Cours 11

Capacités et Limitations de l'IA

Anthropic Academy — Resume detaille

Yoann Boulch | 01/04/2026

Introduction

Construire un modèle mental précis du fonctionnement des systèmes d'IA est essentiel pour les utiliser efficacement. Ce cours explore les capacités réelles et les limitations des modèles de langage, en fournissant une base pour comprendre pourquoi l'IA produit parfois des réponses plausibles mais inexactes.

Qu'est-ce que l'IA?

L'IA englobe de nombreuses technologies différentes. Dans ce cours, nous nous concentrons principalement sur les modèles de langage volumineux (LLMs), une catégorie d'IA générative basée sur la prédiction de jetons.

■ IA générative vs autres types d'IA: L'IA générative crée nouveau contenu, tandis que l'IA traditionnelle classe ou prédit des valeurs discrètes.

Comment l'IA acquiert son caractère

Le processus d'entraînement façonne fondamentalement le comportement d'un modèle d'IA. Trois composantes clés influencent ce processus:

1. Apprentissage supervisé: Le modèle apprend à prédire les jetons suivants à partir de milliards de textes.
2. RLHF (Reinforcement Learning from Human Feedback): Les retours humains affinent les préférences du modèle.
3. IA constitutionnelle: Ensemble de principes qui guident le comportement du modèle.

Prédiction du prochain jeton

Au cœur des modèles de langage se trouve un mécanisme simple mais puissant: la prédiction du prochain jeton. À chaque étape, le modèle analyse le contexte fourni et prédit le jeton le plus probable à venir.

Comment cela fonctionne?

Chaque jeton est converti en représentation numérique. Le modèle traite cette séquence pour générer une distribution de probabilité sur tous les jetons possibles. Puis un jeton est sélectionné (déterministiquement ou de manière aléatoire).

Implications pratiques

■■ Les réponses peuvent être plausibles mais incorrectes: Le modèle prédit ce qui vient généralement après, pas ce qui est vrai.

Comportements émergents: Des propriétés imprévisibles émergent du simple processus de prédiction de jetons.

Exercices pratiques: Essayez de prédire le jeton suivant dans vos propres textes pour développer l'intuition.

Température et variabilité

La température contrôle le caractère aléatoire de la sélection. Une température basse (0.0-0.7) produit des réponses plus déterministes et cohérentes. Une température élevée (0.8+) produit plus de variabilité et de créativité.

Connaissance dans l'IA

Ce que l'IA 'connaît' est entièrement basé sur ses données d'entraînement. Cela crée des forces et des faiblesses distinctes.

Données d'entraînement et limite de connaissance

Claude a été entraîné sur des textes jusqu'à février 2025. Les événements, découvertes ou changements après cette date ne sont pas connus du modèle. Les données d'entraînement incluent du texte qui reflète les biais de ces sources.

Implications pour la fiabilité

✓ Quand faire confiance à la connaissance de l'IA: Pour les informations bien établies, largement couvertes dans les données d'entraînement.

✗ Quand vérifier: Pour les informations récentes, spécialisées, ou sensibles.

Hallucinations

Une hallucination se produit quand un modèle génère du contenu plausible mais factellement incorrect. Cela arrive parce que le modèle prédit ce qui vient généralement, pas ce qui est vrai.

Fenêtre contextuelle

Claude peut traiter jusqu'à 200,000 jetons dans une seule conversation. Au-delà de cette limite, le contexte antérieur est oublié. C'est dû à la façon dont les mécanismes d'attention fonctionnent.

■ Implication: Les très longues conversations peuvent se dégrader en qualité. Structurez les conversations en résumant les points clés.

Dirigibilité (Steerability)

Les petits changements dans les instructions (prompts) peuvent produire de grandes différences dans les résultats. Cela permet une contrôle fin, mais aussi une sensibilité imprévisible.

Techniques de direction

Role prompting: Donner au modèle un rôle spécifique (ex: 'Vous êtes un expert en finance').

Contrôle du format: Spécifier exactement comment la réponse doit être formatée.

Contraintes: Fournir des limites claires sur ce qui doit ou ne doit pas être inclus.

Quand les propriétés entrent en conflit

Parfois, la connaissance et la dirigibilité entrent en conflit. Par exemple, un prompt peut demander au modèle de générer du contenu qui contredit ses connaissances de base. Diagnostiquer quel problème cause un résultat inattendu est la clé pour le corriger.

Tableau de diagnostic

Symptôme	Propriété en cause	Solution
Réponse incorrecte mais plausible	Prédiction vs Précision	Vérifier les faits, ajouter des sources
Réponse en dehors du format demandé	Dirigibilité insuffisante	Clarifier le format avec des exemples
Perte de contexte après longue conversation	Fenêtre contextuelle	Résumer, structurer la conversation
Contenus inappropriés	RLHF insuffisant	Affiner les instructions de sécurité

Continuum Capacité-Limitation

Chaque tâche se situe quelque part sur un spectre allant de très facile à très difficile pour l'IA. Comprendre où se situe votre tâche vous aide à déterminer si l'IA convient.

Tâches où l'IA excelle

Résumer, reformuler, générer du contenu créatif, brainstorming, expliquer des concepts.

Tâches où l'IA struggle

Arithmétique précise, informations très récentes, inférence logique complexe, mathématiques avancées.

■ Évaluation rapide: Demandez-vous si la tâche nécessite de vérifier les faits contre une source. Si oui, l'IA doit être vérifiée.

En comprenant ces propriétés fondamentales, vous pouvez utiliser l'IA efficacement, reconnaître ses limitations et savoir quand appliquer un jugement humain critique.