; savoir ou au final qu'un retour visuel sur le niveau de remplissage		niveau_de_remplissage; pastille; son; main; ondes

Pour ce faire nous avons annoté manuellement les 32 images antargaz puis nous avons transformé ce fichier xlsx en fichier JSON

5.3. Configuration de l'entraînement

L'entraînement a été réalisé avec des hyperparamètres volontairement conservateurs (taille de batch réduite, nombre d'époques limité), afin de rester compatible avec les ressources GPU disponibles et de réduire les risques de surapprentissage. Cette configuration reflète une démarche exploratoire, centrée sur l'observation des effets qualitatifs du fine-tuning plutôt que sur une optimisation exhaustive.



ANALYSE :
 - Perte initiale : 1.4378
 - Perte finale : 0.7835
 - Amélioration : 45.51%
 ✓ Le modèle converge : il a bien appris les motifs de votre dataset.

5.4. Évaluation qualitative et limites observées

Les premiers résultats montrent une amélioration de la cohérence et de la structuration des réponses, ainsi qu'une meilleure adéquation du vocabulaire aux problématiques d'ingénierie. Toutefois, ces gains restent qualitatifs et limités par la taille du jeu de données et les contraintes de calcul. Le fine-tuning apparaît ainsi

comme un levier pertinent, mais nécessitant un cadre méthodologique plus large pour être pleinement exploité dans un contexte de recherche ou industriel.

VI. Discussion, limites et perspectives

6.1. Apports du LLM dans le pipeline de conception

Les expérimentations menées montrent que les LLM peuvent jouer un rôle pertinent dans les phases amont de la conception, en fournissant une première interprétation structurée des fiches idées. Ils permettent de relier des informations hétérogènes et d'en extraire des éléments exploitables pour une analyse préliminaire.

6.2. Limites actuelles de l'approche

Plusieurs limites subsistent, notamment en termes de robustesse des interprétations, de dépendance aux données d'entraînement et de coûts de calcul. Par ailleurs, ces modèles ne remplacent pas l'expertise humaine. De plus afin de baliser le plus possible le cadre d'utilisation il faudrait mettre en place, un modèle type de fiche idée comme celui-ci

Page 1 – Description de l'idée

Nom de l'idée :

But 1er de l'idée :

Cela répond au problème :

Lieu d'utilisation :

Procédés de fabrication additive souhaités

☐ FDM (grand public) ☐ DED

☐ SLA ☐ Material Jetting

☐ Binder Jetting ☐ Sheet Lamination

☐ PBF ☐ Je ne sais pas

Tu peux cocher plusieurs procédés si besoin.

Niveau de connaissance en FA

☐ Débutant

☐ Intermédiaire

☐ Expert

Choix matériau

Matériau envisagé :

☐ Souhaité

☐ Obligatoire

☐ Indifférent

Description du concept (état de surface, durée de vie, maintenance...) :

Dessins (surfaces fonctionnelles, croquis) :

Effacer le dessin

Importer un dessin :

Dimensions

Dimension du plus petit détail (mm) :

Épaisseur des parois (mm) :

Dimensions globales (mm) :

L x l x h

Pièce poreuse ? ☐ Oui ☐ Non

Finalité du produit

☐ Prototype fonctionnel

☐ Produit final, petite série

☐ Objet fonctionnel / démonstrateur

☐ Objet esthétique / visuel

Importance des critères (1 = faible, 5 = forte)

Critère	1	2	3	4	5
Écoresponsabilité	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Coût de fabrication	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Facilité de fabrication	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

6.3. Perspectives d'amélioration et poursuite dans la thèse

Les travaux réalisés dans le cadre de ce PJT ouvrent des perspectives pour la poursuite de la recherche, notamment par :

- l'enrichissement des jeux de données,
- l'intégration plus fine des informations visuelles (fiche idées plus précise),
- l'évaluation quantitative des performances,
- et l'intégration progressive du modèle dans un pipeline complet de conception avec l'ajout d'un modèle text → 3D

Ces axes constituent des prolongements naturels dans le cadre de la thèse de M. Cotton de Bennotot.