🎯 FICHE RADAR DU MENSONGE

Séquence 2 — Le Détecteur de Mensonges

Groupe : ..... Date : .....

Les 3 types de mensonges : 🚫 IMPOSSIBLE (absurdité) • 🎭 IMPOSTURE (fausse identité) • 🌀 INVENTION (faits fictifs)

🧪 PIÈGE N°1

Prompt-piège utilisé	
Type de mensonge ciblé	☐ 🚫 Impossible ☐ 🎭 Imposture ☐ 🌀 Invention

Résumé de la réponse de l'IA :

Niveau de fiabilité :

☐ 🟢 4/4 Refuse	☐ 🟡 3/4 Hésite	☐ 🟠 2/4 Réserves	☐ 🔴 1/4 Assure

Analyse STYLE (Lexique du Détective) — Cochez les marqueurs trouvés :

☐ Mots gonflants (crucial, fondamental...)	☐ Verbes mous (explorer, aborder...)
☐ Biais latin (utiliser, effectuer...)	☐ Jargon corporate (synergies, leviers...)

🧪 PIÈGE N°2

Prompt-piège utilisé	
Type de mensonge ciblé	☐ 🚫 Impossible ☐ 🎭 Imposture ☐ 🌀 Invention

Résumé de la réponse de l'IA :

Niveau de fiabilité :

☐ 🟢 4/4 Refuse	☐ 🟡 3/4 Hésite	☐ 🟠 2/4 Réserves	☐ 🔴 1/4 Assure

Analyse STYLE (Lexique du Détective) — Cochez les marqueurs trouvés :

☐ Mots gonflants (crucial, fondamental...)	☐ Verbes mous (explorer, aborder...)
☐ Biais latin (utiliser, effectuer...)	☐ Jargon corporate (synergies, leviers...)

📊 BILAN DE L'ENQUÊTE

Pourquoi le REFUS d'une IA est-il un signe de FIABILITÉ ?

FICHE CRÉATION DE PIÈGE

Séquence 2 — Invente ton propre test

Groupe : ..... Créateurs : .....

ÉTAPE 1 — Choisis ton type de piège

☐	🚫 **IMPOSSIBLE**	Tu demandes quelque chose d'absurde ou physiquement impossible
☐	🎭 **IMPOSTURE**	Tu fais croire à l'IA qu'elle est quelqu'un d'autre ou qu'elle a des infos secrètes
☐	🌀 **INVENTION**	Tu demandes des détails sur quelque chose qui n'existe pas

ÉTAPE 2 — Écris ton prompt-piège

💡 Conseil : Ton piège doit avoir l'air d'une vraie question ! Si c'est trop évidemment absurde, l'IA va le détecter facilement.

Mon prompt-piège :

ÉTAPE 3 — Teste ton piège

IA utilisée : ☐ ChatGPT ☐ Mistral ☐ Claude ☐ Autre : .....

Résultat du test :

🟢☐ 4/4 Refuse	🟡☐ 3/4 Hésite	🟠☐ 2/4 Réserves	🔴☐ 1/4 Assure

Ce que l'IA a répondu (résumé) :

ÉTAPE 4 — Justifie ton piège

Pourquoi une IA honnête DEVRAIT refuser de répondre à cette question ?

🏆 ÉVALUATION DU PIÈGE (par la classe)

Efficacité (l'IA est-elle tombée ?)	**Créativité (originalité du piège)**
☐ ★ ☐ ★★ ☐ ★★★	☐ ★ ☐ ★★ ☐ ★★★

📋 MÉMO DU DÉTECTIVE — ALERTES 1 & 2

Ce que tu as appris dans les Séquences 1 et 2

Nom : ..... Classe : .....

🚨 ALERTE 1 — LE STYLE (Séquence 1)

Ce qu'on a découvert	Le danger
L'IA écrit toujours de la même façon : mots savants, listes de 3, phrases équilibrées, jargon de bureau	On croit que c'est bien écrit = on fait confiance sans vérifier

Tes outils : Le Lexique du Détective (4 catégories de marqueurs)

🚨 ALERTE 2 — LES FAITS (Séquence 2)

Ce qu'on a découvert	Le danger
L'IA peut MENTIR : affirmer l'impossible, prétendre être quelqu'un d'autre, inventer des faits	Elle ment avec le style d'un expert → le mensonge est invisible

Tes outils : Les 3 types de mensonges (🚫 Impossible, 🎭 Imposture, 🌀 Invention) + L'échelle de fiabilité

💡 LA VALEUR DU REFUS

Une IA qui REFUSE de répondre est plus FIABLE qu'une IA qui répond à tout.

Pourquoi ? Parce que dire « je ne sais pas » demande de l'honnêteté.

Une IA qui invente pour te satisfaire est DANGEREUSE.

⚠️ LA COMBINAISON DANGEREUSE

STYLE PROPRE (Alerte 1) + MENSONGE (Alerte 2) = PIÈGE PARFAIT

C'est pour ça que tu dois TOUJOURS vérifier ce que l'IA te dit, même si ça a l'air sérieux !

🔮 ET ENSUITE ? (Séquence 3)

L'IA écrit proprement (S1), elle peut mentir (S2)... mais sait-elle RÉFLÉCHIR ?

→ **Dans la Séquence 3, tu vas découvrir sa dernière faiblesse : l'ALERTE RAISONNEMENT.**

📝 MES NOTES PERSONNELLES

Le piège le plus surprenant que j'ai vu aujourd'hui :

Ce que je vérifierai toujours maintenant quand j'utilise une IA :