Notes de Certification Google AI Leader

1. Concepts Fondamentaux de l'Intelligence Artificielle

1.1 Intelligence Artificielle Traditionnelle vs Foundation Models

IA Traditionnelle :

- Entraînée pour réaliser une tâche spécifique comme filtrer les spams dans une boîte de messagerie
- Performance limitée au domaine d'entraînement
- Nécessite un réentraînement complet pour chaque nouvelle tâche

Foundation Models :

- Entraînement polyvalent sur des données énormes
- Permet d'adapter le modèle à des tâches spécifiques par la suite
- Exemples de Foundation models :
    - **Gemini** (multimodal)
    - **Imagen** (Image et description d'images)
    - **Chirp** (audio, reconnaissance vocale, assistant vocal, traduction)

1.2 Types d'Apprentissage Machine

Apprentissage Supervisé :

- La data est étiquetée (on décrit sa signification : chat/chien, commentaire positif/négatif/neutre...)
- Permet à la machine de construire des relations et faire des prévisions
- Exemple : utiliser des capteurs pour voir si une machine en usine risque de tomber en panne

Apprentissage Non Supervisé :

- Utile quand on n'a pas de data étiquettée ou qu'on ne peut donner des instructions précises
- Exemple : analyser les données bancaires pour détecter des risques de fraudes lorsque des transactions sortent de la norme

Apprentissage par Renforcement :

- Exemple : analyser les données clients pour proposer des recommandations pour maximiser les ventes et l'engagement

1.3 Deep Learning et Réseaux de Neurones

- Les réseaux neuronaux peuvent ingérer en même temps des données étiquetées et non étiquetées
- Ils apprennent les concepts fondamentaux puis généralisent avec de nouveaux exemples
- Les LLMs sont des foundation models, mais les Diffusion models aussi (pour générer des images, de l'audio et de la vidéo en transformant du non structuré en structuré)

1.4 Caractéristiques Requises pour les Données ML

Pour que le machine learning fonctionne, il faut veiller à ces caractéristiques de la data :

- **Fiables**
- **Volumineuses**
- **Représentatives**
- **Consistance** (ex : parfois, mal étiquetées)
- **Pertinence** (par rapport au but, l'output visé)

Par ailleurs, lors de l'entraînement, la data doit être :

- **Accessible**
- **Au prix souhaité**
- **Dans un format compréhensible** pour l'IA

1.5 Types de Données d'Entreprise

Data Structurée :

- Par exemple collectée lors d'une inscription, lors d'un achat
- Tout est disponible dans un fichier type Excel

Data Non Structurée :

- Textes, posts sur réseaux sociaux, emails, photos, messages vocaux, vidéo, musique...

2. Historique et Évolution de l'IA chez Google

2.1 Timeline Complète

- **2006** : Google Translate utilise le machine learning
- **2014** : Acquisition de DeepMind par Google

- **2015** :
    - Google Search utilise RankBrain pour adapter les SERP à l'utilisateur
    - TensorFlow, framework open source de machine learning utilisé en reconnaissance vocale
- **2016** : TPUs (Tensor Processing Units) : puces silicium spécialement créées pour le machine learning
- **2017** : Publication scientifique de Google "Attention is All You Need" qui introduit une nouvelle architecture neuronale artificielle particulièrement bien adaptée au traitement du langage : les transformers
- **2018** : AI Principles guidelines
- **2019** : BERT, nouvelle technique pour l'entraînement en NLP pour mieux comprendre les demandes de l'utilisateur
- **2023** : Bard, puis Gemini

2.2 Modèles Open Source

- **Gemma** : modèles open légers issus des mêmes modèles que Gemini
- **TensorFlow** : framework open source

3. Plateformes et Outils Google Cloud AI

3.1 Vertex AI - Plateforme Unifiée

Définition : Plateforme qui unifie les capacités de machine learning de Google. Permet de construire, entraîner et déployer des modèles ML et des applications IA. Pour de l'IA multimodale, LLM utilise Model Garden.

Model Garden :

- Banque de 160 modèles provenant de Google et d'autres IA, notamment open source
- Peut être déployé dans nos propres applications
- Possibilité d'entraîner avec nos propres données
- On peut les entraîner en customisant tout (PyTorch, TensorFlow…) ou bien utiliser AutoML pour créer et entraîner un modèle sans trop de connaissances techniques

MLOps in Vertex AI : Permet de gérer le workflow du processus de machine learning

3.2 Vertex AI Search

Permet aux développeurs d'intégrer des fonctionnalités de recherches avancées dans une application utilisant de l'IA générative (style chatbots clients) en allant fouiller dans des bases de données ou des documents spécialisés.

Applications :

- Utiliser l'IA pour réaliser des recherches dynamiques au sein d'une entreprise (par exemple dans toutes les données d'une entreprise, l'IA donne une réponse à une recherche ou fournit un résumé sur ce qu'elle a trouvé)
- Systèmes de recommandations

3.3 Vertex AI Studio vs Google AI Studio

Aspect	Google AI Studio (Google account)	Vertex AI Studio (Google Cloud)
Focus	Interface simple pour explorer les capacités de Gemini, expérimenter avec les paramètres, et générer différents formats de texte créatif	Environnement complet et robuste pour construire, entraîner et déployer des modèles ML qui fait partie de la plateforme Google Cloud Vertex AI

Utilisation :

- **Google AI Studio** : pour générer une clé API
- **Vertex AI Studio** : pour régler les autorisations et les authentifications

3.4 Cycle de Vie d'un Projet ML en Entreprise

1. Data Gathering :

- **Google Cloud Pub/Sub**
- **Cloud Storage** (unstructured)
- **Cloud SQL et Cloud Spanner** (structured)

2. Prepare Data :

- **BigQuery** pour analyser la data et vérifier sa qualité, réparer la data
- **Data Catalog** pour la gouvernance des données (trouver les données dont on a besoin)

3. Train your Model :

- **Vertex AI**, puis Déploiement (on commence à s'en servir pour réaliser des actions, du type des prédictions)

4. Management Model :

- Mise à jour, performance tracking

- Stockage des paramétrages et des modèles
- Automatiser (Vertex Pipeline)

3.5 Sécurité et Gestion des Accès

IAM (Identity and Access Management) : Pour protéger l'accès au Google Cloud

SAIF (Secure AI Framework) : Ensemble d'outils et de principes pour construire des systèmes IA sécurisés

Mandiant : Protège l'IA contre les cyberattaques

4. APIs et Outils Pré-construits

4.1 Pre-built AI APIs

Traitement de la Parole :

- **STT (Speech-to-Text)**
- **TTS (Text-to-Speech)**
- **Translation**
- **Document Translation**

Vision et Vidéo :

- **Cloud Vision**
- **Cloud Video Intelligence**
- **Veo** : le modèle de Google spécialisé en vidéo

Traitement du Langage :

- **Natural Language API**

4.2 Edge AI - LiteRT (Lite Runtime)

Permet d'intégrer des modèles IA dans des appareils, sans passer par le cloud, afin de :

- **Réduire la latence** et les problèmes de connexion
- **Garantir une meilleure sécurité** des données
- **Fonctionnement IA offline**

Modèle de référence : Gemini Nano (par exemple dans les smartphones)

Exemples d'applications :

- Call Notes et Pixel Recorder dans les smartphones Pixel

- Disponible pour les développeurs Android à travers le **AI Edge SDK**

4.3 Intégrations Google Workspace

Gemini for Google Cloud :

- Ne utilise PAS vos prompts ou les réponses de Gemini pour entraîner les modèles
- Vient avec les protections d'entreprise standard de Google Cloud
- **Intégrations :** BigQuery, Google Databases, Looker, Colab Enterprise, Code Assist

Compléments Workspace à inclure :

- **Infosaved** (fonctionnalité Gemini)
- **Les Gems** (assistants personnalisés)
- **Paramètre de rétention dans Gemini Advanced**

5. Techniques de Prompting et Optimisation

5.1 Techniques de Base

Zero-shot Prompting : Demander à un foundation model de compléter une tâche sans exemples préalables, en se basant uniquement sur ses connaissances existantes.

One-shot Prompting : Montrer au foundation model juste un exemple, lui permettant d'apprendre et d'appliquer cette connaissance à des situations similaires.

Few-shot Prompting : Fournir au foundation model plusieurs exemples pour apprendre, ce qui l'aide à mieux comprendre la tâche et améliorer ses performances.

Role Prompting : Technique utilisée pour guider le comportement des LLMs en leur assignant un rôle ou persona spécifique. Cela peut être n'importe quoi, d'un analyste business ou acteur shakespearien à un agent de service client utile. En instruisant le modèle à adopter un rôle particulier, vous influencez le style, le ton et le focus de ses réponses.

Prompt Chaining : Technique puissante pour obtenir des résultats plus complexes et nuancés des grands modèles de langage comme Gemini. C'est comme avoir une conversation avec l'IA où chaque réponse s'appuie sur la précédente, menant à un résultat plus sophistiqué et raffiné.

5.2 Techniques Avancées

Chain of Thought (CoT) :

- Pensez au CoT comme un moyen de rendre les LLMs encore plus intelligents en leur apprenant à penser étape par étape, comme un humain le ferait

- Au lieu de juste donner un prompt au LLM et attendre une réponse, vous le guidez à travers le processus de raisonnement
- Vous fournissez des exemples de comment résoudre des problèmes similaires, montrant les étapes impliquées
- C'est similaire à enseigner à un étudiant à penser à voix haute
- **Focus sur le raisonnement interne**, guidant le LLM à travers une chaîne de pensée

ReAct (Reasoning and Acting) :

- ReAct, qui signifie "reasoning and acting", c'est comme donner au LLM un cerveau et une paire de mains
- Permet au LLM non seulement de penser à un problème mais aussi de prendre des actions pour le résoudre

Pourquoi ReAct est important :

- **Résolution de problèmes dynamique :** permet aux LLMs de s'attaquer à des tâches complexes nécessitant d'interagir avec des ressources externes et de s'adapter à de nouvelles informations
- **Réduction des hallucinations :** en ancrant le raisonnement du LLM dans des données du monde réel, ReAct peut aider à réduire le risque de générer des informations incorrectes ou absurdes
- **Confiance accrue :** la capacité de voir le processus de raisonnement du LLM et comment il interagit avec des sources externes rend ses réponses plus transparentes et dignes de confiance

Composants clés de ReAct :

- **Think :** Le LLM génère une pensée sur le problème, similaire au CoT
- **Act :** Le LLM décide quelle action prendre, comme rechercher sur le web, accéder à une base de données, ou utiliser un outil spécifique ; le LLM spécifie l'input pour l'action, comme une requête de recherche ou commande de base de données
- **Observe :** Le LLM reçoit un retour de l'action, comme des résultats de recherche ou entrées de base de données
- **Respond :** Le LLM génère une réponse, qui pourrait impliquer de fournir une réponse à l'utilisateur, prendre d'autres actions, ou formuler une nouvelle pensée pour la prochaine itération

ReAct en action :

- **Réponse aux questions :** Les LLMs peuvent utiliser ReAct pour accéder à des sources de connaissances externes et répondre aux questions plus précisément
- **Vérification des faits :** Les LLMs peuvent vérifier des affirmations en recherchant des preuves en ligne
- **Prise de décision :** Les LLMs peuvent utiliser ReAct pour rassembler des informations et prendre des décisions éclairées dans des environnements interactifs

Différence CoT vs ReAct :

- **CoT se concentre sur le raisonnement interne**, guidant le LLM à travers une chaîne de pensée
- **ReAct se concentre sur l'interaction externe**, permettant au LLM de rassembler des informations et prendre des actions dans le monde réel

Metaprompting : Le metaprompting consiste à créer des prompts qui guident l'IA à générer, modifier ou interpréter d'autres prompts. C'est un outil puissant pour interagir avec l'IA, permettant une création et interprétation de prompts plus dynamique, flexible et adaptable. C'est une technique clé pour débloquer le plein potentiel des grands modèles de langage.

5.3 Paramètres d'Échantillonnage (Sampling Parameters)

Pour interagir avec le modèle, on peut utiliser les sampling parameters :

Token Count :

- Un token équivaut approximativement à quatre caractères en anglais
- Cent tokens représentent environ soixante à quatre-vingts mots

Temperature :

- Une température plus élevée rend la sortie plus aléatoire et imprévisible
- Une température plus basse la rend plus focalisée, déterministe et répétable

Top-p :

- "Top-p" représente la probabilité cumulative des tokens les plus probables considérés pendant la génération de texte
- Une valeur top-p plus basse mène à des réponses plus focalisées (seulement les tokens les plus probables)

- Une valeur plus élevée permet plus de diversité (s'étendant aux tokens de probabilité plus basse aussi)

Autres paramètres :

- **Safety settings** (filtres contre les biais)
- **Output length**

6. Agents Conversationnels et Architecture

6.1 Types d'Agents

Agents Déterministes :

- Chemins d'actions prédéfinis
- Haut degré de contrôle et prévisibilité
- Arbres de décision
- Exemple : une messagerie vocale automatique, "appuyez sur 1..."
- Peut utiliser de la reconnaissance vocale mais pas de modèle génératif

Agents Génératifs :

- Utilisent des modèles génératifs pour des réponses plus flexibles

Agents Modernes (Hybrides) :

- À la fois déterministes et génératifs
- Combinent structure et flexibilité

6.2 Composants d'un Agent Moderne

Trois composantes principales :

1. **Model** (modèle de base)
2. **Tools** (API, base de données)
3. **Reasoning Loop** (boucle de raisonnement)

Reasoning Loop : Fonctionnement itératif des agents : observe, interprète, planifie, agit

Tools : Permettent à l'agent d'interagir avec son environnement

- **Extensions** (pour communiquer avec d'autres APIs extérieures)
- **Fonctions** (une action spécifique)
- **Data stores**

- **Plugins**

6.3 Agents Multiples

Situations où vous pourriez avoir besoin de "multiples agents" :

- Ces systèmes utilisent plusieurs agents, chacun potentiellement spécialisé pour une tâche spécifique
- Créent des applications plus efficaces et capables
- **Exemple :** une app de réservation de voyage pourrait utiliser un agent pour chercher des vols, un autre pour trouver des hôtels, et un troisième pour suggérer des attractions locales
- Ces agents peuvent travailler indépendamment ou interagir entre eux pour fournir une expérience de réservation fluide
- Cette approche modulaire améliore non seulement l'efficacité mais permet aussi plus de flexibilité et scalabilité

7. RAG et Techniques de Grounding

7.1 Grounding (Ancrage)

Permet de limiter les hallucinations en liant les résultats à des ressources spécifiques pour vérifier ses réponses, donne donc le contexte dont l'IA a besoin.

Techniques pour limiter les hallucinations :

- Qualité du prompt
- Fine-tuning
- Principe HITL (Human In The Loop) pour :
    - Réaliser de la modération de contenu
    - Encadrer les réponses si elles ont des conséquences importantes (finance, santé, juridiction)
    - Prévisioner le résultat
    - Surveiller les résultats

7.2 RAG (Retrieval-Augmented Generation)

Une technique de grounding puissante qui implique :

1. Récupérer des informations pertinentes : Le modèle IA récupère d'abord des informations pertinentes d'une vaste base de connaissances (comme une base de données, un ensemble de documents, ou même tout le web). Ce processus de

récupération est souvent alimenté par des techniques sophistiquées, comme la recherche sémantique ou les bases de données vectorielles.

2. Générer la sortie : Le modèle utilise ensuite ces informations récupérées pour générer la sortie finale. Cela peut être n'importe quoi, de répondre à une question à écrire une histoire créative.

Avantages du RAG :

- **Amélioration de l'explicabilité et transparence :** RAG peut augmenter la transparence et la confiance dans le système IA en montrant les sources spécifiques utilisées pour générer la sortie, s'assurant que les affirmations peuvent être vérifiées pour leur exactitude
- **Réponses basées sur les sources :** Chaque réponse et insight fourni est directement ancré dans vos sources uploadées

Exemple avec NotebookLM :

- **Réponses basées sur les sources :** Toute réponse et insight fourni par NotebookLM est directement ancré dans vos sources uploadées. Cela assure l'exactitude et vous permet de facilement tracer retour à l'information originale
- Si vous posez à NotebookLM une question qui n'est pas couverte dans les matériaux que vous avez fournis, il vous dira honnêtement qu'il ne peut pas répondre. Il n'inventera pas d'informations ou ne spéculera pas. Cela assure que l'information que vous obtenez est toujours ancrée dans vos sources et fiable

8. Outils Google Cloud pour Agents

8.1 Outils de Stockage et Données

Cloud Storage : Service de stockage d'objets hautement évolutif et durable. Utiliser Cloud Storage pour stocker et récupérer des données dont votre agent a besoin.

Bases de Données (Cloud SQL, Cloud Spanner, Firestore) : Google Cloud offre une variété de solutions de base de données pour répondre à vos besoins. Votre agent peut utiliser ces bases de données pour stocker et récupérer des informations, gérer les données utilisateur, ou suivre son propre progrès.

8.2 Outils de Calcul

Cloud Run Functions : Créer des fonctions serverless qui agissent comme outils spécialisés pour votre agent. Cloud Run Functions peut être utilisé pour se connecter aux bases de données, appeler des APIs externes, effectuer des calculs complexes, ou gérer d'autres tâches spécifiques. Elles sont facilement déclenchées par votre agent et s'adaptent automatiquement.

Cloud Run : (conteneurs)

Vertex AI : (pour les capacités IA spécialisées)

9. Solutions Customer Engagement

9.1 Customer Engagement Suite

Google's Customer Engagement Suite a des outils conçus pour soutenir votre entreprise dans l'engagement efficace avec les clients. Les outils sont construits utilisant l'IA conversationnelle, dont une partie est aussi de l'IA générative.

Tout cela peut être construit au-dessus du Contact Center as a Service (CCaaS) de Google, une solution de centre de contact de niveau entreprise qui est native au cloud.

9.2 Approche Hybride

Alors que l'IA déterministe seule peut être rigide, et l'IA générative seule peut manquer de structure, leur combinaison crée une solution puissante. Une approche hybride vous permet de construire des agents conversationnels qui peuvent gérer efficacement une large gamme de besoins clients.

9.3 Agent Assist

Outil qui soutient les agents humains en direct avec une assistance en temps réel, des réponses générées, et un coaching en temps réel pour les aider à résoudre les problèmes clients plus rapidement et avec une plus grande précision.

9.4 Conversational Insights

Que vous utilisiez des agents en direct ou virtuels, ces connexions avec vos clients fournissent beaucoup de valeur et d'insights pour votre entreprise. Conversational Insights analyse les données conversationnelles à travers le parcours client pour fournir aux dirigeants et managers de centres de contact des insights basés sur les données pour augmenter l'efficacité, améliorer la performance des agents, et créer de meilleures expériences client.

9.5 Conversational Agents - Playbooks

Quand vous construisez un agent IA générative avec Conversational Agents, vous créez ce qu'on appelle un playbook pour comment vous voulez que votre agent agisse. Dans le playbook, vous définissez l'objectif de votre agent, comme fournir du support client, répondre aux questions utilisateur, ou même générer du contenu créatif.

System Instructions : Avec les instructions système, vous pouvez fournir contexte, persona, et contraintes à un agent avant qu'aucune input utilisateur soit fournie pour guider le comportement du modèle et s'assurer que ses réponses s'alignent avec votre résultat désiré.

10. Coûts et Modèles de Facturation

10.1 Modèles de Facturation des Modèles IA

Les modèles IA sont facturés :

- **Au token**
- **Au caractère**
- **Au temps d'utilisation**
- **À la requête**

10.2 Types de Tarification

Globalement, on paie :

- **À l'utilisation**
- **En souscription**
- **En licence**
- **En accès libre pour usage non commercial**

10.3 Facteurs Influençant le Prix

Le prix change en fonction de :

- **Du modèle**
- **Du contexte window**
- **De caractéristiques spéciales** (fine tuning)
- **Des modalités de déploiement**

11. Sécurité et IA Responsable

11.1 Sécurisation d'une IA

Étapes de sécurisation :

- Veiller à la qualité de la data ingérée
- Prendre en compte la confidentialité lors de leur préparation, notamment en les anonymisant et enlevant toute anomalie
- Veiller à la sécurisation des accès lors de l'entraînement
- Faire les mises à jour régulières
- Vérifier les résultats générés

11.2 Responsible AI

Principes fondamentaux :

Transparence : Être transparent avec les utilisateurs et clients sur l'utilisation de leurs données, les données doivent être anonymisées.

Aspect Éthique : Prendre en compte l'aspect éthique en limitant les biais générés par le corpus d'entraînement, afin de ne pas produire de discrimination et relayer les biais sociétaux.

Fairness (Équité) et Accountability (Responsabilité)

Human-in-the-Loop (HITL) : Principe fondamental avec exemples :

- Sélection de data
- Design du prompt
- Évaluation des résultats
- Feedback pour amélioration continue

Concept Clé : Différencier automation et augmentation (assistance de l'IA pour une prise de décision humaine)

11.3 Objectifs Finaux

Objectifs finaux d'une IA éthique :

- Ne pas causer de mal
- Permettre une utilisation responsable

12. Intégration de l'IA en Entreprise

12.1 Approches d'Intégration

Top Down : L'IA doit soutenir l'objectif clé de l'entreprise

Bottom Up : L'IA peut résoudre des problèmes rencontrés par tous les salariés avec l'idée d'expérimenter et de gagner en efficacité

12.2 Creative Matrix

Sorte de brainstorming pour envisager des manières d'intégrer l'IA en entreprise :

- On met deux axes : Y = domaines clés d'entreprise et X = outils IA
- À l'intersection des deux, on colle des post-it avec des idées des employés

13. Hiérarchie de l'IA Générative

Structure hiérarchique :

- **Applications**
- **Agents**
- **Plateformes**
- **Models**
- **Infrastructure**

Définition d'un Agent : Un agent réalise une tâche de manière autonome :

- Interagir en utilisant le langage humain
- Accomplir de l'automation
- Personnaliser des services

Types d'Agents Spécialisés :

- **Agent conversationnel**
- **Agent workflow**

Certaines applications sont dites multi-agents, car elles proposent plusieurs fonctionnalités IA.

14. NotebookLM et Agentspace

14.1 NotebookLM

Versions :

- **NotebookLM** (version standard)
- **NotebookLM Plus**
- **NotebookLM Enterprise**

14.2 Différence NotebookLM vs Agentspace

NotebookLM Enterprise : Outil IA spécialisé pour approfondir des documents et sources web spécifiques – poser des questions, résumer, et créer du nouveau contenu basé uniquement sur ces sources.

Agentspace : Assistant IA d'entreprise compréhensif. Utilise des agents IA et une recherche unifiée pour automatiser les tâches et trouver des informations à travers tous vos systèmes d'affaires connectés, pas seulement des documents spécifiques que vous uploadez.

Relation : Agentspace peut se connecter à NotebookLM Enterprise, mais ils servent des objectifs centraux différents.