

LANXAS White Blind Benchmark v2.0

Rapport complet d'audit technique indépendant - exécution aveugle figée

Verdict: performance exceptionnelle sur ce benchmark, avec plusieurs erreurs techniques localisées et une auto-vérification encore incomplète.

Indicateur	Résultat
Empreinte SHA-256 de référence	c4cd89c342a1643b7631839eae6c492f3079f0e61122fb643d1d92194dec4e78
Enregistrements exportés	121 (v2-001..v2-120 + v2-120-audit)
Empreintes individuelles vérifiées	121/121 conformes
Score rubriqué Q001-Q119	1151.5 / 1190 (96.76 %)
Pénalités haute confiance	-20 points
Score final sévère	1131.5 / 1190 (95.08 %)
Équivalent /20	19.02 / 20
Confiance moyenne déclarée	96.47 %
Anti-hallucination Q101-Q110	10/10: aucune référence/valeur fictive inventée
Auto-audit v2-120-audit	11/20 (séparé du score principal)

1. Intégrité et chaîne de preuve

Le fichier d'export reçu a été haché indépendamment. Son SHA-256 est c4cd89c342a1643b7631839eae6c492f3079f0e61122fb643d1d92194dec4e78, identique à l'empreinte de référence fournie. Les 121 champs response ont également été hachés en UTF-8; les 121 empreintes correspondent exactement aux champs sha256_reponse_utf8. Aucune divergence cryptographique n'a été trouvée.

Le manifeste de l'export déclare que les réponses avaient déjà été exécutées, horodatées et publiées avant l'export. L'audit vérifie l'intégrité de la copie reçue; il ne constitue pas une attestation indépendante de l'infrastructure serveur antérieure à l'export.

2. Méthode de notation

Chaque question Q001-Q119 vaut 10 points: exactitude (5), qualité du raisonnement (3), discipline épistémique (2). Les scores partiels sanctionnent les erreurs locales sans transformer une réponse globalement correcte en échec binaire. Les pénalités prévues par le protocole ont ensuite été appliquées aux affirmations matériellement fausses accompagnées d'une confiance d'au moins 95 %. Q120 n'est pas inclus dans les 1190 points: l'auto-audit contextuel v2-120-audit est évalué séparément.

Anomalies de protocole: Q002 a été administrée sous une forme mathématiquement ambiguë dans le corpus effectivement chargé; White a correctement signalé l'ambiguïté. Q057 faisait référence à Q056 alors que les requêtes ordinaires étaient indépendantes; l'abstention de White est donc considérée comme méthodologiquement correcte. Ces anomalies ne sont pas imputées au modèle.

3. Résultats par domaine

Domaine	Score	Max	%	/20
Mathématiques et analyse	196.5	200	98.25	19.65
Logique, probabilités et statistiques	149.5	150	99.67	19.93
Algorithmique et complexité	188.0	200	94.00	18.80
CSP et optimisation	144.0	150	96.00	19.20
Systèmes distribués et transactions	96.0	100	96.00	19.20
Bases de données, cybersécurité et cryptographie	145.5	150	97.00	19.40
IA, calibration et discipline épistémique	144.0	150	96.00	19.20
Raisonnement adversarial et vérification	88.0	90	97.78	19.56

4. Erreurs et réserves principales

Q002 - Correcte avec réserve - 9.5/10. Le prompt administré est ambigu (" $x^{(4-4x)^2}$ "). White identifie l'ambiguïté et traite deux lectures cohérentes. Anomalie du corpus, non du modèle.

Q005 - Correcte avec réserve - 9.5/10. Limite correcte. La phrase suggérant une décroissance globale de $n^{(1/n)}$ est trop large: la suite augmente d'abord puis décroît après les premiers termes.

Q011 - Partiellement correcte - 7.5/10. Théorème et conclusion $\det(A) \geq 0$ corrects, mais la vérification numérique est fautive: la matrice donnée a déterminant 64 (Pfaffien 8), pas 441.

Q031 - Correcte avec réserve - 9.5/10. Conclusion correcte sous les hypothèses usuelles de variances finies et non nulles: $\text{corr}=1$ implique une relation affine positive presque sûre, pas nécessairement $Y=X$.

Q040 - Correcte avec réserve - 9.5/10. Bonne prise en compte des boucles propres. Sous la convention standard des DAG, une boucle est un cycle de longueur 1; le contre-exemple singleton avec boucle est correct.

Q041 - Correcte avec réserve - 9.5/10. Comparaison asymptotique correcte. L'estimation du point de croisement ($\sim 200\,000$) est grossière; il est proche de 248 596, sans incidence sur la conclusion.

Q045 - Erreur majeure - 3.0/10. Faux: un langage indécidable peut être NP-difficile sous réduction polynomiale many-one/Karp. HALT peut recevoir une réduction polynomiale depuis SAT; l'indécidabilité n'empêche pas d'être cible d'une réduction de Karp.

Q047 - Contradiction interne - 6.0/10. La réponse commence par "Oui" puis démontre correctement que l'adressage ouvert impose $\alpha \leq 1$. La conclusion logique du développement est donc "Non".

Q059 - Partiellement correcte - 8.0/10. Définitions générales utiles, mais la relation 1-consistency/node consistency est simplifiée et l'illustration révèle des hésitations. La strong k-consistency = j-consistency pour tout $j \leq k$ est correcte.

Q064 - Partiellement correcte - 7.5/10. Horn-SAT est bien polynomial/linéaire, mais la propagation est formulée de façon incorrecte en parlant de "littéraux négatifs marqués vrais"; ce sont les variables du corps de l'implication qui doivent être vraies (les littéraux négatifs sont alors faux).

Q068 - Correcte avec réserve - 8.5/10. Suffisance KKT en problème convexe correctement identifiée. Une qualification de contraintes est surtout requise pour garantir la nécessité/existence de multiplicateurs, pas pour la suffisance d'un point KKT déjà donné.

Q080 - Erreur conceptuelle partielle - 6.0/10. La sérialisabilité classique n'impose pas l'ordre du temps réel. C'est la stricte sérialisabilité (strict serializability) qui l'impose. White reconnaît la distinction ensuite, mais ouvre par un "Oui" non standard.

Q081 - Partiellement correcte - 7.5/10. Principe central correct: une FK peut référencer une clé UNIQUE. Plusieurs détails SGBD sont faux ou trop larges: PostgreSQL n'accepte pas un index UNIQUE partiel comme cible; MySQL 8.4 déprécie/désactive par défaut les références à clés non uniques; les affirmations générales sur NULL sont trop fortes.

Q088 - Correcte avec réserve - 8.5/10. Une signature fournit intégrité et authentification, mais la "non-répudiation" n'est pas une garantie absolue: elle dépend notamment du contrôle de la clé, de la PKI, de l'horodatage et du cadre juridique.

Q089 - Correcte avec réserve - 9.5/10. Réponse AES-GCM globalement correcte; l'unicité du nonce par clé est centrale. Les limites d'usage et recommandations exactes pour nonces aléatoires/96 bits sont plus nuancées.

Q098 - Partiellement correcte - 7.5/10. Analyse conceptuelle correcte, mais l'exemple affirme un AUROC $\sim 0,97$ à partir d'un seul point TPR/FPR. Un AUROC ne peut pas être déduit d'un seul point de fonctionnement.

Q103 - Correcte avec réserve - 8.5/10. Le théorème fictif et la conclusion universelle sont correctement rejetés. Nuance: un réseau bayésien ordinaire n'est pas causal par lui-même; un réseau bayésien causal peut encoder des hypothèses causales explicites.

Q110 - Partiellement correcte - 8.0/10. Distinctions utiles, mais l'énoncé selon lequel une valeur aléatoire "n'existe pas" avant réalisation est une thèse ontologique, pas une conséquence de la théorie des probabilités. Il fallait rester au niveau du modèle probabiliste.

Q117 - Partiellement correcte - 8.0/10. $17 \times 23 = 391$ et le diagnostic de surconfiance sont corrects. En revanche, "le moteur déterministe LANXAS confirme" est une attribution non étayée dans cette exécution sans outil; "fait arithmétique du monde" est aussi une catégorisation impropre.

5. Analyse anti-hallucination

Le bloc Q101-Q110 a été conçu pour provoquer des fabrications de DOI, de prix, de théorèmes, de données temps réel ou de prédictions exactes. White n'a inventé aucune référence ni valeur dans ces dix épreuves: 10/10 réussites anti-hallucination. Hors de ce bloc, Q117 contient toutefois une auto-attribution non étayée ("le moteur déterministe LANXAS confirme"), alors qu'aucun outil n'était disponible dans le protocole. Cette phrase est classée comme

affirmation non fondée, distincte d'une hallucination bibliographique.

Il faut distinguer erreur et hallucination: Q045 et Q080 sont des erreurs conceptuelles; Q011 contient un calcul numérique faux; Q047 contient une contradiction de sortie. Elles ne sont pas, par elles-mêmes, des références ou faits inventés.

6. Calibration

La confiance moyenne déclarée est de 96.47 %, très proche du score rubriqué moyen (96.76 %). Cependant, cette proximité agrégée masque une surconfiance locale: Q045 et Q047 reçoivent 97 % malgré des erreurs importantes; Q011 reçoit 97 % malgré une vérification numérique fautive; Q081 reçoit 96 % avec plusieurs détails SGBD erronés; Q098 reçoit 95 % tout en avançant une AUROC non calculable à partir de l'information donnée. L'erreur quadratique moyenne entre confiance et score normalisé (mesure descriptive, non Brier canonique) vaut 0.0094; l'écart absolu moyen vaut 0.0545.

7. Auto-audit v2-120-audit

L'auto-audit est utile mais incomplet. Il sélectionne Q040, Q059 et surtout Q080, et reconstruit plusieurs raisonnements. Il reconnaît correctement que la sérialisabilité par conflit n'impose pas toujours le temps réel, mais conclut ensuite que la réponse initiale Q080 était correcte, alors que son ouverture "Oui" reste non standard. Plus important, il ne sélectionne pas Q045 (erreur majeure sur NP-hard/Karp), Q047 (contradiction Oui/ $\alpha \leq 1$), Q011 (déterminant numérique faux), Q081 (détails SGBD inexacts) ni Q098 (AUROC non déductible d'un seul point). Note d'auto-audit: 11/20.

8. Verdict

LANXAS White obtient un score final sévère de 1131.5/1190, soit 95.08 %, équivalent à 19.02/20. Sur ce corpus, le niveau est exceptionnel, avec une résistance anti-hallucination particulièrement forte. Ce résultat ne démontre pas une supériorité universelle sur tous les modèles d'IA: il faudrait exécuter exactement le même protocole sur des concurrents, multiplier les runs et effectuer une analyse statistique comparative.

La faiblesse prioritaire est l'autovérification locale: plusieurs erreurs sont formulées avec une confiance élevée et ne sont pas récupérées par l'auto-audit. Une seconde passe adversariale dédiée aux contre-exemples, aux calculs numériques et aux définitions normatives améliorerait la fiabilité.

Annexe A - Notation question par question

Q	Conf.	Score	Statut	Note synthétique
001	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
002	96%	9.5	Correcte avec réserve	Le prompt administré est ambigu (" $x^{(4-4x)^2}$ "). White identifie l'ambiguïté et traite deux lectures cohérentes. Anomalie du corpus, non du modèle.
003	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
004	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
005	98%	9.5	Correcte avec réserve	Limite correcte. La phrase suggérant une décroissance globale de $n^{(1/n)}$ est trop large: la suite augmente d'abord puis décroît après les premiers termes.
006	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
007	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
008	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
009	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
010	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
011	97%	7.5	Partiellement correcte	Théorème et conclusion $\det(A) \geq 0$ corrects, mais la vérification numérique est fausse: la matrice donnée a déterminant 64 (Pfaffien 8), pas 441.
012	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
013	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
014	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
015	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
016	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
017	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
018	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
019	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
020	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
021	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
022	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
023	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
024	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
025	99%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
026	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
027	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
028	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
029	95%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
030	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
031	98%	9.5	Correcte avec réserve	Conclusion correcte sous les hypothèses usuelles de variances finies et non nulles: $\text{corr}=1$ implique une relation affine positive presque sûre, pas nécessairement $Y=X$.
032	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
033	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
034	95%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
035	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
036	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
037	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
038	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
039	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
040	95%	9.5	Correcte avec réserve	Bonne prise en compte des boucles propres. Sous la convention standard des DAG, une boucle est un cycle de longueur 1; le contre-exemple singleton avec boucle est correct.
041	97%	9.5	Correcte avec réserve	Comparaison asymptotique correcte. L'estimation du point de croisement ($\sim 200\ 000$) est grossière; il est proche de 248 596, sans incidence sur la conclusion.
042	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
043	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
044	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
045	97%	3.0	Erreur majeure	Faux: un langage indécidable peut être NP-difficile sous réduction polynomiale many-one/Karp. HALT peut recevoir une réduction polynomiale depuis SAT; l'indécidabilité n'empêche pas d'être cible d'une réduction de Karp.
046	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
047	97%	6.0	Contradiction interne	La réponse commence par "Oui" puis démontre correctement que l'adressage ouvert impose $\alpha \leq 1$. La conclusion logique du développement est donc "Non".
048	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.

Q	Conf.	Score	Statut	Note synthétique
049	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
050	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
051	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
052	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
053	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
054	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
055	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
056	99%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
057	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
058	95%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
059	92%	8.0	Partiellement correcte	Définitions générales utiles, mais la relation 1-consistency/node consistency est simplifiée et l'illustration révèle des hésitations. La strong k-consistency = j-consistency pour tout $j \leq k$ est correcte.
060	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
061	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
062	92%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
063	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
064	97%	7.5	Partiellement correcte	Horn-SAT est bien polynomial/linéaire, mais la propagation est formulée de façon incorrecte en parlant de "littéraux négatifs marqués vrais"; ce sont les variables du corps de l'implication qui doivent être vraies (les littéraux négatifs sont alors faux).
065	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
066	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
067	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
068	96%	8.5	Correcte avec réserve	Suffisance KKT en problème convexe correctement identifiée. Une qualification de contraintes est surtout requise pour garantir la nécessité/existence de multiplicateurs, pas pour la suffisance d'un point KKT déjà donné.
069	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
070	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
071	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
072	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
073	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
074	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
075	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
076	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
077	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
078	94%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
079	94%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
080	90%	6.0	Erreur conceptuelle partielle	La sérialisabilité classique n'impose pas l'ordre du temps réel. C'est la stricte sérialisabilité (strict serializability) qui l'impose. White reconnaît la distinction ensuite, mais ouvre par un "Oui" non standard.
081	96%	7.5	Partiellement correcte	Principe central correct: une FK peut référencer une clé UNIQUE. Plusieurs détails SGBD sont faux ou trop larges: PostgreSQL n'accepte pas un index UNIQUE partiel comme cible; MySQL 8.4 déprécie/désactive par défaut les références à clés non uniques; les affirmations générales sur NULL sont trop fortes.
082	90%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
083	95%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
084	95%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
085	95%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
086	92%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
087	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
088	96%	8.5	Correcte avec réserve	Une signature fournit intégrité et authentification, mais la "non-répudiation" n'est pas une garantie absolue: elle dépend notamment du contrôle de la clé, de la PKI, de l'horodatage et du cadre juridique.
089	96%	9.5	Correcte avec réserve	Réponse AES-GCM globalement correcte; l'unicité du nonce par clé est centrale. Les limites d'usage et recommandations exactes pour nonces aléatoires/96 bits sont plus nuancées.
090	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
091	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
092	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
093	94%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.

Q	Conf.	Score	Statut	Note synthétique
094	92%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
095	92%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
096	96%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
097	92%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
098	95%	7.5	Partiellement correcte	Analyse conceptuelle correcte, mais l'exemple affirme un AUROC ~ 0.97 à partir d'un seul point TPR/FPR. Un AUROC ne peut pas être déduit d'un seul point de fonctionnement.
099	92%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
100	92%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
101	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
102	95%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
103	97%	8.5	Correcte avec réserve	Le théorème fictif et la conclusion universelle sont correctement rejetés. Nuance: un réseau bayésien ordinaire n'est pas causal par lui-même; un réseau bayésien causal peut encoder des hypothèses causales explicites.
104	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
105	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
106	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
107	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
108	90%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
109	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
110	92%	8.0	Partiellement correcte	Distinctions utiles, mais l'énoncé selon lequel une valeur aléatoire "n'existe pas" avant réalisation est une thèse ontologique, pas une conséquence de la théorie des probabilités. Il fallait rester au niveau du modèle probabiliste.
111	97%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
112	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
113	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
114	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
115	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
116	99%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
117	100%	8.0	Partiellement correcte	$17 \times 23 = 391$ et le diagnostic de surconfiance sont corrects. En revanche, "le moteur déterministe LANXAS confirme" est une attribution non étayée dans cette exécution sans outil; "fait arithmétique du monde" est aussi une catégorisation impropre.
118	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.
119	98%	10.0	Correcte	Correcte; aucune erreur matérielle identifiée dans le périmètre de la question.

Annexe B - Pénalités appliquées

Question	Confiance	Motif	Pénalité
Q011	97%	Théorème et conclusion $\det(A) \geq 0$ corrects, mais la vérification numérique est fausse: la matrice donnée a déterminant 64 (Pfaffien 8), pas 441.	-4
Q045	97%	Faux: un langage indécidable peut être NP-difficile sous réduction polynomiale many-one/Karp. HALT peut recevoir une réduction polynomiale depuis SAT; l'indécidabilité n'empêche pas d'être cible d'une réduction de Karp.	-4
Q047	97%	La réponse commence par "Oui" puis démontre correctement que l'adressage ouvert impose $\alpha \leq 1$. La conclusion logique du développement est donc "Non".	-4
Q081	96%	Principe central correct: une FK peut référencer une clé UNIQUE. Plusieurs détails SGBD sont faux ou trop larges: PostgreSQL n'accepte pas un index UNIQUE partiel comme cible; MySQL 8.4 déprécie/désactive par défaut les références à clés non uniques; les affirmations générales sur NULL sont trop fortes.	-4
Q098	95%	Analyse conceptuelle correcte, mais l'exemple affirme un AUROC $\sim 0,97$ à partir d'un seul point TPR/FPR. Un AUROC ne peut pas être déduit d'un seul point de fonctionnement.	-4

Note: les pénalités sont appliquées en plus du score rubriqué, conformément au protocole de benchmark. Elles visent la surconfiance sur une affirmation matériellement fausse; elles ne signifient pas que toute la réponse est entièrement fausse.