OLORA+: ГИБРИДНЫЙ ПОДХОД К LORA

Давронов Рифкат Рахимович[1], Кушмуратов Самариддин Ибодулла угли[2]

[1]канд. техн. наук, старший научный сотрудник Институт Математики имени В.И.Романовского Академии наук Республики Узбекистан
E-mail: rifqat.davronov@mathinst.uz
[2]младший научный сотрудник Институт Математики имени В.И.Романовского Академии наук Республики Узбекистан
E-mail: bekmezonali@gmail.com

K E Y W O R D S

NLP, PEFT, LLM, LoRA, OLoRA, LoRA+.

A B S T R A C T

В данной статье представлен и проанализирован OLoRA+, новый метод тонкой настройки с эффективным использованием параметров, который улучшает Low-Rank Adaptation (LoRA) путем объединения ортонормированной инициализации OLoRA с оптимизацией дифференциальной скорости обучения LoRA+. В исследовании сравнивалась производительность OLoRA+ со стандартным базовым уровнем OLoRA на модели TinyLlama-1.1B-Chat-v1.0. Используя подмножество набора данных alpaca, коллекцию из 52 тысяч демонстраций следования инструкциям, с 1000 образцами для обучения и 500 для тестирования, производительность модели оценивалась с использованием метрик evaluation loss, BLEU и ROUGE. Полученные результаты показывают, что OLoRA+ стабильно превосходит базовый уровень OLoRA по всем рассмотренным метрикам. Кроме того, исследование показывает, что OLoRA+ эффективен при коэффициентах скорости обучения как больше, так и меньше единицы, раскрывая новый компромисс между стратегиями обучения «Refinement» и «Exploration». Это подтверждает потенциал OLoRA+ как более универсального и мощного подхода для адаптации LLM в условиях ограниченных ресурсов.

1. Введение

Появление предварительно обученных больших языковых моделей (LLM) значительно продвинуло область обработки естественного языка (NLP) [1]. Эти модели демонстрируют мощные возможности в общем понимании и генерации языка. Однако растущий размер LLM представляет значительные проблемы для традиционной полной тонкой настройки, которая обновляет все параметры модели и влечет за собой существенные вычислительные затраты и затраты памяти [2].

В ответ на эти проблемы методы Parameter-Efficient Fine-Tuning (PEFT) стали многообещающим решением. Методы PEFT адаптируют предварительно обученные модели путем выборочной тонкой настройки небольшого количества дополнительных параметров, сохраняя при этом большую часть исходной модели замороженной. Среди них Low-Rank Adaptation (LoRA) [5] стал одним из самых популярных подходов, достигающим высокой производительности без увеличения задержки вывода. Успех LoRA стимулировал разработку экосистемы вариантов, каждый из которых направлен на устранение конкретных ограничений исходного метода.

Исследования по улучшению LoRA развивались по нескольким различным направлениям. С одной стороны, такие

методы, как **OLoRA**, сосредоточены на **инициализации** адаптера [3]. OLoRA использует QR-разложение для создания ортонормированного базиса из предварительно обученных весов, обеспечивая более стабильную и хорошо обусловленную отправную точку для обучения, что приводит к ускоренной сходимости. С другой стороны, подходы, такие как **LoRA+**, сосредоточены на **оптимизации** процесса обучения. LoRA+ демонстрирует, что использование дифференциальных скоростей обучения для двух матриц адаптера (A и B) может значительно ускорить обучение и улучшить конечную производительность [4].

Однако эти направления исследований — улучшение инициализации и оптимизация процесса обучения — в значительной степени развивались параллельно. Потенциальные синергетические эффекты их комбинации остаются неисследованной областью.

В данной статье устраняется этот пробел путем введения **OLoRA+**, нового гибридного метода, который синергетически сочетает структурированную инициализацию OLoRA с ускоренной оптимизацией LoRA+. Основная цель этого исследования — изучить взаимодействие между этими двумя методами и определить, дает ли их комбинация более эффективный и универсальный метод PEFT. Путем эмпирической оценки мы не только сравниваем производительность OLoRA+ с базовым уровнем OLoRA, но и раскрываем новую динамику обучения, связанную с выбором коэффициента скорости обучения. Мы характеризуем эти режимы как стратегии **«Refinement»** и **«Exploration»**, демонстрируя, что OLoRA+ превосходит своих предшественников и предлагает более глубокое понимание взаимодействия между инициализацией адаптера и оптимизацией https://github.com/kushmuratoff/OLoRA-plus.git .

2. Связанные работы

Наша работа напрямую основана на трех фундаментальных методах PEFT: LoRA [5], OLoRA [3] и LoRA+ [4].

2.1. Low-Rank Adaptation (LoRA) LoRA основана на гипотезе, что изменение весов модели во время адаптации имеет низкий внутренний ранг. Он замораживает предварительно обученные веса модели и вводит пару обучаемых низкоранговых матриц A и B в каждый целевой слой, что показано на рисунке 1. Обновление представлено как ΔW = BA.

$$W = W_0 + \Delta W = W_0 + BA \qquad (1)$$

Чтобы обеспечить неразрушающее начало обучения, матрица B инициализируется нулями, делая начальное обновление нулевым. Хотя LoRA очень эффективна по параметрам, она может сходиться медленнее, чем полная тонкая настройка [5].

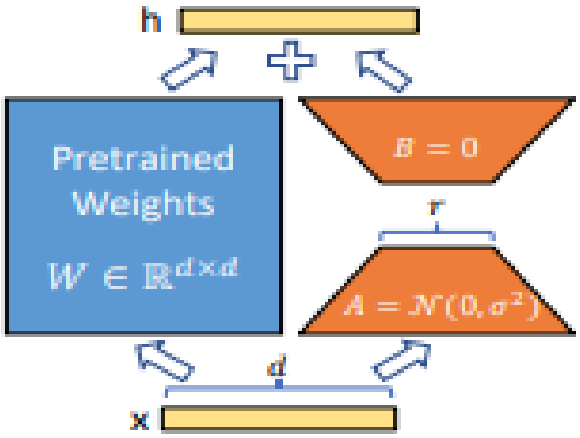

***Рисунок 1**: Перепараметризация весов для обучения только A и B*

2.2. Orthonormal LoRA (OLoRA) OLoRA устраняет ограничения скорости сходимости и стабильности LoRA путем улучшения процесса инициализации. Вместо случайной инициализации OLoRA выполняет QR-разложение предварительно обученной матрицы весов W для получения ортогональной матрицы Q и верхнетреугольной матрицы R [3]. Матрицы адаптера B и A затем инициализируются первыми r столбцами Q и r строками R [12] [13] соответственно, что показано на рисунке 2.

$$W_{adapted} = W + Q_r R_r \qquad (2)$$

Эта ортонормированная инициализация обеспечивает «хорошо обусловленный ландшафт оптимизации», что приводит к более

быстрой сходимости и часто лучшей конечной производительности по сравнению со стандартным LoRA.

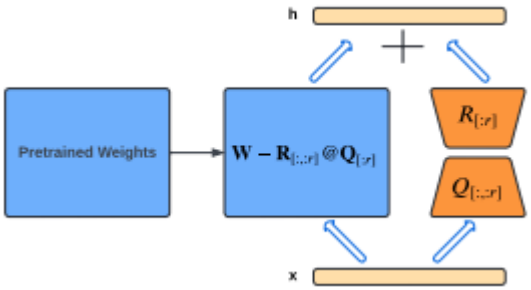

***Рисунок 2**: Иллюстрация метода OloRA*

2.3. LoRA+ LoRA+ выявляет субоптимальность в стандартной процедуре обучения LoRA, где обе матрицы адаптера A и B обновляются с одинаковой скоростью обучения. Авторы утверждают, что это замедляет изучение признаков. LoRA+ исправляет это, устанавливая гораздо более высокую скорость обучения для матрицы B, чем для матрицы A (т.е. ηB = λ * ηA, где λ >> 1). Это простое изменение, как показано, улучшает производительность на 1-2% и ускоряет обучение до 2 раз без дополнительных вычислительных затрат.

$$A \leftarrow A - \eta \times G_A ,$$

$$B \leftarrow B - \lambda\eta \times G_B \qquad (3)$$

$$\lambda \gg 1$$

3. Метод OLoRA+

Наша центральная гипотеза заключается в том, что преимущества превосходной инициализации OLoRA и ускоренной оптимизации LoRA+ являются взаимодополняющими и могут быть объединены для создания более мощного метода PEFT.

3.1. Архитектурная синергия OLoRA+ — это гибридный метод, который объединяет эти две техники в двухфазном концептуальном процессе: структурированная инициализация с последующей дифференциальной оптимизацией. Процесс начинается с **фазы инициализации OLoRA**. Для заданной предварительно обученной матрицы весов W мы сначала выполняем QR-разложение, W=QR. Матрицы адаптера, обозначенные как B (восходящая проекция) и A (нисходящая проекция), затем инициализируются с использованием результатов этого разложения. В частности, B инициализируется Q_r (первыми r столбцами ортогональной матрицы Q), а A инициализируется R_r (первыми r строками верхнетреугольной матрицы R). Это обеспечивает структурированную, ортонормированную отправную точку, которая является низкоранговым приближением исходных весов, создавая «хорошо обусловленный ландшафт оптимизации» [3]. Таким образом, адаптированная матрица весов выражается как $W_{adapted} = W + Q_r R_r$, где B и A начинаются как Q_r и R_r.

$$W = W_0 + \Delta W = W_0 + BA = W_0 + Q_r R_r \qquad (4)$$

После начала обучения применяется **фаза оптимизации LoRA+**. Вместо использования единой скорости обучения η для обеих матриц адаптера мы вводим коэффициент скорости обучения λ. Правила обновления матриц на каждом шаге градиентного спуска определяются следующими формулами:

$$A \leftarrow A - \eta \times G_A ,$$

$$A \leftarrow A - \eta \times G_A ,$$

Здесь G_A и G_B представляют градиенты для матриц A и B соответственно. Базовая скорость обучения — η, а эффективная скорость обучения для матрицы B масштабируется коэффициентом λ. В статье LoRA+ рекомендуется λ≫1 для стандартного LoRA для ускорения изучения признаков. Эта синергия — начало с превосходного начального состояния, а затем следование более эффективному пути обучения — является основой метода OLoRA+.

3.2. Детали реализации Мы реализовали OLoRA+ с использованием библиотеки Hugging Face PEFT.

1. **OLoRA Initialization:** Мы настроили LoraConfig, установив параметр

init_lora_weights="olora". Это указывает библиотеке выполнить QR-разложение и соответствующим образом инициализировать матрицы A и B.

2. **LoRA+ Optimization:** Мы создали пользовательский оптимизатор, разделив обучаемые параметры модели на две группы на основе их имен (A и B). Затем мы создали оптимизатор AdamW, назначив разные скорости обучения каждой группе на основе заданного λ. Этот пользовательский оптимизатор затем был передан в Hugging Face Trainer.

4. Экспериментальная установка

Для проверки нашего метода мы провели контролируемый эксперимент, сравнивая OLoRA+ со стандартным базовым уровнем OLoRA [3].

4.1. Модель

Все эксперименты проводились с использованием модели TinyLlama/TinyLlama-1.1B-Chat-v1.0, компактной, но мощной LLM, подходящей для эффективных экспериментов [10].

4.2. Набор данных

Мы использовали подмножество набора данных tatsu-lab/alpaca, который состоит из 52 000 демонстраций следования инструкциям, сгенерированных движком OpenAI text-davinci-003. Для наших экспериментов мы использовали разделение 1000 образцов для обучения и 500 образцов для оценки, чтобы имитировать сценарий тонкой настройки с ограниченными ресурсами.

4.3. Базовые показатели и гиперпараметры Для обеспечения справедливого сравнения все ключевые гиперпараметры оставались неизменными во всех экспериментах:

Rank (r): 32; **Alpha (α):** 16; **Learning Rate:** 3e-4; **Epochs:** 1; **Baseline (OLoRA):** Модель обучалась с использованием инициализации OLoRA со стандартным оптимизатором AdamW [9]; **Our Method (OLoRA+):** Модель обучалась с использованием инициализации OLoRA и нашего пользовательского оптимизатора LoRA+ с изменяющимся коэффициентом скорости обучения.

4.4. Метрики оценки Производительность модели оценивалась с использованием набора стандартных метрик:

- **Evaluation Loss:** Для измерения способности модели обобщать на невидимые данные во время обучения [11].
- **BLEU:** Для измерения точности и беглости сгенерированного текста [8].
- **ROUGE:** В частности, ROUGE-1, ROUGE-2 и ROUGE-L, для измерения полноты ключевой информации в сгенерированном тексте [7].

5. Результаты и обсуждение

Наши эксперименты подтвердили нашу первоначальную гипотезу и привели к значительному новому открытию относительно динамики обучения инициализированных адаптеров.

5.1. Улучшение производительности

По всем количественным метрикам наш метод OLoRA+ стабильно превосходил базовый уровень OLoRA. Модель, обученная с OLoRA+, достигла более низкого конечного значения evaluation loss и более высоких показателей BLEU и ROUGE, что указывает на превосходную обобщаемость и качество генерации текста, что показано в таблице 1.

Таблица 1

Результаты оценки

Method	Eval Loss	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
OLoRA (Baseline)	88.49	74.53	56.46	71.78	56.81
OLoRA+ ($\lambda << 1$)	**88.22**	**74.56**	56.41	**71.86**	**56.83**
OLoRA+ ($\lambda >> 1$)	88.39	74.52	**56.49**	**71.86**	**56.83**

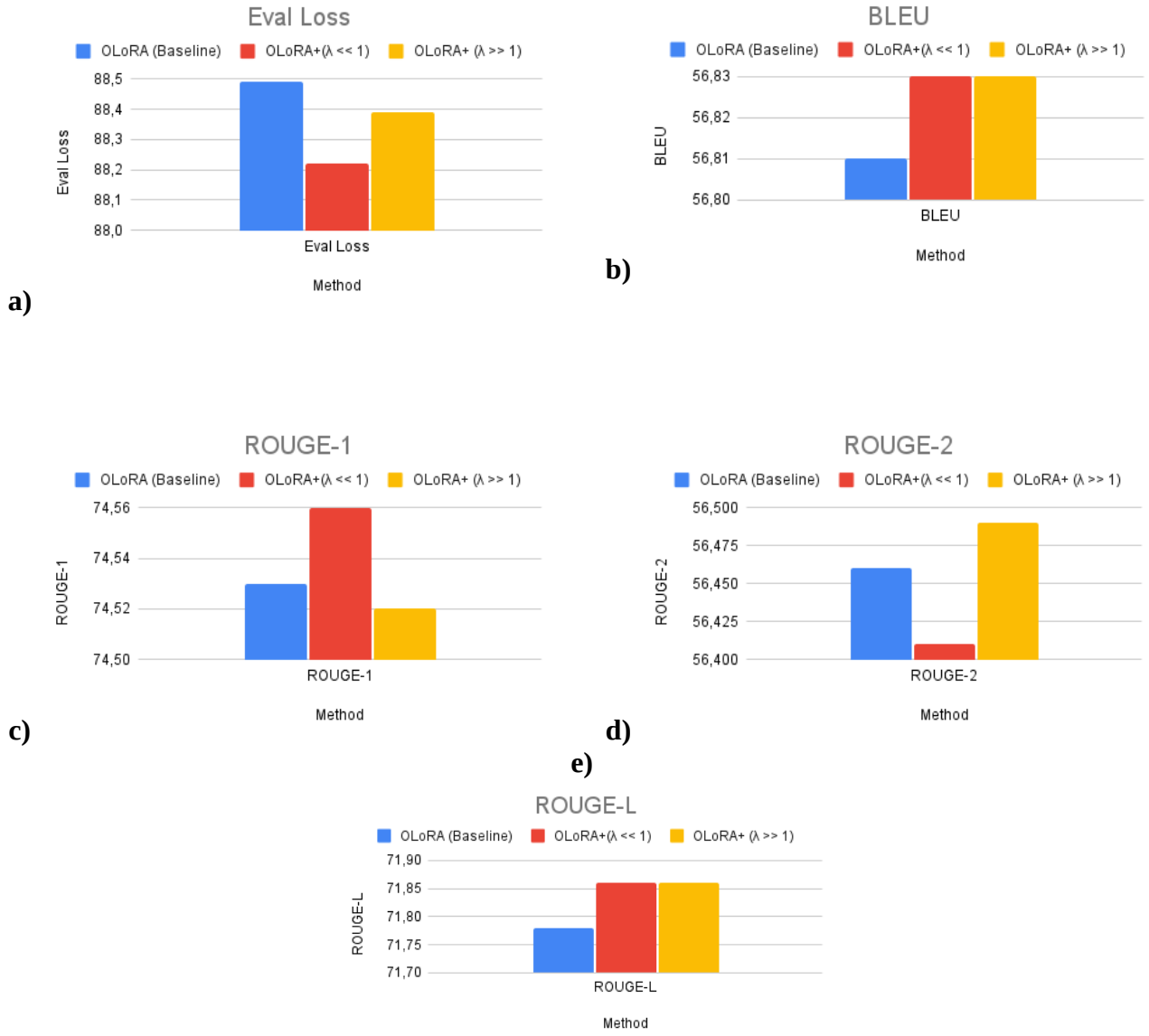

Рисунок 3: Сравнительный анализ производительности методов по ключевым метрикам оценки. a) Оценочные потери (Evaluation Loss), b) Показатель BLEU, c) Показатель ROUGE-1, d) Показатель ROUGE-2, e) Показатель ROUGE-L.

Как показано на Рисунке 3, предложенный нами метод OLoRA+ демонстрирует стабильное преимущество по всем метрикам. На графике **(a)** видно, что обе версии OLoRA+ достигают более низких оценочных потерь по сравнению с базовым OLoRA, при этом стратегия «Refinement» ($\lambda < 1$) показывает наилучший результат, что свидетельствует о превосходной способности к обобщению.

Графики **(b), (c), (d) и (e)** иллюстрируют производительность по метрикам генерации текста BLEU и ROUGE. Во всех случаях оба режима OLoRA+ превосходят или соответствуют результатам базовой модели. В частности, по метрикам ROUGE-L и BLEU обе стратегии OLoRA+ показывают идентичные и более высокие результаты, чем OLoRA. Эти визуальные данные подтверждают нашу гипотезу о том, что гибридный подход, сочетающий структурированную инициализацию с дифференциальной оптимизацией, является более эффективной стратегией для адаптации больших языковых моделей.

5.2. Открытие двойных режимов обучения Наиболее значительным результатом нашего исследования является то, что, вопреки рекомендациям в оригинальной статье LoRA+, OLoRA+ достигает высокой производительности при коэффициенте скорости обучения (λ) как большем, так и меньшем 1. Это открытие предполагает, что оптимальная стратегия оптимизации зависит от метода инициализации. Мы характеризуем эти два эффективных режима как компромисс между «Refinement» и «Exploration».

5.3. Характеристика «Refinement» против «Exploration»

- **The Refinement Strategy (λ < 1):** Когда скорость обучения для матрицы A (полученной из Rr) выше, чем для матрицы B (полученной из Qr), модель принимает консервативную стратегию. Она доверяет высококачественным ортонормированным направлениям, предоставленным инициализацией OLoRA, и фокусирует усилия обучения на поиске правильных величин и комбинаций для этих направлений. Этот подход мягко уточняет сильную начальную структуру. Наши результаты показывают, что это очень эффективная стратегия, предполагающая, что для многих задач основная проблема заключается в обучении *как* использовать начальный базис, а не в поиске нового.
- **The Exploration Strategy (λ > 1):** Когда скорость обучения для матрицы B выше, чем для матрицы A, модель принимает более агрессивную стратегию. Она использует инициализацию OLoRA в качестве отправной точки, но активно исследует новые направления в пространстве параметров, более быстро обновляя матрицу B. Этот подход более охотно отклоняется от исходной ортонормированной структуры в поисках потенциально лучшего решения. Это соответствует исходной логике LoRA+ и эффективно для задач, которые могут потребовать адаптации за пределами исходного подпространства.

6. Заключение

В данной статье мы представили OLoRA+, новый гибридный метод PEFT, который синергетически сочетает ортонормированную инициализацию OLoRA с дифференциальной оптимизацией LoRA+. Наши эмпирические результаты демонстрируют, что OLoRA+ стабильно превосходит базовый уровень OLoRA в задачах следования инструкциям, достигая более низкого значения evaluation loss и более высоких показателей BLEU и ROUGE без каких-либо дополнительных вычислительных затрат.

Ключевым вкладом этой работы является открытие и характеристика двух различных и эффективных режимов обучения для OLoRA+: стратегии «Refinement» (коэффициент < 1) и стратегии «Exploration» (коэффициент > 1). Это открытие показывает, что оптимальная стратегия оптимизации фундаментально зависит от схемы инициализации, предлагая новое измерение для настройки гиперпараметров.

Это исследование не только представляет OLoRA+ как более универсальный и мощный подход для адаптации LLM, но и обеспечивает более глубокое понимание критического взаимодействия между инициализацией адаптера и динамикой оптимизации, открывая путь для более эффективной тонкой настройки в условиях ограниченных ресурсов. В будущем мы планируем протестировать другие модели, такие как Gemma-2B, OPT-1.3B, и увеличить размер набора данных до 10 000.

Список литературы

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K.,

Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P. (2021). On the Opportunities and Risks of Foundation Models. arXiv. https://doi.org/10.48550/arXiv.2108.07258.

2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional trans- formers for language understanding. CoRR, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.
3. Büyükakyüz, K. (2024). *OLORA: Orthonormal Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2406.01775
4. Hayou, S., Ghosh, N., & Yu, B. (2024). *LoRA+: Efficient Low Rank Adaptation of Large Models*. arXiv preprint arXiv:2402.12354.
5. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685.
6. Taori, R., et al. (2023). *Stanford Alpaca: An Instruction-following LLaMA Model*. Stanford Center for Research on Foundation Models (CRFM).
7. Wikipedia contributors. (n.d.). ROUGE (metric). In Wikipedia. Retrieved October 7, 2025, from https://en.wikipedia.org/wiki/ROUGE_(metric)
8. Wikipedia contributors. (n.d.). BLEU. In Wikipedia. Retrieved October 7, 2025, from https://en.wikipedia.org/wiki/BLEU
9. Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. arXiv. https://doi.org/10.48550/arXiv.1711.05101
10. Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An Open-Source Small Language Model. arXiv. https://doi.org/10.48550/arXiv.2401.02385
11. Wikipedia contributors. (n.d.). Evaluation function. In Wikipedia. Retrieved October 7, 2025, from https://en.wikipedia.org/wiki/Evaluation_function
12. Aghajanyan, A., Zettlemoyer, L., & Gupta, S. (2020). Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. arXiv. https://doi.org/10.48550/arXiv.2012.13255
13. Meng, F., Wang, Z., & Zhang, M. (2024). PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2404.02948