Synapsteria

Recherche appliquée / assistants documentaires

Écho

Évaluation de la capacité de réponse des différents modèles

Huit modèles, dix questions, un contexte documentaire commun

Du modèle de 4 milliards de paramètres exécuté sur GPU aux grands modèles accessibles par API : comparer la justesse, le délai de réponse, le coût et l'utilité pour l'entreprise.

Rapport consolidé / Projet ORION

Campagne des 15 et 16 octobre 2026

Résultats des essais existants. Aucune génération supplémentaire pour ce rapport.

1 Ce que l'entreprise peut retenir

Les modèles les plus performants de cette campagne exploitent mieux un même dossier. Le choix ne se résume cependant ni au nombre de paramètres, ni au prix, ni à la qualité apparente du français. Une réponse peut être agréable à lire et perdre une information décisive.

95 % pour Sol et Opus ; **1,4 seconde** pour Groq ; **0,058 $** pour les dix réponses de GPT mini. Ces trois résultats décrivent trois compromis différents.

Un gain de qualité qui se mesure

Sol et Opus satisfont 38 des 40 critères, contre 34 pour GPT mini, 32 pour Groq et 21 pour Qwen 4B. Mistral atteint 37 et Gemini 36. Sol et Opus produisent neuf réponses sur dix entièrement conformes au barème. La question de pilotage Q9 reste limitée par les preuves reçues.

Le modèle le plus cher n'est pas le plus rapide

Sol et Opus obtiennent le même score. Sol répond en médiane en 4,5 secondes contre 8,4 secondes pour Opus, avec une estimation de coût environ 3,25 fois plus faible. Opus rédige davantage : 1 899 mots contre 1 137 pour Sol. Le détail supplémentaire peut être utile, mais il n'apporte pas de point supplémentaire ici. Groq est plus rapide que tous les autres essais mesurés, avec une note de 80 %. Mistral est proche de la tête en pertinence, mais son effort de raisonnement élevé s'accompagne d'une médiane de 14,8 secondes. La nouveauté et le prix ne prédisent donc pas à eux seuls la vitesse.

La valeur se construit à deux niveaux

Écho doit d'abord choisir les bonnes preuves, puis le modèle doit les interpréter. La question Q9 montre une limite commune de sélection ; les différences entre modèles montrent ensuite une capacité variable à exploiter ce qui est présent. Augmenter le budget du modèle ne remplace pas l'amélioration du contexte.

Orientation proposée : une voie rapide pour les demandes simples ; un modèle plus précis pour les rapprochements de dates, de responsabilités et de montants ; une validation humaine pour les décisions engageantes. Le routage doit être évalué avant mise en production.

2 Le protocole de comparaison

Une évaluation de réponses documentaires, construite pour rendre les écarts observables.

Le dossier ORION comprend 18 documents d'un projet industriel fictif : exigences, devis, contrat, plannings, réception, risques et finances. L'état documentaire va jusqu'à juin 2025. Dix questions couvrent faits simples, chronologie, calculs, lecture de tableau, contradiction documentaire, plan d'action et abstention.

Des preuves communes, une grille fixe

La référence est l'essai GPT mini avec recherche hybride et embeddings activés. Les messages et passages ainsi préparés ont été repris pour les comparaisons : vérification des messages enregistrés pour les essais API et Qwen, saisie vérifiée au moment du test dans AI Studio pour Gemini. Les synthèses WikiRAG accompagnent certains passages ; elles restent des sources secondaires. Chaque question comporte quatre critères binaires. Un critère satisfait vaut un point ; un critère composé incomplet vaut zéro. Le total est converti en pourcentage :

$$\text{Pertinence} = \frac{\sum_{q=1}^{10} \sum_{c=1}^{4} \text{point}_{q,c}}{40} \times 100.$$

Ce pourcentage représente la satisfaction de la grille. Il ne signifie ni probabilité universelle de vérité, ni pourcentage de phrases correctes. Le corrigé et les réponses attendues ne sont pas fournis au modèle testé. La correction existante est réutilisée sans changer le barème.

Ce qui est comparable, ce qui varie

Les questions et le contexte documentaire sont communs. Le fournisseur, le matériel, la quantification, l'effort de raisonnement et le plafond de sortie varient. Les petits essais avaient généralement 1 000 tokens de sortie ; Gemini 4 096 ; Sol, Opus et Mistral 6 144. Une seule réponse par question et par configuration est évaluée : il n'y a ni répétition statistique, ni intervalle de confiance.

Indicateur	Définition retenue
Pertinence	Critères satisfaits sur 40, convertis en pourcentage.
Latence	Médiane du temps jusqu'à la réponse complète, réseau compris. GPT mini comprend aussi la recherche. Pour Gemini, repère estimatif de 3,4 s fondé sur une seule durée conservée.
Coût / 10	Valorisation des générations payantes, ou zéro pour les essais sous quota gratuit / exécution Colab gratuite. Les coûts du serveur et de l'exploitation ne sont pas inclus.
Français	Appréciation qualitative : clarté, construction des phrases, densité, répétitions. Aucun score vocal.

La portée des résultats

Ce dispositif isole surtout l'exploitation d'un contexte déjà sélectionné. Il ne classe pas les moteurs de recherche, les embeddings ou les solutions RAG complètes. Les scores ne préjugent pas du résultat sur les documents d'une autre entreprise. Un écart d'un point sur 40 mérite une nouvelle mesure avant décision d'achat.

3 Les résultats consolidés

Ordre croissant de pertinence, identique au comparatif en ligne.

Modèle	Pertinence	/40	Latence	Coût / 10
Qwen3 4B	52,5 %	21	11,1 s	0 $
Llama 3.3 70B	62,5 %	25	9,8 s	0 $
gpt-oss-120B	80 %	32	1,4 s	0 $
GPT-5.4 mini	85 %	34	2,3 s	0,058 $
Gemini 3.8 Flash	90 %	36	≈ 3,4 s	0 $
Mistral Large 4	92,5 %	37	14,8 s	0,163 $
GPT-6.1 Sol	95 %	38	4,5 s	0,186 $
Claude Opus 5.5	95 %	38	8,4 s	0,603 $

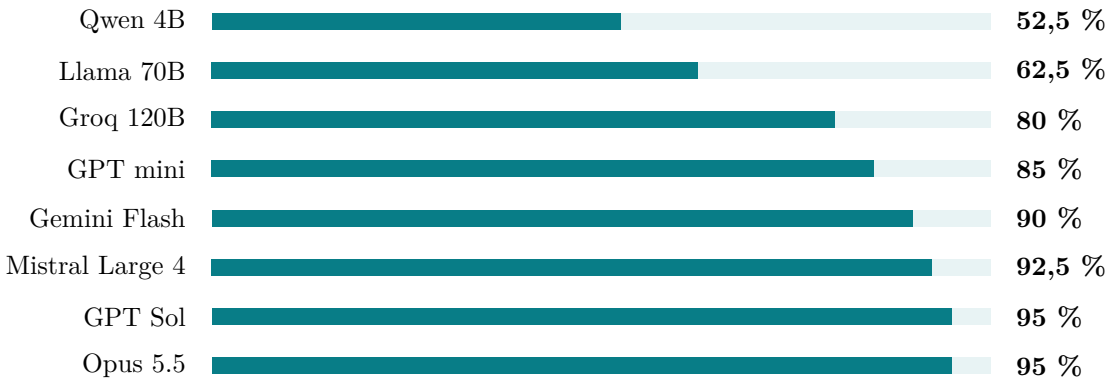

Figure 1. Satisfaction des critères sur les dix questions. Les écarts de rang sont ceux de ce lot, pas un classement universel.

Quatre groupes se dessinent

Sol et Opus (95 %) dominent le barème avec des réponses complètes sur neuf questions. **Mistral et Gemini (92,5 et 90 %)** sont proches ; leur position demande de tenir compte des réglages et des omissions spécifiques. **GPT mini et Groq (85 et 80 %)** offrent un compromis intéressant pour une interface réactive. **Llama et Qwen (62,5 et 52,5 %)** rencontrent davantage de difficultés sur les rapprochements documentaires et les citations. Le faible score de Qwen ne vient pas d'un arrêt prématuré de cette campagne : ses dix réponses sont complètes. Le GPU T4 a rendu l'essai réalisable, sans corriger ses confusions de versions documentaires.

4 Rapidité et coût : lire le compromis

La vitesse dépend aussi de l'hébergement et des réglages de génération.

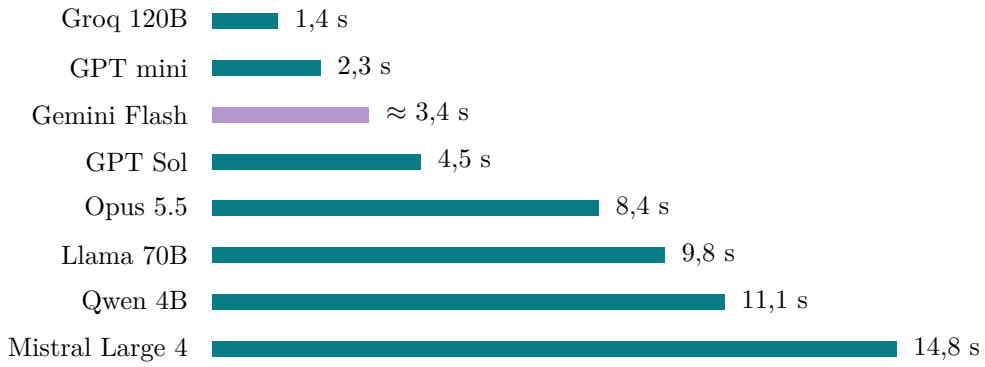

Figure 2. Latence de réponse complète. En violet : Gemini, estimation fondée sur une observation ; les autres valeurs sont des médianes sur dix questions.

Le coût utile est celui d'un service complet

Un modèle librement hébergeable ne supprime pas le coût des GPU, de la mémoire, de l'énergie, de la disponibilité et de la maintenance. Les zéros affichés correspondent aux conditions gratuites de ces essais, pas à une promesse de production gratuite. De même, les estimations API n'intègrent pas ingestion, embeddings, recherche, synthèse vocale ni support.

Modèle API valorisé	10 réponses	1 000 équivalentes	Pertinence
GPT-5.4 mini	0,058 $	5,76 $	85 %
Mistral Large 4	0,163 $	16,28 $	92,5 %
GPT-6.1 Sol	0,186 $	18,55 $	95 %
Claude Opus 5.5	0,603 $	60,33 $	95 %

L'extrapolation à 1 000 réponses multiplie le lot par 100 à longueur et tarifs constants. Elle sert à cadrer un ordre de grandeur ; ce n'est pas une prévision de facture. Le détail de la valorisation figure en annexe.

À score égal, l'écart de coût compte : Opus coûte environ 3,25 fois l'estimation de Sol sur ces réponses. Mistral est un peu moins cher que Sol ici, mais plus lent. GPT mini conserve 85 % pour environ 0,058 $ le lot.

Ce qu'on peut réellement conclure sur les nouveaux modèles

Les modèles de tête améliorent ici la complétude, généralement à un coût supérieur à GPT mini. Certains sont plus rapides que Qwen et Llama dans leurs configurations testées. En revanche, ils ne sont pas tous plus rapides que Groq ou GPT mini. Le meilleur compromis doit être choisi par type de tâche, et non par génération commerciale du modèle.

5 Où se situent les écarts ?

Une matrice par question évite de masquer les faiblesses derrière une moyenne.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Qwen 4B	3	4	3	0	1	2	1	3	0	4
Llama 70B	4	4	4	1	3	2	1	3	0	3
Groq 120B	4	4	4	4	3	4	3	3	0	3
GPT mini	4	4	4	4	4	4	3	3	0	4
Gemini Flash	4	4	4	4	4	4	4	3	1	4
Mistral Large 4	4	4	4	4	3	4	4	4	2	4
GPT Sol	4	4	4	4	4	4	4	4	2	4
Opus 5.5	4	4	4	4	4	4	4	4	2	4

Figure 3. Points obtenus par question, de 0 à 4. Q9 concentre les pertes, mais ne les explique pas toutes.

Q9 : une limite de la chaîne Écho autant que du modèle

La question demande quatre actions couvrant problème, responsable et preuve. Les passages transmis omettent notamment les éléments D45/D91 associant les fournisseurs à la reprise MES, et D82 qui nomme Thomas Rivière pour la présentation de l'arbitrage actuel. Le corpus contient donc des preuves que le modèle ne reçoit pas dans cette question. Sol et Opus obtiennent 2/4 ; Mistral également. Leurs réponses sont largement utiles, mais deux familles ne sont pas entièrement satisfaites. Opus reconnaît explicitement qu'un responsable n'est pas nommé dans ses extraits. Cette abstention vaut mieux qu'une invention, même si le critère de complétude reste à zéro.

La grille est volontairement exigeante

Une famille d'action peut contenir plusieurs informations correctes tout en perdant son point faute d'un responsable. Un zéro à Q9 ne signifie donc pas que toute la réponse est fausse. Inversement, 4/4 ne signifie pas qu'aucune phrase secondaire n'est discutable : des formulations fragiles de Mistral en Q3 et Q7 restent signalées dans les corrections.

Les autres écarts restent bien liés aux réponses

Qwen confond les résultats de la réception en usine avec ceux de la réception sur site pour Q4. Llama omet des mesures et des dates pourtant demandées. Mistral ne rattache pas explicitement l'incident Q5 au contrôle pneumatique. L'amélioration de la recherche ne corrige donc pas toutes les erreurs de lecture ou de synthèse.

6 Une décision d'entreprise, pas un podium

Choisir un service fiable, maîtrisable et utile aux équipes.

Priorité	Lecture des essais	Décision à instruire
Réactivité	Groq répond le plus vite ; GPT mini garde une latence faible et un score supérieur.	Mesurer le premier texte visible et le temps de réponse complète sur le trafic réel.
Complétude	Sol et Opus obtiennent 95 %, avec des réponses de longueur différente.	Tester un niveau renforcé pour les questions de décision, dates, montants et responsabilités.
Budget	Le prix d'une génération ne couvre pas le coût total du service.	Fixer quotas, plafonds, suivi par session et politique de reprise après erreur.
Confidentialité	Qwen a été exécuté sur Colab ; cela ne constitue pas un test d'hébergement interne.	Choisir le lieu de traitement, les droits d'accès, la conservation et les modalités fournisseur.
Traçabilité	Les dates et les citations sont décisives pour éviter une réponse obsolète.	Conserver le lien entre affirmation, extrait, document, version et date.
Exploitation	Ni concurrence, ni disponibilité, ni montée en charge ne sont mesurées ici.	Éprouver le service sous charge et prévoir un repli contrôlé.

Le français fait partie de l'expérience

Sol est concis ; Opus développe davantage ; Mistral structure des phrases complètes ; Gemini reste mesuré. Groq est plus dense et segmenté. Qwen et Llama sont lisibles, mais répétitions ou omissions peuvent rendre la réponse moins utile. Une belle formulation ne doit jamais remplacer la vérification du fait. La voix exige une évaluation distincte : retirer les citations à l'oral, conserver les liens à l'écran, maintenir des phrases naturelles et vérifier la compréhension. Cette campagne n'évalue pas la synthèse vocale.

Trois prochaines étapes prioritaires

1. **Corriger la sélection des preuves.** Sur les questions de pilotage, vérifier la couverture des quatre dimensions demandées avant génération : problème, responsable, décision et preuve.
2. **Mesurer la chaîne complète.** Évaluer séparément la récupération, la réponse et l'expérience, avec des reformulations et des relances de conversation.
3. **Tester un routage sobre.** Réserver le raisonnement coûteux aux demandes qui en bénéficient, sur un jeu plus large et répété, avec validation humaine des décisions sensibles.

Conclusion : le dossier montre une base solide et un gain de qualité réel avec les modèles de tête. La performance d'entreprise viendra de l'association d'une bonne sélection documentaire, d'un modèle adapté à la tâche et d'un contrôle explicite des coûts, des sources et des décisions.

7 Les questions 1 et 2

Questions originales, éléments attendus et résultats par modèle.

Q1 / Fait explicite et compréhension du projet

Question. Que fabrique Mécaval à Chalon, quel contrôle ORION doit-il remplacer et quelles sont la cadence initiale et la cadence nominale visée ?

Réponse attendue. Mécaval fabrique des corps de vannes. À Chalon, ORION remplace le contrôle final manuel des corps de vannes P40 par une cellule de contrôle d'étanchéité et de traçabilité. La cadence initiale moyenne est de 42 pièces par heure ; la cible nominale est d'au moins 60 pièces par heure, avec un cycle maximal de 60 secondes.

Résultats /4. Qwen 4B 3 ; Llama 70B 4 ; Groq 120B 4 ; GPT mini 4 ; Gemini Flash 4 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. Les modèles de tête identifient le projet et distinguent objectif et résultat. Qwen restitue les faits, mais ne fournit pas la citation attendue. La différence porte ici autant sur la vérifiabilité que sur le contenu.

Sources du corrigé : ORION-CDC-001 p. 1 ; ORION-CDC-001 p. 2.

Q2 / Fraîcheur, version et statut financier

Question. Combien coûte la formation au dernier état du dossier ? Précise ce que représente ce montant et ce qui a changé depuis le budget initial.

Réponse attendue. Au 20 juin 2025, la prévision finale du poste Formation est de 19 000 EUR HT : 7 000 comptabilisés, 3 100 engagés non facturés et 8 900 estimés non engagés. Le budget initial du 21 mars était de 14 000 EUR HT ; la prévision augmente donc de 5 000 EUR HT. Les 19 000 EUR sont une prévision à terminaison, pas un coût final constaté ni la preuve d'un relèvement formel du budget autorisé.

Résultats /4. Qwen 4B 4 ; Llama 70B 4 ; Groq 120B 4 ; GPT mini 4 ; Gemini Flash 4 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. La mise à jour de 14 000 à 19 000 euros est correctement exploitée dans cette campagne. Ce cas montre que le contexte actualisé peut résoudre le défaut observé dans la première démonstration, sans suffire à garantir tous les rapprochements futurs.

Sources du corrigé : ORION-FIN-002 p. 1 ; ORION-FIN-001 p. 1.

8 Les questions 3 et 4

Questions originales, éléments attendus et résultats par modèle.

Q3 / Calcul de délai et choix de la bonne référence

Question. Quel est le retard de livraison par rapport au contrat initial ? Montre le calcul et explique pourquoi le planning révisé peut donner un autre écart.

Réponse attendue. Le contrat prévoyait la livraison le 12 mai 2025 et la livraison a eu lieu le 9 juin 2025 : 9 juin moins 12 mai donne 28 jours calendaires. Le planning B du 23 mai prévoyait le 26 mai ; par rapport à ce jalon, l'écart est de 14 jours. Ce planning de coordination ne modifie pas à lui seul le délai contractuel, et aucun avenant de délai accepté n'est joint au dossier.

Résultats /4. Qwen 4B 3 ; Llama 70B 4 ; Groq 120B 4 ; GPT mini 4 ; Gemini Flash 4 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. La distinction entre date contractuelle et planning de coordination est déterminante. Mistral conserve les bons chiffres mais ajoute une causalité trop affirmative. Le score doit être lu avec cette réserve de formulation.

Sources du corrigé : ORION-CTR-001 p. 2 ; ORION-RIS-002 p. 1.

Q4 / Lecture comparative d'un tableau technique

Question. À la réception sur site, quelles différences observe-t-on entre P40 et P60 pour le cycle, la lecture des codes et la traçabilité ? Peut-on conclure que toute la cellule est acceptée ?

Réponse attendue. À la SAT du 17 juin, P40 réalise un cycle de 58 secondes, lit 200 codes sur 200 (100 %) et ne présente aucune pièce sans acquittement MES lié. P60 réalise 72 secondes pour une limite de 60, lit 194 codes sur 200 (97 %, sous le seuil de 99,5 %) et présente six pièces sans acquittement MES lié. Les deux formats rejettent bien les 12 témoins défectueux sur 12. La réception globale bi-format est refusée ; la conformité des résultats P40 ne vaut pas acceptation de toute la cellule.

Résultats /4. Qwen 4B 0 ; Llama 70B 1 ; Groq 120B 4 ; GPT mini 4 ; Gemini Flash 4 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. Cette question discrimine la lecture des versions. Qwen reprend 75 secondes et 99 % issus de la réception en usine, au lieu de 72 secondes et 97 % à la réception sur site. Une source présente ne garantit pas le choix de la bonne date.

Sources du corrigé : ORION-QUAL-002 p. 1.

9 Les questions 5 et 6

Questions originales, éléments attendus et résultats par modèle.

Q5 / Cause documentée et chronologie fournisseur

Question. Quel incident fournisseur explique la replanification de la livraison au 9 juin, et quelle séquence de dates avait été annoncée ?

Réponse attendue. Luc Garnier a signalé le 14 mai une rupture de sonde de pression chez Nordis. Il annonçait un approvisionnement au 30 mai, des tests du 2 au 4 juin puis une livraison au 9 juin. L'incident concerne le contrôle pneumatique Nordis. Il ne faut pas le transformer en panne de caméra VisiTrace ni attribuer, sans preuve supplémentaire, la totalité des écarts du projet à une cause unique.

Résultats /4. Qwen 4B 1 ; Llama 70B 3 ; Groq 120B 3 ; GPT mini 4 ; Gemini Flash 4 ; Mistral Large 4 3 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. Le lien fournisseur / composant / calendrier doit rester complet. Mistral obtient 3/4 car il omet le contrôle pneumatique ; plusieurs réponses plus faibles perdent d'autres précisions de la séquence annoncée.

Sources du corrigé : ORION-RIS-002 p. 1.

Q6 / Contradiction apparente et décision opérationnelle

Question. Le dernier rapport annonce un statut VERT, 92 % d'avancement et un démarrage cible au 23 juin. Cela permet-il de lancer la production en série ? Qui doit autoriser le démarrage et sous quelles conditions ?

Réponse attendue. Non. Les 92 % correspondent à 46 tâches closes sur 50 et le 23 juin est une cible de coordination. Le rapport de SAT refuse la réception globale et ne valide pas la date de production. Le démarrage exige l'autorisation conjointe de Camille Perrin et d'Anaïs Roux, après levée des écarts de traçabilité, vérification du cycle P60 et achèvement de la formation requise, notamment l'équipe de nuit. Le statut de reporting ne constitue pas cette autorisation.

Résultats /4. Qwen 4B 2 ; Llama 70B 2 ; Groq 120B 4 ; GPT mini 4 ; Gemini Flash 4 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. Le statut de reporting ne vaut pas décision qualité. Une réponse utile doit maintenir le refus de démarrage et identifier les autorités et les preuves requises, même lorsque le tableau général est rassurant.

Sources du corrigé : ORION-AVT-001 p. 1 ; ORION-QUAL-002 p. 2.

10 Les questions 7 et 8

Questions originales, éléments attendus et résultats par modèle.

Q7 / Calcul financier et fraîcheur réelle des données

Question. Quel dépassement budgétaire prévoit-on au dernier arrêté financier, en euros et en pourcentage ? Pourquoi ne faut-il pas simplement reprendre les +4 % du rapport général ?

Réponse attendue. Au 20 juin, la prévision à terminaison est de 472 500 EUR HT face à une autorisation de 420 000 EUR HT. Le dépassement prévu est donc de 52 500 EUR HT, soit 52 500 / 420 000 Œ 100 = 12,5 %. Le rapport général publié le 23 juin reprend un export financier arrêté au 17 juin et affiche +4 % ; sa publication plus récente ne rend pas ses données financières plus fraîches. Le plafond autorisé n'a pas été relevé et les 472 500 EUR ne sont pas les seules dépenses déjà comptabilisées.

Résultats /4. Qwen 4B 1 ; Llama 70B 1 ; Groq 120B 3 ; GPT mini 3 ; Gemini Flash 4 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. Un montant juste peut accompagner une explication fausse. Groq propose une cause non démontrée des +4 % ; Mistral inverse verbalement la soustraction malgré le résultat juste. Les dates d'arrêté comptent davantage que la seule date de publication.

Sources du corrigé : ORION-FIN-002 p. 1 ; ORION-AVT-001 p. 1.

Q8 / Statut d'engagement et distinction décision technique / achat

Question. L'option CHG-07 de 24 000 EUR HT peut-elle être considérée comme un achat formellement engagé parce que sa fabrication a commencé ? Explique son statut au dernier état documenté et indique qui doit régulariser l'achat.

Réponse attendue. Non. L'accord de fabrication a été transmis le 9 mai et les travaux ont débuté le 13 mai, mais aucun avenant bilatéral n'est signé au 19 juin. Le contrat exige un avenant approuvé avant les travaux ; un accord technique ne le remplace pas. Le suivi financier inclut donc les 24 000 EUR HT dans le reste estimé non engagé, et non dans les engagements d'achat signés. Julien Armand doit régulariser cet achat.

Résultats /4. Qwen 4B 3 ; Llama 70B 3 ; Groq 120B 3 ; GPT mini 3 ; Gemini Flash 3 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. Le commencement des travaux ne prouve pas l'engagement formel. Gemini donne le bon statut mais omet les dates du 9 et du 13 mai ; la famille chronologique perd donc un point.

Sources du corrigé : ORION-RIS-002 p. 2 ; ORION-FIN-002 p. 2 ; ORION-CTR-001 p. 2.

11 Les questions 9 et 10

Questions originales, éléments attendus et résultats par modèle.

Q9 / Synthèse multi-documents orientée actions

Question. Prépare un point de pilotage en quatre actions prioritaires : pour chacune, indique le problème, le responsable et la preuve attendue. Couvre la traçabilité, la cadence P60, la préparation des équipes et la situation financière/achats.

Réponse attendue. 1. Traçabilité P60 : Hugo Martin et les fournisseurs doivent sécuriser la reprise en Mode Secours et prouver le lien entre identifiant, résultat et acquittement MES, avec blocage de sortie tant que la trace est incomplète. 2. Cadence P60 : Nadia Benali doit faire recontrôler le cycle complet après les réglages et fournir un chronométrage conforme à la limite de 60 secondes ; les 72 secondes constatées restent une réserve. 3. Équipe de nuit : Anaïs Roux avec Malik Ferrand doivent former les 12 opérateurs de nuit encore non formés et remettre les présences ainsi qu'une preuve de prise en main ; les 12 opérateurs de jour sont déjà formés. 4. Finances et achats : Julien Armand doit obtenir la formalisation de CHG-07 ; Thomas Rivière doit présenter l'arbitrage sur la prévision totale et le dépassement de 52 500 EUR HT. Les preuves attendues sont l'avenant régularisé et une décision budgétaire formalisée, pas un simple accord oral. Les essais du 26 juin sont prévus, non démontrés comme réussis.

Résultats /4. Qwen 4B 0 ; Llama 70B 0 ; Groq 120B 0 ; GPT mini 0 ; Gemini Flash 1 ; Mistral Large 4 2 ; GPT Sol 2 ; Opus 5.5 2.

Lecture. Le résultat combine deux difficultés : des preuves décisives absentes de la sélection et des éléments présents mais omis par certaines réponses. La synthèse opérationnelle exige davantage qu'un résumé plausible. Le diagnostic détaillé figure dans la section sur Q9.

Sources du corrigé : ORION-RIS-002 p. 1 ; ORION-RIS-002 p. 2 ; ORION-QUAL-002 p. 2 ; ORION-FIN-002 p. 2 ; ORION-FIN-002 p. 1.

Q10 / Abstention et donnée absente

Question. Quelle était la marge nette de Mécaval en 2024, en pourcentage ? Réponds uniquement à partir du dossier ORION. Si ce n'est pas déterminable, indique quelles données manquent.

Réponse attendue. Le dossier ORION ne permet pas de déterminer la marge nette de Mécaval en 2024. Il décrit les coûts, engagements et prévisions du projet en 2025, sans fournir le chiffre d'affaires et le résultat net de l'entreprise pour 2024. Il faudrait les comptes correspondants pour calculer ce pourcentage.

Résultats /4. Qwen 4B 4 ; Llama 70B 3 ; Groq 120B 3 ; GPT mini 4 ; Gemini Flash 4 ; Mistral Large 4 4 ; GPT Sol 4 ; Opus 5.5 4.

Lecture. Refuser de calculer sans données est un résultat positif. Llama s'abstient mais laisse un faux repère de citation et ne précise pas les deux données attendues. L'abstention doit être claire, utile et sans source inventée.

12 Annexe : réglages, tokens et provenance

Modèle	Entrée	Sortie	Plafond sortie
Qwen 4B	79 089	1 916	1 000
Llama 70B	75 447	2 213	1 000
Groq 120B	66 655	3 529	1 000
GPT mini	66 015	1 801	1 000
Gemini Flash	non relevé	non relevé	4 096
Mistral Large 4	63 658	18 238	6 144
GPT Sol	66 015	2 048	6 144
Opus 5.5	108 690	8 427	6 144

Les tokens sont ceux déclarés par chaque fournisseur. Les unités de découpage et le traitement des tokens de réflexion diffèrent ; ces nombres ne sont pas une mesure commune de longueur du français. Gemini a été testé dans AI Studio sans relevé API de tokens.

Configurations

Qwen3 4B : Q4_K_M, GPU T4 Colab, sans réflexion. Llama 3.3 70B : Cloudflare FP8 Fast. gpt-oss-120B : Groq. GPT mini : parcours Écho initial avec embeddings. Gemini 3.8 Flash : AI Studio, effort faible. Sol : effort moyen. Opus : raisonnement adaptatif, effort moyen. Mistral Large 4 : préversion Large 4-0, effort élevé.

Convention de coût et de latence

Les valeurs payantes sont reprises des compteurs et tarifs enregistrés pendant les essais. Sol est valorisé prudemment au tarif d'écriture du cache pour toutes les entrées ; Opus inclut la réflexion ; GPT mini porte sur la génération seule. Mistral est recalculé à 1,36 $/million de tokens d'entrée et 4,18 $/million de sortie, sans la marge budgétaire de 30 % utilisée pour sécuriser le test. Il s'agit d'estimations, pas de factures. Groq, Cloudflare et AI Studio utilisent ici leurs quotas gratuits ; Qwen utilise Colab gratuit. Les 0 $ ne représentent pas un tarif permanent. Gemini est affiché à environ 3,4 s sur la base d'une seule durée conservée ; sa médiane n'est pas mesurée. GPT mini comprend le temps de recherche, contrairement aux générations rejouées.

Sources primaires de cette consolidation

Questions et corrigé : `questions.json`.

Barème : `grading-policy.json`. Réponses et corrections : huit campagnes des 15 et 16 octobre 2026. Scores et graphiques sont recalculés depuis ces fichiers, sans nouvel appel aux modèles. Les messages sont comparés à la référence ; Gemini repose sur son journal de saisie vérifiée. Les données et empreintes des corrections sont conservées avec la source de fabrication.

Version publique du comparatif : synapsteria.com/lab-echo/.