Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Les données de calibration comptent-elles vraiment pour LoRA?

Benedictus Kent Rachmat[1, 2] Thomas Gerald[1] Zheng Zhang[2] Cyril Grouin[1]
(1) Université Paris-Saclay, CNRS, LISN, Orsay, France
(2) Embedded AI Lab, SLB, Clamart, France
rachmat@lisn.fr

RÉSUMÉ

Les méthodes récentes d'adaptation efficace en paramètres, notamment LoRA, utilisent parfois quelques exemples du domaine cible pour configurer les adapteurs avant l'entraînement. Nous étudions ce que mesurent réellement ces données de calibration à travers une décomposition de variance menée sur plusieurs modèles et jeux de données. Nos résultats montrent que les signaux d'activation sont fortement dominés par l'identité du module, qui explique jusqu'à 95 % de la variance, tandis que le choix du jeu de données a un effet limité. Les signaux de gradient suivent une dynamique différente : ils présentent de fortes interactions module–jeu de données, suggérant une sensibilité plus directe au domaine cible. Nous montrons en outre que les activations et les gradients ne conduisent pas aux mêmes choix de modules : les modules à forte énergie d'activation tendent à avoir une faible énergie de gradient. Ces résultats remettent en question l'idée selon laquelle les activations suffisent à guider la configuration de LoRA et ouvrent la voie à des méthodes d'adaptation davantage fondées sur les gradients.

ABSTRACT

Do Calibration Data Really Matter for LoRA ?

Recent parameter-efficient adaptation methods, notably LoRA, sometimes use a small set of target-domain examples to configure adapters before training. We study what these calibration data actually measure through a variance decomposition across multiple models and datasets. Our results show that activation-based signals are strongly dominated by module identity, which explains up to 95% of the variance, while dataset choice has only a limited effect. Gradient-based signals follow a different pattern : they exhibit strong module–dataset interactions, suggesting a more direct sensitivity to the target domain. We further show that activations and gradients do not lead to the same module choices : modules with high activation energy tend to have low gradient energy. These results challenge the idea that activations alone are sufficient to guide LoRA configuration and open the way to adaptation methods more strongly grounded in gradients.

MOTS-CLÉS : Méthodologie d'évaluation, LoRA, décomposition de variance.

KEYWORDS: Evaluation methodology, LoRA, variance decomposition.

1 Introduction

Les grands modèles de langue fondés sur l'architecture Transformer (Vaswani *et al.*, 2017) sont aujourd'hui adaptés à des tâches ou domaines cibles au moyen de méthodes efficaces en paramètres, qui évitent de réentraîner l'ensemble du modèle. Parmi elles, LoRA (*Low-Rank Adaptation*) (Hu

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 111–121.

et al., 2022) est devenue une approche centrale, en recherche comme en pratique, car elle permet d'adapter de grands modèles avec un faible coût d'entraînement et sans surcoût notable à l'inférence.

LoRA consiste à figer les poids préentraînés et à apprendre une correction de faible rang pour certaines transformations linéaires. Pour une matrice préentraînée $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA remplace la transformation Wx par

$$(W + \Delta W)x, \qquad \Delta W = BA,$$

où $A \in \mathbb{R}^{r \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$, et $r \ll \min(d_{\text{in}}, d_{\text{out}})$. La capacité d'adaptation dépend donc directement du choix des modules adaptés, du rang r, de l'initialisation des matrices A et B, ainsi que des hyperparamètres d'entraînement.

Ces choix sont souvent fixés uniformément ou par heuristique. Plusieurs travaux récents cherchent au contraire à les déterminer automatiquement, soit à partir des poids préentraînés eux-mêmes, soit à partir d'un petit ensemble d'exemples du domaine cible. Nous appelons cette seconde étape *calibration* : le modèle reste figé, mais les exemples servent à extraire des statistiques utilisées pour configurer LoRA. Ces statistiques peuvent provenir de la propagation avant, par exemple les activations, ou de la propagation arrière, par exemple les gradients.

Une hypothèse implicite de nombreuses méthodes calibrées est que ces signaux contiennent une information spécifique au domaine cible. Si cette hypothèse est correcte, changer le jeu de données de calibration devrait modifier de manière substantielle les modules considérés comme importants. Pourtant, plusieurs observations récentes remettent cette intuition en question. Des méthodes telles que PiSSA (Meng *et al.*, 2024) et MiLoRA (Wang *et al.*, 2025), qui reposent principalement sur la structure spectrale des poids préentraînés, atteignent des performances compétitives sans utiliser de données de calibration. De même, des travaux sur la dimensionnalité intrinsèque (Aghajanyan *et al.*, 2021) et la stabilité des représentations (Huh *et al.*, 2024) suggèrent qu'une part importante de la structure du fine-tuning est déjà déterminée par le modèle lui-même.

Ces résultats soulèvent une question simple mais peu étudiée systématiquement :

Les signaux de calibration utilisés pour configurer LoRA reflètent-ils réellement le domaine cible, ou sont-ils principalement déterminés par l'architecture et l'identité des modules ?

Pour répondre à cette question, nous analysons six signaux de calibration issus des activations et des gradients, sur trois modèles décodeurs et sept jeux de données couvrant des domaines variés. Pour chaque modèle, chaque jeu de données et chaque module linéaire, nous extrayons des statistiques de propagation avant et arrière, puis nous décomposons leur variance selon deux facteurs : l'identité du module et le jeu de données utilisé pour la calibration.

Nos contributions sont les suivantes :

- Nous proposons un protocole de diagnostic permettant de mesurer si les signaux de calibration pour LoRA dépendent principalement du module ou du jeu de données cible.
- Nous montrons que les signaux basés sur les activations sont très majoritairement expliqués par l'identité du module, tandis que l'effet principal du jeu de données reste faible.
- Nous observons que les signaux de gradient présentent davantage d'interactions module–jeu de données, suggérant qu'ils capturent une information plus liée au besoin d'adaptation.

Nos résultats suggèrent que les statistiques d'activation utilisées pour la calibration reflètent principalement des propriétés architecturales déjà présentes dans le modèle préentraîné. Elles ne doivent donc pas être interprétées automatiquement comme des indicateurs spécifiques au domaine cible. Les

gradients, en revanche, semblent offrir un signal plus sensible à l'interaction entre module et données, ce qui motive leur étude pour de futures méthodes d'adaptation.

Le reste de l'article est organisé comme suit. La section 2 présente les travaux connexes sur l'initialisation et la configuration de LoRA. La section 3 décrit notre protocole expérimental, les modèles, les jeux de données et les signaux étudiés. La section 4 présente les résultats, suivis d'une discussion sur leurs implications et leurs limites.

2 Travaux connexes

Méthodes étudiées. Afin de répondre à notre question de recherche, nous nous appuyons sur trois familles d'extraction de statistiques. La première, basée sur les activations, regroupe les méthodes fondées sur l'extraction du rang des représentations intermédiaires (capturées lors de la propagation avant du réseau de neurones). La seconde famille regroupe les méthodes s'appuyant sur les gradients (capturés lors de la rétropropagation du gradient) ; en particulier, exploitées dans LoRA-GA (Wang *et al.*, 2024), LoRA-One (Zhang *et al.*, 2025a), et GoRA (haonan he *et al.*, 2025). La troisième famille repose uniquement sur les poids/paramètres des modèles (des transformations linéaires) regroupant les méthodes telles que PiSSA (Meng *et al.*, 2024), MiLoRA (Wang *et al.*, 2025) et DoRA (Liu *et al.*, 2024). Pour l'adaptation, ces dernières approches ne se basent pas sur un jeu de données du domaine cible (jeux de données de calibration), mais uniquement sur les informations extraites de l'architecture et/ou des paramètres.

Indices partiels d'indépendance aux données. Plusieurs travaux observent l'indépendance des données au modèle sans la quantifier. CorDA (Yang *et al.*, 2024) rapporte, dans son étude d'ablation sur la source de calibration, des performances quasi identiques entre Wikitext-2, NQ Open et TriviaQA (variation inférieure à 1,5 % de score moyen). Concernant l'initialisation, Zhang *et al.* (2025b) montrent que le bénéfice de la décomposition spectrale réside dans l'amplification de la magnitude des mises à jour, et Lee *et al.* (2026) montrent qu'un taux d'apprentissage adapté suffit pour LoRA standard pour atteindre les performances des variantes spectrales. Biderman *et al.* (2024) montrent que LoRA opère à un rang effectif 10–100$\times$ plus faible que le fine-tuning de l'intégralité des paramètres du modèle, et Shuttleworth *et al.* (2025) identifient des « dimensions inattendues » propres à LoRA, soulignant le caractère spécifique du module.

Structure déterminée par le modèle. Plusieurs résultats convergent vers l'idée que la structure des réseaux est déterminée par le modèle. Aghajanyan *et al.* (2021) montrent que le fine-tuning occupe un sous-espace de faible dimension déterminé par le modèle, l'hypothèse des représentations stables (Huh *et al.*, 2024) suggère une convergence des représentations à travers différents modèles, domaines et modalités, malgré des objectifs d'apprentissage distincts, et Martin & Mahoney (2021); Martin *et al.* (2021) montrent que les spectres des poids suffisent à sélectionner les dimensions à entraîner. Staats *et al.* (2025) établissent un chevauchement entre vecteurs singuliers des poids et vecteurs propres des matrices de covariance des activations, et Jaiswal *et al.* (2025) distinguent les modules (Q, K, V) selon leur structure spectrale. Concernant les gradients, Zhao *et al.* (2024) montrent que leurs sous-espaces se stabilisent rapidement.

3 Protocole expérimental

3.1 Modèles et jeux de données

Nous testons trois décodeurs (transformeurs) préentraînés, en considérant deux architectures (Llama et Qwen) avec des nombres de paramètres différents (Tableau 3.1). Chaque couche comporte sept projections linéaires Q, K, V, O pour l'attention, ainsi que $\mathrm{Gate}, \mathrm{Up}, \mathrm{Down}$ pour le perceptron multicouche. Nous considérons chacune de ces transformations linéaires comme un *module* distinct, identifié par son indice de couche et son type de projection.

Modèle	**Dimensions**	**Couches**	**Proj.**	**Modules**
Llama-3.2-1B-Instruct	2048	16	7	112
Qwen-2.5-7B-Instruct	3584	28	7	196
Llama-3.1-8B-Instruct	4096	32	7	224
			Total	**532**

TABLE 1 – Les trois modèles utilisés. Modules = couches $\times$ 7 projections.

Nous considérons 7 jeux de données couvrant des domaines variés, comme présenté dans le Tableau 2. Ces domaines sont choisis pour être aussi distincts que possible (domaine ou tâche), afin de maximiser la probabilité d'observer un éventuel signal de calibration dépendant du domaine. Notons que nos expériences ne portent que sur l'anglais.

Jeu de données (réf.)	**Domaine**
Financial PhraseBank (Malo *et al.*, 2014)	Finance
AG News (Zhang *et al.*, 2015)	Actualités
SciQ (Welbl *et al.*, 2017)	Questions scientifiques
ACL-ARC (Jurgens *et al.*, 2018)	Intention de citation (NLP)
BoolQ (Clark *et al.*, 2019)	Lecture (domaine ouvert)
PubMed QA (Jin *et al.*, 2019)	Biomédical
GSM8K (Cobbe *et al.*, 2021)	Mathématiques

TABLE 2 – Jeux de données couvrant des domaines variés en anglais, sélectionnés pour maximiser la diversité des distributions et évaluer la dépendance au domaine des signaux de calibration.

Pour chaque paire (modèle, jeu de données), nous effectuons une seule passe de 1 024 exemples dans le modèle préentraîné, et nous étudions également la convergence des signaux à différents budgets d'exemples $n \in \{32, 64, 256, 512, 1024\}$. Nous prenons à chaque fois les n premiers exemples du dataset.

3.2 Extraction des signaux

Pour chaque cellule (modèle, jeu de données), nous faisons passer 1024 exemples non annotés à travers le modèle, calculons la fonction de perte de génération auto-régressive (log-vraisemblance négative), puis faisons une étape de rétropropagation du gradient. Pour chacun des 7 modules linéaires

de chaque couche ($Q, K, V, O, \mathrm{Gate}, \mathrm{Up}, \mathrm{Down}$), nous capturons deux quantités au cours de chacun de ces passages :

— Lors de la **propagation avant**, nous capturons la représentation d'entrée du module, $z_i \in \mathbb{R}^{d_i}$ pour estimer la variance par dimension.

— Lors de la **propagation arrière**, le gradient de la fonction de perte par rapport à cette même entrée, $\nabla_{z_i}\mathcal{L}$, est capturé et est utilisé pour estimer la variance par dimension. Nous capturons donc le gradient de la représentation cachée fournie en *entrée* du module, et non celui des poids du module.

À partir de ces statistiques, nous calculons six signaux scalaires par module, résumés dans le Tableau 3, Nous notons Σ_T (resp. Σ_∇) la matrice de covariance calculée sur l'ensemble des représentations capturées lors de la propagation avant (resp. propagation arrière) pour un module, une couche pour les n exemples.

Signal	**Formule**	**Interprétation**
Trace d'activation	$\mathrm{tr}(\mathbf{\Sigma}_T)$	Énergie totale des entrées du module
Norme moyenne	$\|\boldsymbol{\mu}_T\|^2$	Composante du décalage moyen
Ratio act./poids	$\mathrm{tr}(\mathbf{\Sigma}_T)/\|W\|_F^2$	Énergie des activations relative aux poids
Ratio d'outliers	$\mathrm{std}\|z_k\|/\mathrm{mean}\|z_k\|$	Coefficient de variation des normes de tokens
Trace du gradient	$\mathrm{tr}(\mathbf{\Sigma}_\nabla)$	Énergie de l'erreur rétropropagée au module
Trace de Fisher	$\mathrm{tr}(\mathbf{\Sigma}_T)\cdot\mathrm{tr}(\mathbf{\Sigma}_\nabla)$	Proxy diagonal du bloc de Fisher par module

TABLE 3 – Les six signaux extraits par module. Les quatre premiers proviennent du passage avant, les deux derniers du passage arrière.

Les signaux d'activation résument ce que le module *reçoit*. Ce sont notamment les quantités que CorDA (Yang *et al.*, 2024) utilise pour allouer le rang. Les signaux de gradient résument plutôt l'erreur rétropropagée au module, et fournissent donc un indicateur plus direct du besoin potentiel d'adaptation. Ces deux familles sont parfois utilisées comme critères d'importance, sans que leur relation soit toujours explicitement analysée.

3.3 Décomposition par ANOVA

Pour chaque couple (modèle, signal), nous construisons une matrice $M \times D$ notée S, où M est le nombre de modules et $D = 7$ le nombre de jeux de données. S_{ij} correspond à la valeur du signal pour le module i lorsque le corpus de calibration est le jeu de données j. Une ANOVA à deux facteurs sans répétition décompose la somme totale des carrés (avec $\bar{S}$ la moyenne arithmétique) :

$$\mathrm{SS}_{\mathrm{tot}} = \sum_{i,j}(S_{ij} - \bar{S})^2, \tag{1}$$

$$\mathrm{SS}_{\mathrm{mod}} = D\sum_{i}(\bar{S}_{i\cdot} - \bar{S})^2, \qquad \mathrm{SS}_{\mathrm{data}} = M\sum_{j}(\bar{S}_{\cdot j} - \bar{S})^2, \tag{2}$$

$$\mathrm{SS}_{\mathrm{res}} = \mathrm{SS}_{\mathrm{tot}} - \mathrm{SS}_{\mathrm{mod}} - \mathrm{SS}_{\mathrm{data}}. \tag{3}$$

Exprimées en pourcentage de $\mathrm{SS}_{\mathrm{tot}}$, ces quantités permettent d'identifier les principales sources de la variance. Une valeur élevée de $\mathrm{SS}_{\mathrm{mod}}$ indique que le signal est principalement déterminé par l'identité

du module, tandis qu'une valeur importante de $\mathrm{SS}_{\mathrm{data}}$ reflète une influence du jeu de données. À l'inverse, une contribution élevée du résidu $\mathrm{SS}_{\mathrm{res}}$ suggère que le signal ne peut pas être expliqué par une unique dimension (module ou jeu de données) étudiée, mais dépend d'interactions entre les deux dimensions.

4 Résultats

4.1 Signaux extraits

Dans la figure 1 nous reportons les résultats de la trace de la matrice de covariance des activations moyennés sur les différents modèles et les différents jeux de données pour chaque module. Nous observons que les tendances de $\mathrm{tr}(\mathbf{\Sigma}_T)$ sont similaires, indépendamment du modèle ou du jeu de données avec une variance faible (les bandes min–max se confondant presque avec la courbe moyenne). En suivant la distinction proposée par Jaiswal *et al.* (2025), nous distinguons les projections LRC (*Low-Rank Components*) et N-LRC (*Non-Low-Rank Components*). Les LRC correspondent à des matrices de poids bien approximées par une structure de faible rang, tandis que les N-LRC présentent une structure plus diffuse.

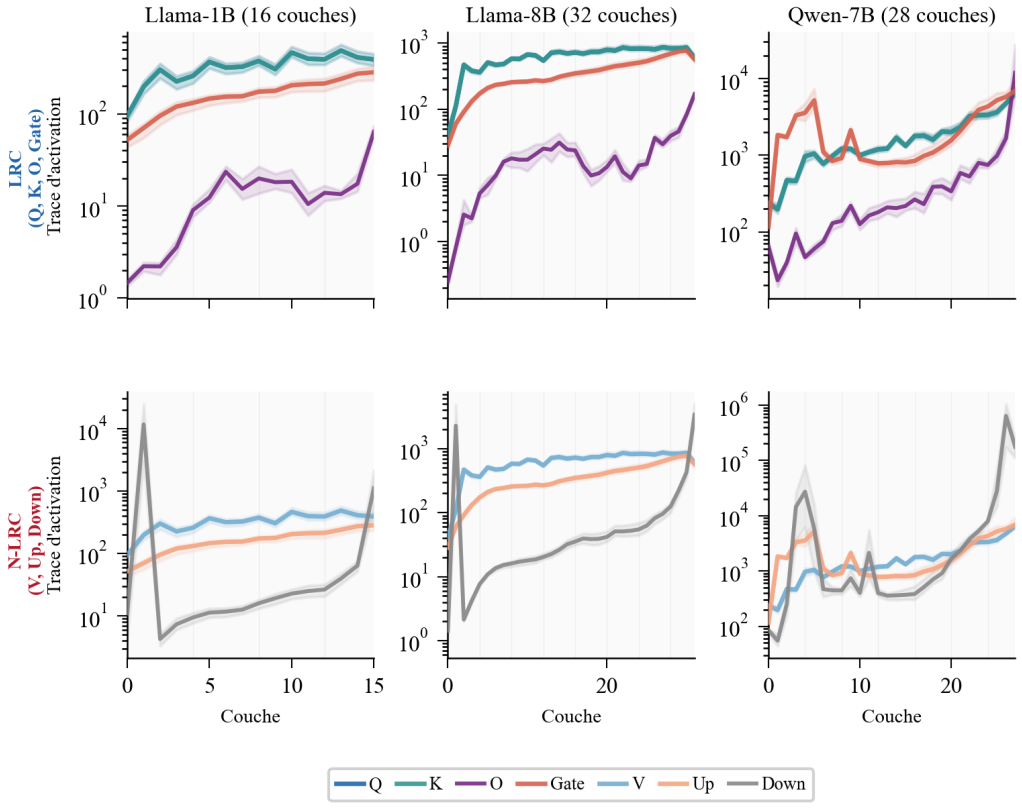

FIGURE 1 – Profils de $\mathrm{tr}(\mathbf{\Sigma}_T)$ par couche pour les trois modèles. Haut : projections LRC (Q, K, O, Gate). Bas : projections N-LRC $(V, \mathrm{Up}, \mathrm{Down})$ (Jaiswal *et al.*, 2025). Bandes : min–max sur 7 jeux de données

Dans la Figure 2, pour chaque module et chaque couche, nous représentons la valeur de la trace pour n=32 et n=1024 dans la sous-figure de gauche. Dans la sous-figure droite, nous représentons la

corrélation de Spearman entre le classement des modules obtenu avec différents budgets n et celui obtenu avec $n = 1024$. Cette mesure évalue la stabilité de l'ordre relatif des modules, et non une corrélation linéaire entre les valeurs brutes.

On observe une très forte stabilité du classement des modules dès $n = 32$. Aussi, tous les modèles atteignent une valeur $\rho \geq 0{,}99$ dès n=32. L'observation de 32 exemples semble représentatif des activations ; il est donc, pour les jeux de données étudiés, non nécessaire d'augmenter le nombre d'exemples pour la calibration dans le cadre de l'adaptation de modèles.

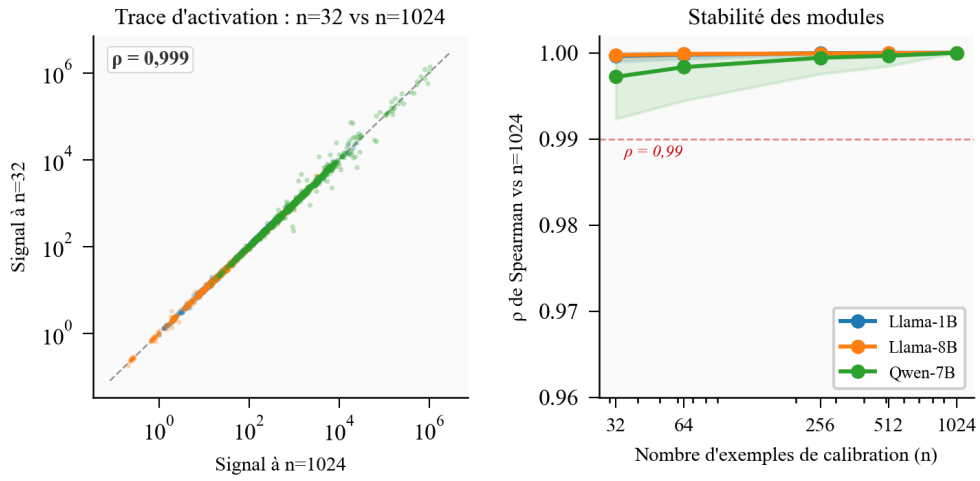

FIGURE 2 – Stabilité des modules selon $\mathrm{tr}(\mathbf{\Sigma}_T)$. Gauche : signal à n=32 vs n=1024 ($\rho = 0{,}999$). Droite : ρ vs n=1024 en fonction de n, seuil ρ=0,99 atteint dès n=32

4.2 L'identité du module explique le signal

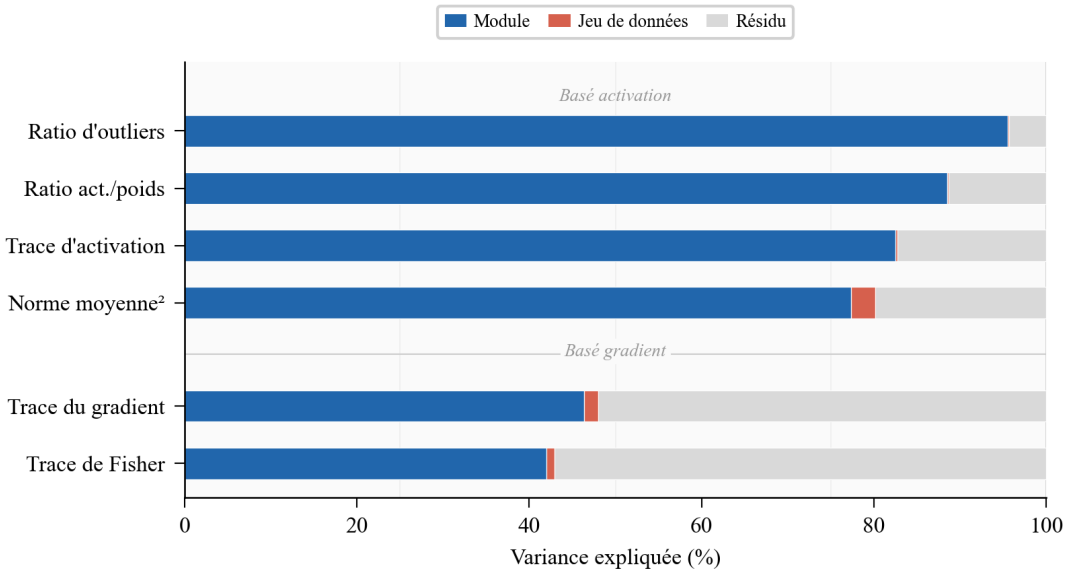

FIGURE 3 – ANOVA à deux facteurs sur six signaux (moyenne sur 3 modèles). Bleu : module. Rouge : jeu de données. Gris : résidu (interaction)

Si les signaux de calibration reflétaient principalement le domaine cible, la part de variance expliquée par le jeu de données devrait être importante. Or, la Figure 3 montre l'inverse : pour tous les signaux, l'effet principal du jeu de données reste quasi nul (0–3 %). Pour les signaux d'activation, l'identité du module domine largement la variance expliquée, avec 77–95 % selon le signal. Le résidu reste limité, ce qui indique que ces signaux varient peu avec le jeu de données de calibration. Les signaux de gradient suivent un comportement différent. La part expliquée par le module diminue fortement, autour de 42–47 %, tandis que le résidu devient majoritaire, jusqu'à 57 %. Comme l'effet principal du jeu de données reste faible, cette variance résiduelle suggère surtout une interaction module–jeu de données : certains modules deviennent importants pour certains jeux de données, sans qu'un jeu de données soit globalement dominant.

Ces résultats indiquent que les signaux d'activation se comportent principalement comme des empreintes architecturales du modèle, plutôt que comme des indicateurs spécifiques au domaine. Cela peut expliquer pourquoi des méthodes fondées sur la structure du modèle, comme PiSSA (Meng *et al.*, 2024) ou MiLoRA (Wang *et al.*, 2025) peuvent être compétitives sans calibration spécifique. À l'inverse, les gradients semblent offrir un signal plus directement lié aux besoins d'adaptation, car ils capturent davantage l'interaction entre modules et données.

4.3 Les signaux d'activation et de gradient divergent

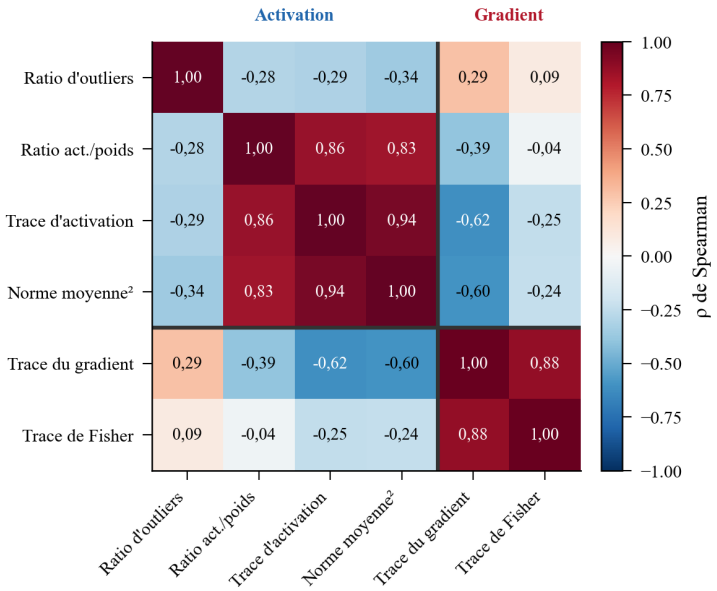

FIGURE 4 – Corrélation de Spearman entre les six signaux, moyennée sur les couples (modèle, jeu de données). Les blocs hors diagonale révèlent une corrélation négative entre familles activation et gradient

Une hypothèse naturelle, lorsque des méthodes d'importance basées sur les activations et sur les gradients sont considérées comme interchangeables, est que ces deux familles estiment la même importance des modules, éventuellement avec des coûts d'échantillonnage ou de calcul différents. Dans ce cas, les corrélations inter-familles devraient être proches de $+1$. La Figure 4 montre que les signaux d'activation sont fortement redondants entre eux (corrélations jusqu'à $0{,}94$), de même que les signaux de gradient (ρ=0,88). En revanche, entre les deux familles, la corrélation est négative

($\rho = -0{,}62$ entre $\mathrm{tr}(\mathbf{\Sigma}_T)$ et $\mathrm{tr}(\mathbf{\Sigma}_\nabla)$, $\rho = -0{,}60$ entre $\|\boldsymbol{\mu}_T\|^2$ et $\mathrm{tr}(\mathbf{\Sigma}_\nabla)$) : les modules à forte énergie d'activation ont tendance à présenter une faible énergie de gradient.

5 Discussion et conclusion

Notre analyse montre que les signaux de calibration fondés sur les activations sont principalement déterminés par l'identité du module : celle-ci explique 77–95 % de la variance, contre seulement 0–3 % pour le jeu de données. Ces signaux apparaissent donc davantage comme des empreintes architecturales que comme des indicateurs propres au domaine cible. À l'inverse, les signaux de gradient présentent une interaction module–données plus marquée (51–57 %), suggérant qu'ils capturent une information plus liée à l'erreur et au besoin d'adaptation.

Ces résultats éclairent le rôle de la calibration dans LoRA. Ils suggèrent que l'allocation non uniforme du rang peut être pertinente, mais que les statistiques d'activation ne constituent pas nécessairement le meilleur critère pour l'estimer, puisqu'elles sont largement prédictibles à partir de la structure du modèle. La corrélation négative entre signaux d'activation et de gradient ($\rho = -0{,}62$) renforce cette conclusion : un module fortement activé n'est pas forcément un module qui nécessite une adaptation importante. Les futures méthodes LoRA pourraient donc s'appuyer davantage sur les gradients pour guider l'allocation du rang, l'initialisation des matrices A et B, ou des taux d'apprentissage sélectifs par module.

Notre travail ne propose pas encore une nouvelle méthode LoRA complète, mais fournit un diagnostic empirique des signaux de calibration utilisés pour configurer ces adaptateurs. Les conclusions doivent donc être interprétées comme une motivation pour des méthodes plus sensibles à la tâche, et non comme la validation d'une nouvelle méthode d'adaptation. Une étape naturelle sera de comparer, par fine-tuning, des méthodes fondées uniquement sur les poids, des méthodes basées sur les activations et des méthodes exploitant les gradients.

Limites. Cette étude analyse des signaux extraits d'un modèle figé avant entraînement ; elle ne mesure pas encore leur effet sur les performances après fine-tuning LoRA. De plus, les expériences portent uniquement sur des jeux de données en anglais et sur des modèles décodeurs de type Transformer. Il reste donc à vérifier si les mêmes tendances se maintiennent dans des contextes multilingues, multimodaux ou avec d'autres architectures.

Aspects éthiques, reproductibilité et remerciements. Les jeux de données utilisés dans ce travail sont des ressources publiques couramment employées pour l'évaluation des modèles de langue. Comme nos expériences ne portent que sur l'anglais, elles ne permettent toutefois pas de conclure sur le comportement de ces méthodes dans des langues moins représentées. Ce travail a bénéficié d'un accès aux ressources de calcul haute performance de GENCI–IDRIS dans le cadre de l'allocation AD011016508.

Références

AGHAJANYAN A., GUPTA S. & ZETTLEMOYER L. (2021). Intrinsic dimensionality explains the

effectiveness of language model fine-tuning. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 7319–7328, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.568.

BIDERMAN D., PORTES J., ORTIZ J. J. G., PAUL M., GREENGARD P., JENNINGS C., KING D., HAVENS S., CHILEY V., FRANKLE J., BLAKENEY C. & CUNNINGHAM J. P. (2024). LoRA learns less and forgets less. *Transactions on Machine Learning Research*.

CLARK C., LEE K., CHANG M.-W., KWIATKOWSKI T., COLLINS M. & TOUTANOVA K. (2019). BoolQ : Exploring the surprising difficulty of natural yes/no questions. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 2924–2936, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1300.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R., HESSE C. & SCHULMAN J. (2021). Training verifiers to solve math word problems. *CoRR*, **abs/2110.14168**.

HAONAN HE, YE P., REN Y., YUAN YUAN, LUYANGZHOU, SHUCUNJU & LEI CHEN (2025). GoRA : Gradient-driven adaptive low rank adaptation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

HU E. J., YELONG SHEN, WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). LoRA : Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

HUH M., CHEUNG B., WANG T. & ISOLA P. (2024). The platonic representation hypothesis.

JAISWAL A. K., WANG Y., YIN L., LIU S., CHEN R., ZHAO J., GRAMA A., TIAN Y. & WANG Z. (2025). From low rank gradient subspace stabilization to low-rank weights : Observations, theories, and applications. In *Forty-second International Conference on Machine Learning*.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). PubMedQA : A dataset for biomedical research question answering. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1259.

JURGENS D., KUMAR S., HOOVER R., MCFARLAND D. & JURAFSKY D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, **6**, 391–406. DOI : 10.1162/tacl_a_00028.

LEE Y.-A., KO C.-Y., CHEN P.-Y. & YEH M.-Y. (2026). Learning rate matters : Vanilla LoRA may suffice for LLM fine-tuning. arxiv.org/abs/2602.04998.

LIU S.-Y., WANG C.-Y., YIN H., MOLCHANOV P., WANG Y.-C. F., CHENG K.-T. & CHEN M.-H. (2024). DoRA : weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria.

MALO P., SINHA A., KORHONEN P., WALLENIUS J. & TAKALA P. (2014). Good debt or bad debt : Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, **65**(4), 782–796. DOI : 10.1002/asi.23062.

MARTIN C. H. & MAHONEY M. W. (2021). Implicit self-regularization in deep neural networks : Evidence from random matrix theory and implications for training. *Journal of Machine Learning Research*, **22**(165), 1–73.

MARTIN C. H., PENG T. & MAHONEY M. W. (2021). Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, **12**(1). DOI : 10.1038/s41467-021-24025-8.

MENG F., WANG Z. & ZHANG M. (2024). PiSSA : Principal singular values and singular vectors adaptation of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

SHUTTLEWORTH R., ANDREAS J., TORRALBA A. & SHARMA P. (2025). LoRA vs full fine-tuning : An illusion of equivalence. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*.

STAATS M., THAMM M. & ROSENOW B. (2025). Small singular values matter : A random matrix analysis of transformer models. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Red Hook, NY, USA : Curran Associates Inc.

WANG H., LI Y., WANG S., CHEN G. & CHEN Y. (2025). Milora : Harnessing minor singular components for parameter-efficient llm finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 4823–4836.

WANG S., YU L. & LI J. (2024). LoRA-GA : Low-rank adaptation with gradient approximation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

WELBL J., LIU N. F. & GARDNER M. (2017). Crowdsourcing multiple choice science questions. In L. DERCZYNSKI, W. XU, A. RITTER & T. BALDWIN, Éds., *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 94–106, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4413.

YANG Y., LI X., ZHOU Z., SONG S. L., WU J., NIE L. & GHANEM B. (2024). CorDA : Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

ZHANG X., ZHAO J. & LECUN Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, p. 649–657, Cambridge, MA, USA : MIT Press.

ZHANG Y., LIU F. & CHEN Y. (2025a). LoRA-one : One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Forty-second International Conference on Machine Learning*.

ZHANG Z., LI H., ZHANG Y., GONG G., WANG J., PENGZHANG LIU, JIANG Q. & HU J. (2025b). The primacy of magnitude in low-rank adaptation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

ZHAO J., ZHANG Z., CHEN B., WANG Z., ANANDKUMAR A. & TIAN Y. (2024). GaLore : Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning*.