# Chatbot Intelligent ENIAD

## Résumé

Ce rapport discute en profondeur des travaux antérieurs concernant l'implémentation ou le développement d'un chatbot conversationnel destiné aux élèves de l'ENIAD (École Nationale d'Intelligence Artificielle et du Digital). Pour relever les problèmes de scolarité liés à la communication et à l'information, nous proposons un système technologique interfaced avec un algorithme efficace pour le traitement du langage naturel et le machine learning. Cette étude s'attaque à l'ensemble des approches de développement et les classifie comme étant de modélisation des règles : RAG, Fine-tuning, ainsi que les approches d’intelligence artificielle possiblement utilisables comme MBFT ou QLoRA et enfin, les pragmatiques, sur l'ENIAD. Le rapport compile en outre une proposition de l'architecture technique et étape du déploiement.

## Introduction au Projet

### Contexte et Vision

Le secteur de l'enseignement supérieur est en pleine évolution, progrès oblige, il faut apporter de nouvelles manières de gérer la communication et la dissémination de l'information. À ce titre, l’École Nationale de L'Intelligence Artificielle Du Digital (ENIAD) se fixe comme objectif de se placer en tête de la liste sur les questions de technologie au service du suivi des étudiants. Le chatbot ENIAD se présente tel un assistant virtuel sur mesure, disponible à tout moment, 24 heures sur 24, 7 jours sur 7, et capable de répondre aux besoins variés des étudiants dans le contexte de la institution et cet environnement académique particulier.

### Définition de la Problématique

Les établissements d'enseignement supérieur comme l'ENIAD font face à un certain nombre de problèmes liés à la communication :

* Surcharge informationnelle - Les étudiants sont confrontés à un volume important d'informations dispersées sur différentes plateformes.
* Accessibilité limitée - Les services administratifs ont des horaires restreints et sont souvent surchargés.
* Personnalisation insuffisante - Les systèmes actuels offrent rarement des réponses adaptées aux besoins individuels.
* Inefficacité des canaux de communication - Les emails et les annonces générales ne garantissent pas une diffusion optimale de l'information.

|  |  |
| --- | --- |
| État Actuel | Solution Chatbot |
| Informations fragmentées | Point d’accès unique |
| Délais de réponse longs | Disponibilité 24/7 |
| Dépendance aux horaires | Réponses instantanées |
| Difficulté d’accès | Personnalisation avancée |
| Service non personnalisé | Évolution continue |

TABLE 1 – Transformation de l’accès à l’information grâce au chatbot ENIAD

**1.3 Objectifs Stratégiques**

Notre chatbot vise à révolutionner l’expérience étudiante à l’ENIAD par :

1. Centralisation de l’information académique

* Fournir des documents ou des informations à la demande (emplois du temps, annonces, règlements)
* Créer un point d’accès unifié aux ressources pédagogiques
* Consolider les informations provenant de différentes sources institutionnelles

2. Automatisation des interactions administratives

* Répondre automatiquement aux questions fréquentes des étudiants
* Guider les utilisateurs dans leurs démarches administratives
* Réduire la charge de travail du personnel administratif

3. Personnalisation de l’expérience utilisateur

* Adapter les réponses au profil académique de chaque étudiant
* Envoyer des notifications ciblées aux utilisateurs sur notre plateforme
* Proposer des recommandations pertinentes selon le parcours de l’étudiant

4. Enrichissement de la vie étudiante

* Informer sur les activités, clubs et opportunités
* Alerter sur les événements, conférences et deadlines importantes
* Faciliter l’intégration des nouveaux étudiants

5. Innovation technologique

Intégrer la synthèse et reconnaissance vocale via des solutions open-source

Mettre en œuvre des technologies d’IA de pointe accessibles

Créer un système évolutif capable d’apprendre continuellement

**1.4 Contraintes et Défis**

Le développement de notre solution fait face à plusieurs obstacles :

|  |  |
| --- | --- |
| Catégorie | Défis identifiés |
| Données | Manque de données fiables et instabilité des services (modules, rôles des professeurs) |
| Ressources | Absence de financement dédié et ressources matérielles limitées pour l’hébergement et le traitement |
| Complexité | Nécessité de couvrir un large éventail de questions académiques et administratives |
| Multilinguisme | Besoin de gérer efficacement le français, l’arabe et potentiellement l’anglais [16] |
| Évolutivité | Obligation d’adaptation aux changements fréquents des programmes et règlements [18] |
| Sécurité | Protection des données personnelles des étudiants conformément aux réglementations [17] |

Table 2 – Défis et contraintes du projet ENIAD

Les défis techniques et organisationnels identifiés nécessitent une approche méthodique et progressive, avec une phase pilote avant un déploiement complet. Le manque de données représente l’obstacle le plus important à surmonter

**2 État de l’Art des Technologies de Chatbot**

**2.1 Évolution des Systèmes Conversationnels**

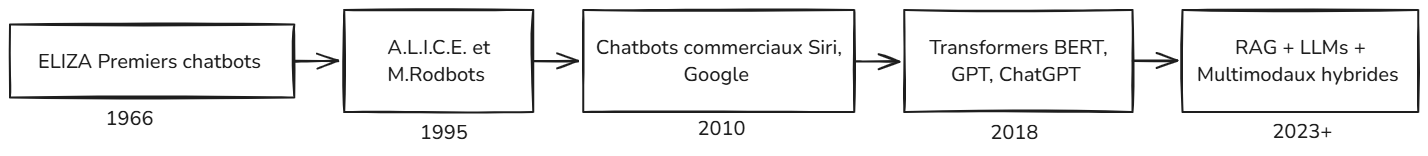
L’histoire des chatbots a connu une évolution spectaculaire depuis les premiers systèmes comme ELIZA [6] dans les années 1960 jusqu’aux assistants intelligents actuels basés sur des architectures Transformers [8] avancées

Figure 2 – Évolution historique des technologies de chatbot

Ce schéma illustre l'évolution historique des systèmes conversationnels artificiels. ELIZA (1966) représente la genèse des chatbots avec ses règles de pattern-matching simples. A.L.I.C.E. et les Rodbots (1995) ont introduit le traitement du langage naturel plus sophistiqué et les bases de connaissances structurées. L'ère commerciale (2010) marque l'intégration des assistants virtuels dans les technologies grand public. La révolution des Transformers (2018) constitue une rupture paradigmatique avec l'avènement des architectures d'attention et des modèles pré-entraînés à grande échelle. La période actuelle (2023+) se caractérise par l'hybridation des approches, combinant Retrieval-Augmented Generation (RAG), Large Language Models (LLMs) et capacités multimodales pour des systèmes plus robustes et contextuellement informés.

2.2 Taxonomie des Approches de Chatbots

Les chatbots modernes peuvent être classifiés selon différentes dimensions technologiques et fonctionnelles :

2.2.1 Classification par Architecture

|  |  |  |  |
| --- | --- | --- | --- |
| Méthode | Description | Description | Description |
| Rule-based | Systèmes basés sur des règles définies manuellement avec pattern matching | — Simple à implémenter — Prédictible et contrôlable — Faible coût de calcul | — Peu flexible — Limité à des scénarios connus — Difficulté à gérer la complexité |
| Retrievalbased | Sélectionne la réponse la plus pertinente dans une base de données prédéfinie | — Fiabilité des réponses — Contrôle sur les sorties — Relative simplicité | — Manque de personnalisation — Difficulté avec les requêtes complexes — Base de données limitée |
| Generative | Génère des réponses originales à partir de modèles statistiques ou neuronaux | — Flexibilité maximale — Capacité à gérer l’imprévu — Réponses plus naturelles | — Risque d’hallucinations — Contrôle limité — Coût computational élevé |
| Hybride (RAG) | Combine la recherche d’information et la génération de texte | — Précision factuelle — Flexibilité contextuelle — Réponses fondées sur des sources | — Complexité d’implémentation — Latence potentielle — Dépendance à la qualité des documents |

Table 2 – Comparaison des architectures de chatbot

2.2.2 Classification par Technique d’Apprentissage

Ce diagramme taxonomique présente les principales approches d'intelligence artificielle employées dans le développement des chatbots contemporains. L'apprentissage supervisé englobe les techniques de fine-tuning complet et les méthodes PEFT (Parameter-Efficient Fine-Tuning), optimisant des modèles préexistants sur des données étiquetées. L'apprentissage non-supervisé exploite la clusterisation et les auto-encodeurs pour identifier des patterns sans supervision humaine. L'apprentissage par renforcement se manifeste via RLHF (Reinforcement Learning from Human Feedback) et DPO (Direct Preference Optimization), permettant d'aligner les modèles sur les préférences humaines. Les méthodes hybrides combinent ces approches: RAG (Retrieval-Augmented Generation) intègre des connaissances externes aux capacités génératives, tandis que Chain-of-Thought améliore le raisonnement séquentiel. Cette classification reflète l'évolution des chatbots vers des systèmes multi-paradigmes, fusionnant diverses techniques pour optimiser la performance conversationnelle.

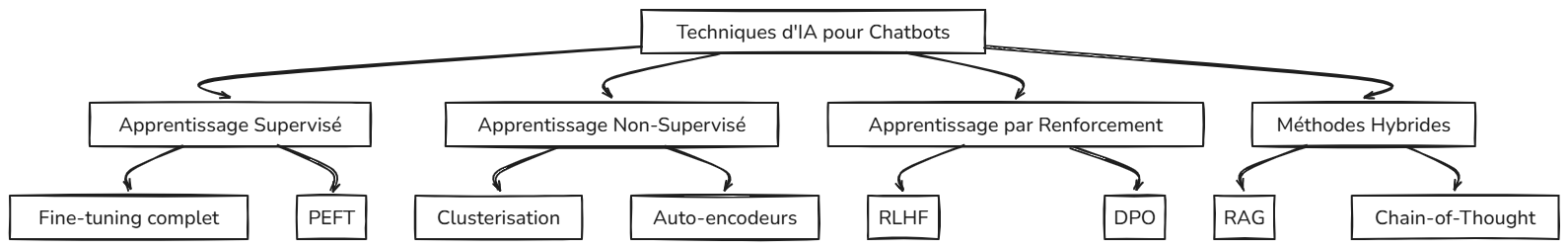


Figure 3 – Classification des techniques d’apprentissage applicables aux chatbots

2.3 Méthodes de Fine-tuning pour les LLMs

Les modèles de langage préentraînés (LLM s) constituent aujourd’hui le fondement des chatbots les plus avancés. Leur adaptation à des domaines spécifiques comme celui de l’ENIAD peut être réalisée par différentes méthodes :

2.3.1 Paradigme MBFT (Model Balancing Helps Low-data Training and Finetuning)

Le MBFT représente une avancée significative dans les techniques d’adaptation des modèles de langage pour des applications spécifiques. Cette approche, présentée par Chen et al. (2024), introduit une stratégie qui équilibre l’apprentissage entre les connaissances générales du modèle et les besoins spécifiques du domaine cible, particulièrement efficace dans des scénarios à faible volume de données.

Le MBFT organise l’adaptation des paramètres en tenant compte de plusieurs aspects :

1. Équilibrage des connaissances : Maintient les capacités générales du modèle tout en apprenant des données spécifiques.

2. Optimisation pour faibles données : Utilise des techniques avancées pour maximiser l’efficacité de l’entraînement avec des ensembles de données limités.

3. Adaptation flexible : Permet une spécialisation fine pour des cas d’usage précis tout en évitant la sur-optimisation. Cette approche permet une adaptation robuste même avec des ressources de données limitées, ce qui est idéal pour le contexte de l’ENIAD. Avantages du MBFT pour notre projet :

— Efficacité dans l’utilisation de données limitées propres à l’ENIAD

— Prévention du sur-ajustement (overfitting) grâce à un équilibrage intelligent

— Maintien des capacités générales du modèle pour des réponses polyvalentes

— Réduction des problèmes de catastrophic forgetting

Selon les travaux récents de Chen et al. (2024), le MBFT surpasse les méthodes traditionnelles de fine-tuning dans des contextes à faible volume de données, avec des améliorations significatives pour les tâches nécessitant une expertise contextuelle.

2.3.2 Techniques PEFT (Parameter-Efficient Fine-Tuning)

Face aux limitations en ressources de calcul, les méthodes PEFT offrent des alternatives prometteuses

|  |  |  |  |
| --- | --- | --- | --- |
| Méthode | Paramètres modifiés | Mémoire requise | Performance relative |
| Full Fine-tuning  LoRA (Low-Rank Adaptation) [5]  Q-Lora  Prompt-tuning | 100%  0.1-1%  0.1-1%  <0.1% | Très élevée  Basse  Très basse  Minimale | 100%  95-99%  90-97%  85-95% |

Table 3 – Comparaison des méthodes de fine-tuning en termes d’efficacité

Comparaison pour le projet de chatbot ENIAD :

Fine-tuning : Offre une personnalisation élevée et des performances optimales avec QLoRA/MBFT, bien adapté au multilinguisme, mais dépend fortement de données structurées et est techniquement complexe. Idéal pour des réponses contextuelles, mais limité par un manque de données.

RAG : Assure une précision factuelle avec une base de connaissances facilement mise à jour, mais dépend de la qualité des documents et peut introduire de la latence. Parfait pour des questions factuelles, mais nécessite un prétraitement rigoureux des données.

Prompting : Simple, rapide, et peu coûteux, permet une mise en œuvre immédiate avec des modèles multilingues. Moins précis, limité pour les cas complexes ou personnalisés. Idéal pour un prototype rapide, mais insuffisant pour une personnalisation avancée à l’ENIAD.

2.3.3 Q-Lora : Une Solution Optimale pour nos Contraintes

Pour le projet de chatbot ENIAD, la méthode Q-Lora (Quantized Low-Rank Adaptation) apparaît comme particulièrement adaptée :

La méthode Q-Lora, introduite par Dettmers et al. (2023), combine la quantification à 4 bits du modèle de base avec l’adaptation de rang faible, permettant ainsi de fine-tuner des LLM s sur des ressources GPU modestes tout en maintenant des performances proches du fine-tuning complet.

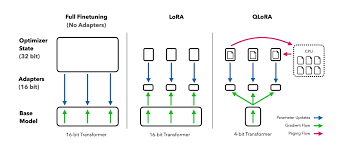
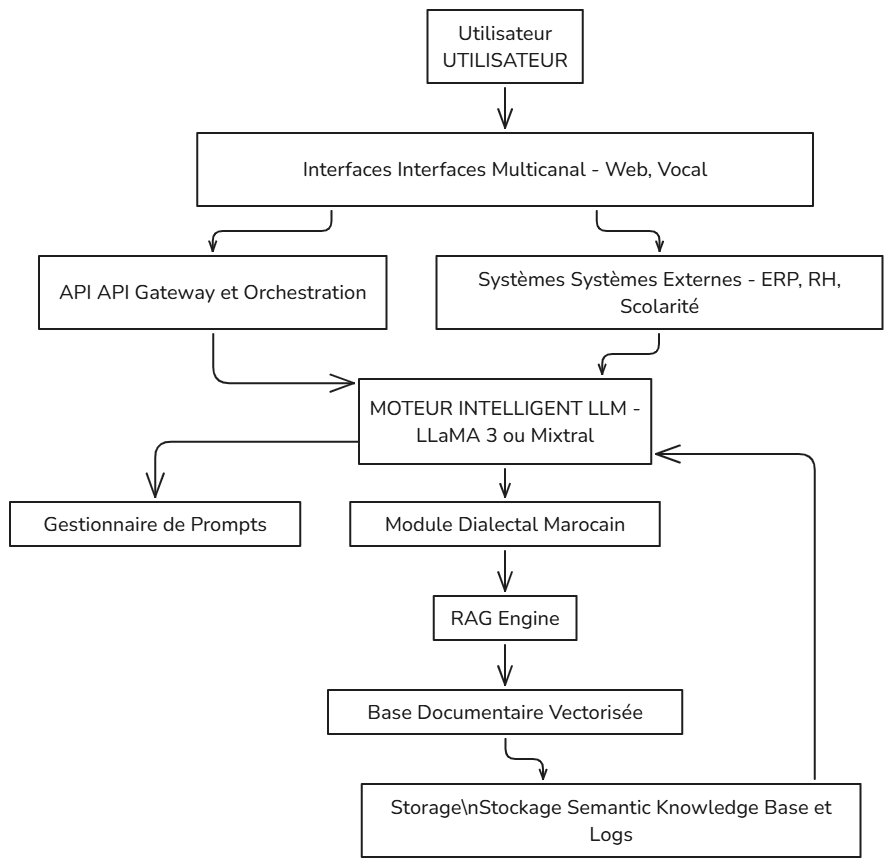


Figure 4 – Architecture Q-Lora

Mise en œuvre pratique de Q-Lora pour ENIAD :

1. Sélectionner un LLM de base adapté au français/arabe (ex : BLOOMZ [14], Mistral [15], XLM-R [16])



Ce diagramme représente l'architecture fonctionnelle d'un système conversationnel hybride combinant recherche d'information et génération de texte. Le processus s'initie par la question utilisateur (1) qui déclenche deux flux parallèles: un traitement direct vers le modèle génératif (ligne pointillée) et une interrogation du système de recherche (2). Ce dernier extrait les informations pertinentes de la base de connaissances ENIAD (3), constituant une source documentaire structurée. Le modèle génératif (4) fusionne ensuite les entrées provenant de la question originale, des résultats de recherche et des connaissances contextuelles pour produire une réponse cohérente. Cette architecture illustre le paradigme RAG (Retrieval-Augmented Generation), permettant d'ancrer les capacités génératives dans des sources factuelles vérifiables tout en préservant la fluidité conversationnelle, aboutissant à une réponse finale (5) contextuellement informée et factuelle.

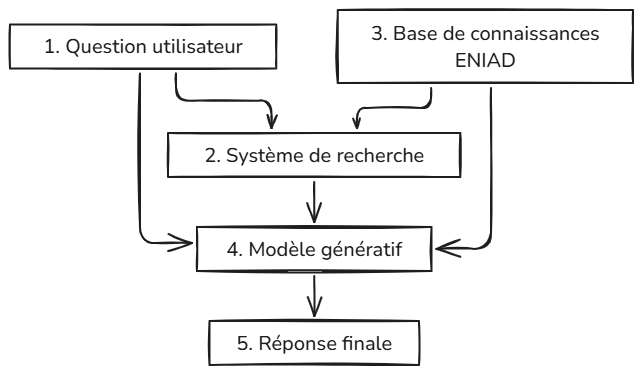


Figure 5 – Architecture RAG pour le chatbot ENIAD

2. Quantifier le modèle à 4 bits pour réduire l’empreinte mémoire

3. Définir les matrices de rang faible (r=8 ou r=16) pour l’adaptation

4. Préparer un jeu de données d’entraînement spécifique à l’ENIAD 5. Effectuer le fine-tuning sur un GPU à mémoire modeste (>= 8 Go)

3 Retrieval Augmented Generation (RAG [1])

Pour garantir la précision factuelle et contextuelle de notre chatbot, l’intégration d’une architecture RAG constitue une approche privilégiée.

Le RAG combine la puissance des modèles génératifs avec des mécanismes de recherche d’information pour produire des réponses précises et contextuelles.

3.1 Construction de la Base de Connaissances

La création d’une base de connaissances robuste pour le chatbot ENIAD nécessite les étapes suivantes :

1. Collecte des documents : Rassembler les règlements, guides, FAQ, emplois du temps, et autres ressources institutionnelles.

2. Prétraitement : Nettoyage des données, déduplication, et structuration en formats exploitables.

3. Chunking : Segmentation des documents en unités sémantiquement cohérentes pour une recherche efficace.

4. Vectorisation : Création d’embeddings sémantiques pour chaque segment à l’aide de modèles comme SentenceBERT [11] ou GTE-Large [12].

5. Indexation : Organisation des embeddings dans une base de recherche vectorielle (FAISS [13], Pinecone) pour une récupération rapide. Sources de données potentielles pour l’ENIAD : — Documents académiques : Règlements intérieurs, syllabus, guides d’études. — FAQ existantes : Questions fréquemment posées par les étudiants. — Ressources administratives : Formulaires, calendriers académiques, procédures. — Données temporelles : Emplois du temps, dates d’examens, événements. — Contenu dynamique : Annonces, actualités, mises à jour institutionnelles.

4 Architecture Proposée pour le Chatbot ENIAD

Sur la base de l’état de l’art, nous proposons une architecture hybride combinant RAG et Fine-tuning (via Q-Lora et MBFT) pour maximiser les performances tout en respectant les contraintes de ressources

L’architecture est structurée en plusieurs modules interconnectés :

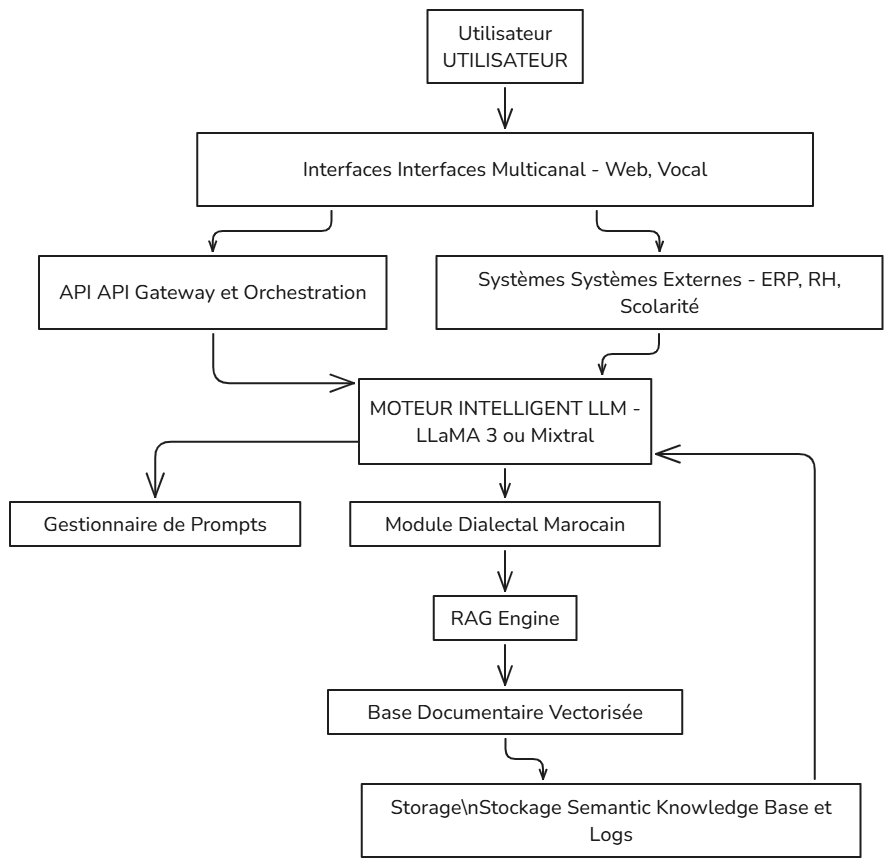


Figure 6 – Architecture fonctionnelle proposée pour le chatbot ENIAD

5 Critique et Limites

Malgré les avancées technologiques, plusieurs défis persistent :

— Fusion MBFT et Q-Lora : La combinaison de ces approches augmente la complexité d’implémentation et nécessite une expertise technique approfondie.

— Données non brutes : Les données extraites (FAQ, documents) ne sont pas toujours structurées, ce qui peut affecter la fiabilité des réponses.

— Conflits d’information : Des informations similaires mais contradictoires dans la base de connaissances peuvent entraîner des erreurs.

— Catastrophic forgetting : Lors du Fine-tuning, le modèle risque de perdre des connaissances générales, limitant sa robustesse.

— Hallucinations : Les modèles génératifs peuvent produire des réponses incorrectes ou non fondées, particulièrement en l’absence de données fiables.

— Sécurité des données : Les interactions avec les étudiants impliquent des données sensibles, nécessitant des mécanismes robustes de protection [17]. Le risque d’hallucinations nécessite une validation humaine des réponses critiques (ex : informations administratives) et une base de connaissances bien curated.

Conclusion

Le développement d’un chatbot intelligent pour l’ENIAD représente une opportunité majeure de moderniser la communication et l’accès à l’information pour les étudiants. En combinant RAG pour la précision factuelle et Fine-tuning (via Q-Lora et MBFT) pour la personnalisation, notre solution répond aux besoins spécifiques de l’ENIAD tout en surmontant les contraintes de ressources. Une approche progressive, avec un prototype initial et des phases de test rigoureuses, garantira un déploiement réussi et évolutif, avec une attention particulière portée au multilinguisme [16] et à la sécurité des données [17].

Références

[1] Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In Proceedings of Neural Information Processing Systems (NeurIPS). https:// arxiv.org/abs/2005.11401

[2] Dettmers, T., et al. (2023). "QLoRA : Efficient Finetuning of Quantized LLMs". In Proceedings of Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/ 2305.14314

[3] Mangrulkar, S., et al. (2023). "PEFT : Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware". In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). <https://arxiv.org/abs/2203.15556>

[4] Chen, Z., et al. (2024). "Model Balancing Helps Low-data Training and Fine-tuning". arXiv preprint. <https://arxiv.org/abs/2410.12178>

[5] Hu, E., et al. (2021). "LoRA : Low-Rank Adaptation of Large Language Models". In Proceedings of the International Conference on Learning Representations (ICLR). https: //arxiv.org/abs/2106.09685

[6] Weizenbaum, J. (1966). "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine". Communications of the ACM, 9(1), 36-45. <https://dl.acm.org/doi/10.1145/365153.365168>

[7] Wallace, R. (1995). "A.L.I.C.E. and AIML : A Brief History". In Artificial Intelligence Markup Language (AIML) Specification. https://www.alicebot.org/articles/wallace/ alice.html

[8] Vaswani, A., et al. (2017). "Attention Is All You Need". In Proceedings of Neural Information Processing Systems (NeurIPS). <https://arxiv.org/abs/1706.03762>

[9] Devlin, J., et al. (2019). "BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding". In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). https://arxiv.org/abs/ 1810.04805

[10] Brown, T., et al. (2020). "Language Models are Few-Shot Learners". In Proceedings of Neural Information Processing Systems (NeurIPS). <https://arxiv.org/abs/2005.14165>

[11] Reimers, N., Gurevych, I. (2019). "Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks". In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). <https://arxiv.org/abs/1908.10084>

[12] Wang, Y., et al. (2023). "GTE : General Text Embeddings for Large Language Models". arXiv preprint. https://arxiv.org/abs/2308.03281 [13] Johnson, J., et al. (2019). "Billion-scale similarity search with GPUs". IEEE Transactions on Big Data. https://arxiv.org/abs/1702.08734 [14] Le Scao, T., et al. (2022). "BLOOM : A 176B-Parameter Open-Access Multilingual Language Model". arXiv preprint. <https://arxiv.org/abs/2211.05100>

[15] Jiang, A., et al. (2023). "Mistral 7B". arXiv preprint. [https://arxiv.org/abs/2310. 06825](https://arxiv.org/abs/2310.%2006825)

[16] Conneau, A., et al. (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). <https://arxiv.org/abs/1911.02116>

[17] Li, X., et al. (2023). "Privacy-Preserving Conversational AI : A Survey". In Proceedings of the IEEE Symposium on Security and Privacy. <https://arxiv.org/abs/2302.07519>

[18] Hoffmann, J., et al. (2022). "Training Compute-Optimal Large Language Models". In Proceedings of Neural Information Processing Systems (NeurIPS). https://arxiv.org/ abs/2203.15556