VOL. 13 NO. 9,
AUGUST, 2026
E-ISSN 3026-958X
P-ISSN 3027-1169

African Scholar Publications & Research International
www.africanscholarpub.com

Journal of Science Innovation & Technology Research (JSITR)

The Grounding Tax: How Domain Fine-Tuning Suppresses Context Utilisation in Retrieval-Augmented Generation

Labaran Usman Yunusa[1]; Rufai Yauri Aliyu[1]; Isah Charles Saidu[1]; Rislana A. Kanya[2]; & Ahmad N. Tambaya[3]

[1]Department of Computer Science, Baze University, Abuja, Nigeria. [2]Cosmopolitan University, Abuja, Nigeria. [3]Cyber Safety Alliance (CSA), Nottingham, United Kingdom.
Corresponding Author: usman8194@bazeuniversity.edu.ng
DOI: https://doi.org/10.70382/ajsitr.v13i9.096

Abstract

Large language models (LLMs) are routinely adapted to specialised domains through fine-tuning and grounded in external evidence through retrieval-augmented generation (RAG). The prevailing assumption is that these methods are complementary even though minimal empirical work examines whether domain-specific fine-tuning influences a model's responsiveness to retrieved context. Thus, this study examines the interaction between fine-tuning and RAG through a controlled evaluation in a high-stakes public-finance domain. The evaluation is conducted through 12 model configurations built on two open-weight model families (LLaMA-3.1-8B and Qwen2.5-7B), using a frozen 100-query benchmark covering factual question-answering (factual-QA), risk assessment, and summarisation under both closed-book and context-matched conditions. The evidence confirms successful knowledge injection as domain fine-tuning raised closed-book numeric-fact accuracy (NFA) from 0.148 to 0.248. However, when retrieved context is introduced, the benefit collapses as the base model gained +0.183 from context while the fine-tuned model gained only +0.021. Across the backbone pairs, the change in context gain ranged from -0.161 to -0.348 and was statistically significant for the LLaMA Instruct and Qwen Base pairs. Further ablation analysis shows that prompt-template and instruction alignment can amplify this effect. We term this phenomenon the grounding tax: domain fine-tuning imposes a measurable cost on the model's ability to use retrieved evidence because the adapted model relies on internally encoded knowledge rather than faithfully using the evidence supplied to it. Thus, these findings show that conventional closed-book evaluation can mask degradation in context utilisation and that fine-tuned RAG systems must be evaluated jointly under both closed-book and context-

conditioned settings. Therefore, we argue for treating domain adaptation, instruction alignment, and retrieval grounding as distinct and potentially competing design dimensions.

Keywords: Large Language Models; Retrieval-Augmented Generation; Fine-Tuning; Retrieval Grounding; Parametric Knowledge; Domain Adaptation.

Introduction

Retrieval-augmented generation has become the standard way of grounding LLM outputs in external evidence. RAG allows a model to answer from documents that did not form part of its training corpus without the need for retraining by simply conditioning generation on text fetched from a curated corpus (Lewis et al., 2020; Karpukhin et al., 2020). This approach is also a safety guardrail that reduces hallucination due to retrieval of relevant evidence during inference (Shuster et al., 2021). These retrieved texts can equally serve as a provenance trail that links outputs to their source in applied deployments. Another dominant route to conduct domain adaptation is through fine-tuning as parameter-efficient fine-tuning (PEFT) methods like low-rank adaptation (LoRA) and quantised LoRA (QLoRA) have made it practical to inject domain knowledge into open-weight models on modest hardware (Hu et al., 2021; Dettmers et al., 2023). Additionally, instruction-style fine-tuning on domain corpora is now routine across specialised fields with finance offering a well-developed example of adapted models running from early domain-specific encoders to open instruction-tuned financial LLMs (Araci, 2019; Xie et al., 2023; Liu et al., 2023). The motivation for this is simple as the model becomes a better native speaker of its target domain after training on domain text and formatted tasks.

In applied settings, domain fine-tuning and RAG frequently combined without rigorous evaluation. This "fine-tune-then-retrieve" architecture common within institutional knowledge systems implicitly assumes that parametric fluency and external evidence are purely additive (Zhang et al., 2023; Efeoglu and Paschke, 2024; Lee et al., 2024). However, the interaction between parametric knowledge and retrieved evidence remains poorly characterised. Thus, when conflict happens, a model often favours its parametric memory over provided evidence (Longpre et al., 2021). Furthermore, fine-tuning introduces systemic behavioural shifts that extend beyond the target task (Wu et al., 2023; Jain et al., 2023). Consequently, literature currently lacks a controlled measurement of how domain fine-tuning alters a model's capacity to utilise retrieved information while everything else is held constant. This question matters the most in high-stakes environments where answers must be traceable to provided sources. In sectors such as finance, law, and public administration, an answer is only acceptable if it can trace back to its source document with even a fluent response recalled from training data not able to substitute for a grounded one (Chen et al., 2024; Nie et al., 2024). Thus, a model that over relies on memorised domain knowledge may perform well in closed-book evaluations while failing to utilise the actual evidence provided during inference. This undermined the core objective of RAG architectures and if fine-tuning is responsible for this behaviour, the standard approach of combining fine-tuning with retrieval requires careful re-evaluation.

Consequently, this paper investigates this interaction empirically through a controlled evaluation in a public finance setting by utilising the public document corpus of a sovereign wealth fund (SWF). The frozen 100-query benchmark spanning factual-QA, risk assessment, and

summarisation was constructed from ~175 public documents collated. Subsequently, we evaluated 12 system configurations including a core four-way ablation (base model, base+retrieval, fine-tuned model, fine-tuned+retrieval), matched base-checkpoint controls, an Instruct-backbone ablation, and a cross-backbone replication on Qwen2.5-7B against the benchmark. All configurations answered identical prompts under closed-book and context-matched conditions with the latter supplying the exact same top-4 retrieved chunks to all relevant systems using deterministic decoding and a fixed seed throughout. Additionally, we introduced a context-gain measure $CG_M = Score(M_{context}) - Score(M_{closed})$ which captures the benefit a system derives from retrieved evidence whereas $\Delta CG = CG_{FT} - CG_{Base}$ captures how that benefit changes post-fine-tuning to help analyse the results.

The central finding is a grounding tax imposed by domain fine-tuning on retrieval responsiveness. Fine-tuning delivered as evidenced during closed-book tests where NFA on factual QA rose from 0.148 to 0.248 to show that institutional knowledge was injected into the model's parameters. However, this same fine-tuning collapsed the model's responsiveness to retrieved evidence as context gain on the same metric degraded from +0.183 to +0.021 to produce a negative ΔCG on NFA across all the three fine-tuned model families we trained, with the reduction statistically significant for the LLaMA Instruct and Qwen Base pairs. Thus, under the evaluated training recipe and experimental conditions, domain-specific fine-tuning degraded context gain as fine-tuned models ignored relevant supplied evidence to answer from parametric memory. However, the scope of this claim must be emphasised as it is only an empirical observation based on a single training recipe and institutional corpus rather than a universal law of fine-tuning. Consequently, this paper makes three contributions. Firstly, it empirically identifies and quantifies a parametric override effect in domain-adapted LLMs where fine-tuning improves closed-book institutional recall while reducing responsiveness to retrieved context. Secondly, it introduces context-gain analysis as an analytical device for explicitly measuring the change in retrieval benefit following domain adaptation. Thirdly, through a controlled 12-configuration benchmark spanning architectural ablations, it demonstrates the importance of evaluating fine-tuned LLMs under both closed-book and context-matched conditions.

RELATED WORKS

Retrieval-Augmented Generation

Pretrained language models store a large body of factual knowledge and can answer many questions without external support (Brown et al., 2020; Petroni et al., 2019). This parametric knowledge is fixed at training time, incomplete for specialised domains, and difficult to verify. Consequently, RAG addresses these limitations by conditioning the generator on documents fetched at inference time. This was facilitated by dense text retrieval which showed that learned embeddings outperformed sparse lexical matching for open-domain question answering (Karpukhin et al., 2020). Thus, Lewis et al. (2020) built the present standard RAG architecture by combining a dense retriever with a sequence-to-sequence generator, demonstrating accuracy gains over closed-book baselines on knowledge-intensive tasks.

Grounding offers three core benefits with the most important being that answers can show information not present or recent than the model's training corpus. Secondly, specific texts can be traced using provenance and lastly, the usage of retrieved text can reduce hallucination. The last benefit was further confirmed by Shuster et al. (2021) empirically through a finding that demonstrated retrieval augmentation reduces hallucination in knowledge-grounded dialogue even though it did not eliminate it. Further studies extended the pattern to structured knowledge sources and to pipelines where the

generator is adapted to the retrieval format (Efeoglu and Paschke, 2024; Pan et al., 2023). However, across this literature, the generator is typically a base pretrained or instruction-tuned model under the implicit assumption that a competent model will use whatever relevant context it receives. What it fails to address is how training interventions on the generator influence what retrieval delivers.

Domain Adaptation and Fine-Tuning of LLMs

General-purpose models are being adapted in a parallel line of research to follow instructions and aid specific domains. This instruction tuning conducted on demonstrations and usually followed by reinforcement learning from human feedback (RLHF) aligns model outputs with user intent (Ouyang et al., 2022; Ziegler et al., 2019). Furthermore, replicable recipes such as Alpaca showed that modest instruction datasets are enough to instil this behaviour in open models (Taori et al., 2023). This adaptation became routine outside large industrial or commercial settings due to the advent of openly released backbones like the LLaMA and Qwen model families (Bai et al., 2023; Grattafiori et al., 2024; Touvron et al., 2023). The cost of adapting these models was then further lowered by PEFT methods such as LoRA which limits updates to only low-rank matrices inserted into attention layers (Hu et al., 2021). Additionally, quantisation which reduces the memory footprint of frozen weights (Dettmers et al., 2022) was combined with LoRA to produce QLoRA which allows the fine-tuning of billion-parameter models on single GPUs (Dettmers et al., 2023). Further library support standardised these recipes in practice (Mangrulkar et al., 2022; Parthasarathy et al., 2024).

This trend proliferated especially with domain-specific adaptation spreading in science and medicine (Luo et al., 2022; Singhal et al., 2022; Taylor et al., 2022), and most importantly in finance. We saw the adaptation of a pretrained encoder to financial sentiment through FinBERT (Araci, 2019), the release of FinMA along with its financial instruction data and an evaluation benchmark (Xie et al., 2023), and the proposal of an open and data-centric pipeline for financial LLMs through FinGPT (Liu et al., 2023). Further comparative studied demonstrate that fine-tuned financial models can outperform few-shot prompting of larger general models on classification tasks (Fatemi and Hu, 2023; Fatemi et al., 2024) while numerous surveys document the rapid growth of financial applications for fine-tuning (Nie et al., 2024). However, the literature is primarily focused on whether fine-tuning raises task accuracy which is usually measured in a single evaluation condition instead of what happened to the model's behaviour when retrieved evidence is supplied but not addressed.

Retrieval Faithfulness and Context Utilisation

The question of model's using supplied context is different from whether retrieval finds the right text required. Pretrained language models function as implicit knowledge bases as established by Petroni et al. (2019), but this creates an avenue for conflict especially when retrieved text contradicts with what the model has already memorised during training. This phenomenon was studied by Longpre et al. (2021) using entity-substitution probes to discover that question-answering models frequently ignore supplied context in favour of parametric memory especially when the model knowledge increases. Additionally, Shuster et al. (2021) found that retrieval reduces rather than removes hallucination as it can persist even when relevant context is supplied.

This delicate balance can be altered by training intervention although the present direction is not clear. Xu et al. (2023) show that fine-tuned models know more and hallucinate less on semantic-parsing tasks while Jain et al. (2023) highlight through mechanistic analyses that fine-tuning often wraps existing capabilities rather than adding knowledge. Instruction tuning has also been noted to measurably shift the model's behaviour and attention allocation relative to the base model (Wu et al., 2023). Further

surveys of knowledge mechanisms in LLMs pinpoint the interplay between parametric and contextual knowledge as an open problem (Wang et al., 2024) while related work on calibration demonstrates that models can learn when to abstain under domain adaptation (Kamath et al., 2020). Alignment methods have also been proposed to make domain models prefer supplied context in question answering (Zhang et al., 2023). However, this literature characterises knowledge conflict in base or generically instruction-tuned models under synthetic perturbations while omitting the measurement of whether domain-specific fine-tuning on an institutional corpus alters a model's responsiveness to genuinely retrieved context from the same corpus.

Evaluation of RAG and Fine-Tuned LLMs

The evaluation of RAG-enabled and fine-tuned LLMs has largely developed along two distinct tracks with general-capability benchmarks such as GLUE, SuperGLUE, MMLU, BIG-bench, and HELM evaluating models in a fixed prompting condition (Hendrycks et al., 2020; Liang et al., 2022; Srivastava et al., 2022) while generated text is assessed using lexical-overlap metrics such as BLEU and ROUGE (Lin, 2004; Papineni et al., 2001). In addition, embedding-based metrics like BERTScore provide complementary semantic sensitivity (Haque et al., 2022; Zhang et al., 2019). However, known weaknesses of automatic metrics for open-ended generation are well documented (Gehrmann et al., 2022; Goyal et al., 2022).

The dimensions omitted by overlap metrics is introduced through RAG-specific evaluation such as RAGAS (Es et al., 2023). RAGAS operationalises faithfulness to retrieved context and answer relevance as automated measures over the question-context-answer structure. In finance, domain competence is assessed through dedicated benchmarks such as FLUE with the FLANG model, ConvFinQA for numerical reasoning over dialogue, and the PIXIU evaluation suite (Chen et al., 2022; Shah et al., 2022; Xie et al., 2023). Further surveys of LLM evaluation documents the proliferation of task-specific protocols (Tahmid et al., 2024). Consequently, much of the evaluation literature remains grounded in the closed-book assumption inherited from few-shot evaluations where models answer questions solely from information stored in their parameters (Brown et al., 2020) while retrieval-augmented systems are typically assessed separately in open-book condition. However, the literature rarely examines the relationship between these two conditions within the same model as evaluation studies compare systems using one condition at a time rather than as a pair. Thus, the extent to which a model's performance changes when external context is supplied and how adaptation methods such as fine-tuning influence that change remains unmeasured.

The Gap

Existing evaluation studies establish whether fine-tuning improves task performance or whether RAG improves grounding. In comparison, minimal attention has been given to whether fine-tuning alters a model's responsiveness to retrieved evidence despite the importance of this question for knowledge-intensive institutional applications. Prior research on knowledge conflict shows that parametric memory can override externally supplied context while the domain-adaptation literature shows that fine-tuning can embed substantial domain knowledge within model parameters. However, the interaction between these phenomena has rarely been examined directly. Few studies compare matched base and fine-tuned models on the same queries and retrieved text to determine how adaptation affects the use of external evidence. Therefore, this paper addresses this gap by providing a controlled measurement of the interaction between fine-tuning and retrieval by introducing context gain as an analytical measure which

quantifies the extent to which retrieved context improves model performance under matched evaluation conditions.

METHODOLOGY

Benchmark

To ensure strict reproducibility, all systems were evaluated on a frozen expert-curated 100-query benchmark comprising 50 factual-QA queries, 25 risk-assessment queries, and 25 summarisation queries. The benchmark was formed from a real SWF document corpus containing ~175 documents which produced ~1.78 million words after cleaning. The corpus consisted of annual reports, regulatory filings, press releases, and audit summaries to ensure that reference answers are anchored in verifiable institutional text rather than synthetic material. The 100-query benchmark was frozen before any system was trained or evaluated to ensure that every configuration faces the same task. The queries existed in both closed-book conditions where the model receives only the query and must answer from parametric knowledge and in context conditions where the prompt additionally supplies retrieved evidence through RAG. This retrieval step was executed once before its outputs are frozen to ensure that every system under context conditions receives the same top-4 retrieved chunks for a given query. Therefore, any differences between the systems under context conditions reflects differences in how the models use evidence rather than differences in the supplied evidence.

Furthermore, retrieval was implemented using a standard dense-retrieval architecture (Karpukhin et al., 2020) while the document corpus was segmented into 2,866 chunks of 800 tokens with 18.75% overlap. These chunks were embedded using all-MiniLM-L6-v2 and indexed in FAISS with cosine similarity using a retrieval depth of top-k = 4. On the benchmark, the mean top-1 similarity score was 0.637, and retrieval achieved a hit rate of 0.426 across 94 evaluable queries where hit rate is defined as the proportion of evaluable queries for which the gold source document appears among the four retrieved chunks (40 of 94). This modest hit rate is expected in real-world deployments, and it naturally limits overall context gain, but it cannot by itself produce the pattern this study measures. Retrieval failures therefore impact the models we measure equally since every system receives the identical four chunks for every query. Consequently, any difference in context gain reflects how each model uses the same evidence and this trend pattern holds even when retrieval succeeds. This is evident for the 10 factual questions where the reference figures were in the provided context, the LLaMA Instruct and Qwen Base models improved their NFA by +0.600 and +0.767, but their fine-tuned versions only gained -0.033 and +0.067.

Models and Configurations

The core design is a four-way ablation of a single model family. The configurations are M_Base (base model with closed-book), M_RAG (base model with context), M_SFT (fine-tuned model with closed-book), and SFT+RAG (fine-tuned model with context). This structure isolates the three contrasts the research question requires.

1. M_Base to M_RAG which measures the gain from retrieved evidence.
2. M_Base to M_SFT which measures the knowledge added by fine-tuning.
1. M_SFT to SFT+RAG which measures how much the fine-tuned model gains from evidence.

Furthermore, three more tiers were added to prevent confounding variables. Firstly, because the fine-tuned checkpoints were trained from the non-instruct base model, matched base-checkpoint controls were added to evaluate the exact starting checkpoint of the fine-tuned model. Secondly, an Instruct-backbone ablation was added to apply the same fine-tuning recipe to the instruction-tuned LLaMA-3.1-

8B backbone. This separates the effect of domain fine-tuning from prior instruction alignment. Thirdly, a cross-backbone replication was added to apply the same recipe, data, and serving stack to Qwen2.5-7B (Base). This tests if the behaviour is specific to one model family. Table 1 summarises the various configuration tiers.

Table 1: Evaluated Configuration Tiers

Tier	Configurations	Purpose
Core Four-Way Ablation	M_Base, M_RAG, M_SFT, and SFT+RAG	Isolate the base-to-RAG, base-to-FT, and FT-to-FT+RAG contrasts
Matched Base-Checkpoint Controls	LLaMA-3.1-8B Base Closed + Context	Compare fine-tuning against its exact initialisation checkpoint
Instruct-Backbone Ablation	LLaMA-3.1-8B-Instruct + Identical Fine-Tune Closed + Context	Separate domain fine-tuning from prior instruction alignment
Cross-Backbone Replication	Qwen2.5-7B Base and Fine-Tuned Closed + Context	Test whether the effect generalises beyond one model family

Fine-Tuning Recipe

The same parameter-efficient recipe of QLoRA (Dettmers et al., 2023) with LoRA adapters (Hu et al., 2021) of rank r = 16, scaling alpha = 32, and dropout 0.1 to give 41.94M trainable parameters (0.52% of the model) was adopted for all fine-tuned configurations. The training run lasted for 3,036 optimisation steps (two epochs) at learning rate 2e-4 with cosine decay and an effective batch size of 8 while seed 3407 was fixed for data splitting, training, and evaluation resampling throughout.

The training data was made up of 15,196 instruction examples generated deterministically from the document corpus through script-based extraction to ensure no examples was produced by a language model before the data was split 80/10/10 using seed 3407 into 12,143 training, 1,517 validation and 1,536 test examples. The dataset comprises 9,259 factual-QA, 4,370 summarisation, 876 risk-assessment and 691 stakeholder-communication examples. Additionally, every training example was screened with an 8-gram overlap filter against the frozen benchmark to prevent the model from memorising benchmark answers. This led to the removal of four records that met this criterion to prevent train-test leakage. Thus, any residual advantage of the fine-tuned models on the benchmark reflects learned domain knowledge rather than direct answer leakage at the surface level.

All locally hosted systems were served using an identical protocol of Q4_K_M quantisation and greedy decoding (temperature 0, top_k 1, seed 3407, and repeat penalty 1.1 over a 64-token window) with a 220-token completion budget. Maintaining an identical serving stack and decoding configuration across all served systems ensures that observed performance differences cannot be attributed to variations in sampling behaviour, decoding strategy, or quantisation settings. This experimental control also isolates model architecture and adaptation strategy as the primary sources of between-system variation. The LLaMA-3.1-8B Instruct baseline was the one exception in both closed and context conditions. We queried it through the OpenRouter API using the model's native chat template with temperature set at 0 and a 300-token completion budget, using the same query text and the same retrieved context as every other system.

Prompt Templates

We gave all locally served systems the raw Alpaca template below without a chat template, generation stopped at the end-of-sequence token using the "###" delimiter or 220-token budget or whichever came

first. Under closed-book conditions, the Input field is left empty while under context conditions it includes a fixed instruction followed by the four retrieved chunks. Each prompt was preceded by its source-document name before being truncated at 1,800 characters and the assembled context block truncated at 6,000 characters. Identical 6,000-character blocks were supplied to every context-condition system. The prompt template we used is as follows:

```
Below is an instruction that describes a task, paired with an input that
provides further context. Write a response that appropriately completes the
request.

### Instruction: {query}

### Input: {closed book: empty | context: "Use ONLY the following context to
answer.\n\nContext:\n" + context block}

### Response: Context block: [Source: {document 1}]\n{chunk 1} \n\n … \n\n
[Source: {document 4}]\n{chunk 4} (each chunk ≤ 1,800 characters; block ≤
6,000 characters)
```

We gave the LLaMA-3.1-8B Instruct baseline the exact same query along with the identical context block through its chat template for the context condition:

```
System: You are an assistant answering questions about the Nigeria Sovereign
Investment Authority (NSIA) for financial analysts. Answer concisely and
factually.
User (closed book): {query}
User (context): Use ONLY the following context to
answer.\n\nContext:\n{context block}\n\nQuestion: {query}
```

Metrics and the Context-Gain Analysis

Task performance is evaluated using the following five metrics aligned with the benchmark's underlying task types.

1. **Token-Level F1** to measure lexical overlap between generated and reference answers.
2. **Numeric-Fact Accuracy** to assess the proportion of normalised numerical values in the reference answer that are correctly reproduced in the generation. It is also treated as the primary grounding-sensitive measure.
3. **ROUGE-L** for evaluating summarisation tasks (Lin, 2004).
4. **SBERT** for assessing semantic similarity.
5. **Risk-Category Recall** for evaluating risk-assessment queries.

In institutional settings, the correctness of an answer often depends on the accurate reproduction of specific figures, including allocations, percentages, monetary values, and dates. This information is unlikely to be reproduced reliably unless it has either been retained in the model's parametric memory or incorporated from retrieved context. Thus, changes in NFA across evaluation conditions provides a more direct indication of evidence use than surface-overlap metrics which may reward fluent paraphrasing even when underlying facts are incorrect.

Consequently, to quantify how fine-tuning alters a model's responsiveness to retrieved evidence, this study introduces a simple analytical measure termed context gain. Thus, for any model M, context gain (CG) is defined as:

$$CG_M = Score(M_{context}) - Score(M_{closed}) \quad (1)$$

where Score represents performance on any given evaluation metric. Therefore, context gain captures the improvement in performance attributable to the addition of retrieved context. Hence, the primary quantity of interest is the change in context gain (ΔCG) after fine-tuning:

$$\Delta CG = CG_{FT} - CG_{Base} \quad (2)$$

This difference is computed within the same model family so that the underlying architecture, retrieval pipeline, serving configuration, and evaluation conditions remain constant. This design isolates the effect of adaptation on the model's use of retrieved evidence. A negative ΔCG indicates that the fine-tuned model benefits less from the same retrieved context than its base counterpart, suggesting reduced responsiveness to external evidence. Furthermore, CG and ΔCG are presented as descriptive analytical measures introduced for this study rather than as proposed general-purpose evaluation metrics. They are both simple differences between scores obtained under closed-book and context conditions and therefore inherit the measurement properties, strengths, and limitations of the underlying task metrics on which they are based.

Statistical Procedure

A paired bootstrap procedure with 10,000 resamples (B = 10,000) over the 100 benchmark queries was adopted to assess inter-system differences on the task metrics. This paired design exploits the fact that every system is evaluated on the same set of queries to aid in controlling for variation in task difficulty and providing a direct estimate of performance differences between systems. The Holm-Bonferroni correction (Holm, 1979) which is ideal given the large number of comparisons performed across systems, conditions, and metrics is applied across the complete test family to control the family-wise error rate (FWER). This conservative approach reduces the likelihood that reported pairwise differences arise from multiple-testing effects rather than genuine performance variation.

Furthermore, we estimate CG and ΔCG using the same paired bootstrap. Since all served systems answer the same queries, we directly resample the per-query differences (context minus closed-book score for CG, and the fine-tuned model's CG minus the base model's CG for ΔCG). This gives us 95% percentile confidence intervals and two-sided p-values. We then apply a Holm-Bonferroni correction to the six ΔCG tests (three backbone pairs × two factual-QA metrics) as a single family, summarising the results at the end of the context-gain analysis. Lastly, any interpretation of ΔCG applies only to the specific training procedures, model families, and experimental conditions tested in this study.

RESULTS AND DISCUSSION

Fine-Tuning Injects Institutional Knowledge

The first question is whether the fine-tuning intervention succeeded on its own terms by transferring institutional knowledge into the model's parameters. The closed-book condition where no retrieved context was provided offers a direct test of this proposition and the results indicate that it succeeded. On the matched LLaMA-3.1-8B backbone, closed-book NFA increased from 0.148 for the base model to 0.248 following domain-specific fine-tuning (difference +0.100, 95% CI [0.040, 0.172], $p < 0.001$). This represents a relative improvement of approximately two-thirds in the absence of retrieval. Similarly, this pattern is observed for factual-QA token-level F1 as the fine-tuned model achieved 0.245 in the closed-book condition against 0.214 for the zero-shot Instruct variant. This same effect is observed in the Qwen model family as token-level F1 increased from 0.170 on the base checkpoint to 0.220 after fine-tuning.

Therefore, under the training procedure evaluated, these results demonstrate that deterministic instruction data derived from the institutional corpus was sufficient to inject a measurable quantity of

domain-specific knowledge within model parameters. This finding is important for interpreting the results that follow as the evidence presented in the next subsection cannot be explained by an ineffective or substantively neutral fine-tuning intervention. Therefore, since the adapted models demonstrably acquired institutional knowledge and used that knowledge in the absence of external context, the central question is not whether fine-tuning worked but rather how the resulting parametric knowledge affected the model's responsiveness to retrieved context.

The Retrieval Benefit Collapses After Fine-Tuning

This section examines whether gains in memorised institutional knowledge were accompanied by changes in the model's willingness to use information supplied at inference time.

Table 2: Numeric-Fact Accuracy by Backbone Pair and Condition

Backbone Pair	System	Closed	Context	CG	ΔCG
LLaMA-3.1-8B (Base)	Base	0.148	0.331	+0.183	-0.161
	Base + Domain Fine-Tuning	0.248	0.269	+0.021	
LLaMA-3.1-8B-Instruct	Base	0.249	0.559	+0.310	-0.348
	Base + Domain Fine-Tuning	0.224	0.185	-0.039	
Qwen2.5-7B (Base)	Base	0.241	0.591	+0.350	-0.337
	Base + Domain Fine-Tuning	0.221	0.233	+0.012	

Table 2 reports NFA for the base and fine-tuned member of each of the three backbone pairs in closed-book and context conditions, along with each configuration's CG and the pair's ΔCG while Figure 1 plots the same comparison.

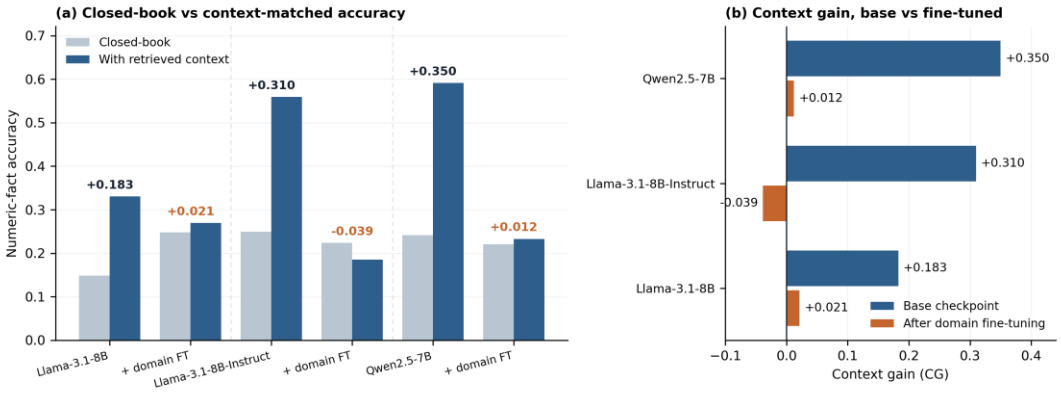


Figure 1: Grounding Effect of Retrieval on NFA Across Three Backbone Pairs

The grounding collapse pattern is consistent across all three matched model pairs as the base model translates retrieved context into a substantial improvement in NFA (+0.183 for the LLaMA base model, +0.310 for the Instruct model, and +0.350 for Qwen) against each fine-tuned counterpart deriving minimal to no benefit from the same retrieved context (+0.021 for the LLaMA base fine-tune, -0.039 for the Instruct model, and +0.012 for the Qwen fine-tune). These results respectively yield corresponding ΔCGs of -0.161, -0.348, and -0.337. Therefore, because each comparison is conducted within a matched experimental framework along with identical prompts, retrieved text, retrieval settings, and decoding parameters, the most plausible explanation under the evaluated conditions is that

the fine-tuning intervention itself reduced the model's responsiveness to retrieved evidence.
Furthermore, the absolute performance levels reinforce this conclusion as the highest context-condition NFA achieved by any fine-tuned model (0.269) remains below the lowest score achieved by any base model (0.331) under the same retrieval condition. This simply means that none of the retrieval benefit observed in the unadapted systems is recovered after fine-tuning. This shows that it raises the closed-book performance floor while weakening responsiveness to information provided in the context window. To better articulate this claim, Table 3 repeats the analysis for factual-QA token-F1 to highlight a more nuanced finding.

Table 3: Factual-QA Token-F1 by Backbone Pair and Condition

Backbone Pair	System	Closed	Context	CG	ΔCG
LLaMA-3.1-8B (Base)	Base	0.230	0.244	+0.014	+0.013
	Base + Domain Fine-Tuning	0.245	0.273	+0.028	
LLaMA-3.1-8B-Instruct	Base	0.214	0.315	+0.101	-0.120
	Base + Domain Fine-Tuning	0.205	0.186	-0.019	
Qwen2.5-7B (Base)	Base	0.170	0.235	+0.064	-0.117
	Base + Domain Fine-Tuning	0.220	0.167	-0.053	

Table 3 demonstrates that on token-level F1, the reduction in CG is evident in the Instruct (ΔCG -0.120) and Qwen (ΔCG -0.117) model families where the fine-tuned models perform worse when retrieval is added than when operating in the closed-book condition to respectively yield context gains of -0.019 and -0.053. However, the LLaMA base pair differ from this pattern as ΔCG is +0.013 meaning the fine-tuned model derived a slightly larger benefit from retrieved context (+0.028) than its base counterpart (+0.014). These results when looked at collectively suggest that the reduction in responsiveness to retrieved evidence is most pronounced on NFA which is the metric most sensitive to whether a model reproduces information contained in the retrieved text while remaining only partially visible on the more permissive lexical-overlap measure.
This pattern is consistent with the grounding tax interpretation because token-level F1 rewards the generation of topically relevant language and semantically appropriate phrasing which can both be produced from domain knowledge acquired during fine-tuning. However, accurate reproduction of numerical facts often requires the model to attend to and incorporate specific information from the retrieved context. Consequently, a model that increasingly relies on parametric memory may continue to perform well on lexical-overlap measures while showing markedly reduced gains on grounding-sensitive metrics. Therefore, this divergence between token-F1 and NFA suggests that fine-tuning affects both what the model knows and the extent to which it consults externally supplied evidence when generating answers.
However, while the 100-query benchmark includes risk assessment and summarisation tasks, our analysis focuses primarily on factual-QA (NFA and Token-F1). This was adopted because factual-QA offers the most direct and unambiguous measurement of parametric override since the model must make a discrete choice between reproducing a specific retrieved figure or relying on memorised weights. In contrast, summarisation and risk assessment are harder to measure because they reward paraphrasing and general domain knowledge, a fine-tuned model can score well just by relying on what it already knows, making it impossible to tell if it actually used the provided context. Accurately measuring context reliance in those tasks would require more granular metrics like claim-level faithfulness, an iteration which we leave for future work.

Statistically, a paired bootstrap (B = 10,000) corrected using Holm-Bonferroni across the six factual-QA tests confirms that fine-tuning significantly reduces CG for both the LLaMA Instruct and Qwen Base pairs. The Instruct model's NFA dropped by 0.348 (95% CI [-0.609, -0.104], p = 0.019) and its token-F1 also fell by 0.120 (95% CI [-0.182, -0.058], p < 0.001). The Qwen Base pair showed a similar pattern with NFA dropping by 0.337 (95% CI [-0.593, -0.099], p = 0.019) and token-F1 falling by 0.117 (95% CI [-0.178, -0.054], p = 0.004). The LLaMA Base pair was the exception as its NFA dropped (-0.161, 95% CI [-0.328, 0.000], p = 0.102). This implies it did not reach statistical significance and its token-F1 barely moved (+0.013, 95% CI [-0.070, 0.093], p = 0.736).

Component Ablation and Alignment Interaction

The closed-book factual-QA results also help separate the contributions of instruction alignment and domain adaptation. On token-level F1, the base-backbone fine-tune (0.245) outperforms both the zero-shot Instruct model (0.214) and the Instruct-backbone fine-tune (0.205). This implies that further application of the same domain-specific fine-tuning procedure to an instruction-aligned checkpoint yielded a system that performed slightly worse in the closed-book condition than the unadapted Instruct model. In contrast, applying the same procedure to the base checkpoint produced the strongest closed-book performance of the three models. Therefore, the effects of instruction alignment and domain fine-tuning were not additive under the evaluated training recipe.

The interaction becomes more apparent in the context condition as the Instruct-backbone pair exhibits the largest reduction in CG observed in the study (-0.348 on NFA) despite the instruction-tuned baseline displaying a substantial context gain of +0.310. Furthermore, abstention behaviour reinforces this pattern as the instruction-tuned RAG baseline declined to answer 6 of 100 queries while the fine-tuned RAG variant declined on 24 of 100 queries. Similarly, the Instruct-backbone fine-tuned model recorded an even higher abstention rate as it declined 45 of 100 queries. These results suggest that instruction alignment did not provide a neutral foundation for subsequent domain adaptation. Instead, it was associated with the strongest reduction in responsiveness to retrieved evidence and with a substantial increase in abstention behaviour when combined with the present fine-tuning procedure.

Collectively, the three matched backbone comparisons present a consistent story under the training recipe and experimental conditions evaluated in this study. The knowledge-injection results demonstrated that fine-tuning successfully injects institutional knowledge within model parameters. However, the context-gain results shows that it reduced the benefits derived from retrieved context on the study's primary grounding-sensitive metric while the Instruct-based models reversed those benefits entirely. The component ablation further indicates that the strongest effect occurred when domain adaptation was layered on an already instruction-aligned checkpoint. This overall pattern is consistent with what we term the grounding tax where domain fine-tuning extracts a measurable cost from the model's retrieval responsiveness. The underlying mechanism is parametric override where the adapted model preferentially answers from memorised parametric knowledge rather than from the institutional evidence explicitly supplied. Therefore, this interpretation provides the most coherent explanation for the observed combination of improved closed-book performance, diminished context gain, and elevated abstention rates within the limits of the evaluated design.

Discussion

The results point to a simple conceptual model where domain fine-tuning writes knowledge into model parameters to improve closed-book performance. However, this fine-tuning creates a second effect where a model that receives retrieved evidence has its parametric memory and the supplied context

competing. It was observed that base models used the context to gain up to +0.350 in NFA while the fine-tuned models gained almost nothing. This phenomenon is the grounding tax, driven by parametric override and under our training recipe, fine-tuning reduced CG which indicates that the model relied on memorised knowledge instead of the supplied context. This reading matches the literature on knowledge conflict as Longpre et al. (2021) already showed us that models prefer memorised sources when the two disagree. Further surveys also describe this dual-source structure and the difficulty of controlling it (Pan et al., 2023; Wang et al., 2024). Thus, our contribution is showing that a standard domain-adaptation recipe can shift this balance toward the parametric side while using the same queries, retrieval pipeline, and decoding settings.

A further qualitative inspection of the model's outputs supports our claim as the fine-tuned model preferred memorised formulations of facts even when the correct figure was in the retrieved text. Whereas the instruction-tuned RAG baseline deferred to the context more reliably. This is because instruction-tuned models are trained to follow user framing including instructions to use provided evidence (Ouyang et al., 2022). However, our fine-tunes appear to have partially overwritten this behaviour. This issue has three main implications as follows.

Closed-Book Evaluation is Insufficient: A fine-tuned model can improve closed-book accuracy as ours did on numeric facts while simultaneously losing the ability to use retrieved evidence. Thus, evaluating a domain adaptation on only closed-book condition hides this degradation as the evaluation rewards precisely the internalisation that produces the override. This is observed in our results as the closed-book comparison compliments the fine-tune (0.248 against 0.148 on NFA) while the context condition reverses the ordering (0.2693 against 0.3307). Therefore, closed-book scores can conceal a loss that only appears when the model is asked to use evidence. This means that the two conditions must be reported together for any system that wants to be retrieval integrated.

Fine-Tuning and RAG are not Automatically Synergistic: The two concepts are often treated as stackable improvements where retrieval supplies fresh evidence and reduces hallucination (Lewis et al., 2020; Shuster et al., 2021) while fine-tuning and instruction tuning supply domain competence and compliance with the task framing (Ouyang et al., 2022; Parthasarathy et al., 2024). However, our results indicate that these contributions are separable and can interfere with each other. Additionally, the strongest context user was a base instruction-tuned model rather than a fine-tuned variant. This shows that adding fine-tuning to the Instruct backbone drove context gain negative. Therefore, combining these components requires balancing them and the combined system must be evaluated holistically rather than assumed to inherit the strengths of its parts.

High-Stakes Applications Need Context-Conditioned Evaluation: In regulated settings, where answers must be traceable to authoritative documents, the operative question is not "does the model know the answer" but "what does the model do when the correct evidence is explicitly supplied". Thus, a model that overrides supplied evidence with a memorised approximation fails this test even when its memory is often right because its outputs can no longer be audited against the documents it was given. Therefore, we suggest that any fine-tuned RAG system should report CG alongside accuracy. This metric is cheap to compute from the closed and context conditions that a rigorous evaluation already runs, and a small or negative CG clearly warns that the system ignores its supplied evidence. This complements existing groundedness measures for RAG pipelines (Es et al., 2023) by attributing the failure to the training recipe rather than the retriever. Finally, future work could test retrieval-aware fine-tuning or calibrated abstention to fix this behaviour.

The grounding tax observed in this study has a plausible explanation rooted in the composition of the training data. The 15,196-example fine-tuning dataset consisted of 9,259 factual-QA items with only

about 26.6% rewarding the model for copying figures from a supplied context. Whereas, the rest of the training data trained the models to answer without relying on supplied context, including 3,688 examples that had no context at all. Similarly, 997 refusal examples were paired with unrelated context while 1,541 of the 3,459 grounded question-answer examples had no target figures in the shown context. Collectively, our training mixture teaches the model to answer first and check context second which is the opposite of what a retrieval-grounded system requires. This shows the direction of the training signal rather than proving that it is the cause. Crucially, this behaviour is best assumed as an artefact of the training data mixture rather than an inherent universal flaw of fine-tuning.

This finding carries a concrete recommendation for researchers building domain-adapted RAG systems because if the primary value proposition is querying a live document store, the safest architecture to deploy is retrieval over a strong instruction-tuned checkpoint without domain fine-tuning. Whereas, where domain fine-tuning is needed for closed-book institutional recall, it should be treated as a targeted intervention evaluated under both closed-book and context-conditioned settings before deployment. Additionally, training recipes for retrieval-grounded systems should explicitly include retrieval-aware examples where the model is rewarded for copying figures from supplied text, including texts that contradict its parametric knowledge. As a resolution, Direct Preference Optimisation (DPO) offers a promising alternative training signal that could decouple safety-based abstention from grounding behaviour even though empirical validation of this approach remains future work.

Limitations

Several features of the experimental design bound what can be claimed. However, the most important is that all fine-tuning runs used a single random seed (3407), a single training recipe, and no exhaustive hyperparameter sweep was conducted. Thus, the override effect applies only to our specific recipe rather than to all domain fine-tuning. This makes it possible that a different data mixture or lighter adaptation might preserve or even improve context responsiveness. The benchmark comprises 100 frozen queries and it is sufficient to detect the large context-gain reversals reported, but it limits any statistical resolution for smaller effects. Similarly, the study draws on one institutional corpus in a single domain which makes the findings untested across domains with different terminology density, document styles, or numeric content.

The three fine-tuned adapters were only tested across two backbone families (LLaMA and Qwen), and the direction of the context-gain reduction is replicated across all three to strengthen the finding. However, broader architectural coverage is needed to generalise beyond 7-8B decoder models. All local systems were served with Q4_K_M quantisation and greedy decoding. Because this applies identically to every local configuration, it cannot explain differences between them even though absolute scores may differ from full-precision serving. The LLaMA Instruct baseline was the exception and was served through an API using its chat template, so the Instruct model's CG contrast also includes a difference in serving stack and prompt format. Furthermore, since the retrieval hit rate of 0.426 caps the maximum context gain, the retrieved text lacked the reference answer for more than half of the queries. However, this cap applies equally all comparisons as the fine-tuned and non-fine-tuned systems received identical top-4 chunks. This implies that imperfect retrieval did not create the observed asymmetry, but it still compresses the measurable range of the effect and in any realistic deployment, a rational model should fall back on parametric memory when the retriever fails.

Furthermore, two measurement caveats apply with the Token-F1 showing negative ΔCG in only two of the three backbone pairs. Thus, the conclusion rests mostly on NFA as the effect is clearest there. Finally, the high abstention rates of the fine-tuned systems complicate reference-overlap metrics as they

score a declined answer as a miss irrespective of whether declining was appropriate. In conclusion, the findings can be strengthened by multi-seed replication of the fine-tuning runs, recipe ablations isolating which component of the training data drives the effect, and replication on additional institutional corpora and domains.

CONCLUSION

This study examined whether domain-specific fine-tuning improves a language model's use of retrieved institutional evidence, or whether it shifts the model towards its memorised parametric knowledge. We tested this with a controlled ablation on a single institutional corpus. This involved the evaluation of three backbone pairs (LLaMA-3.1-8B Base, its Instruct variant, and Qwen2.5-7B Base) in matched closed-book and context-conditioned settings on a frozen 100-query benchmark after identical QLoRA fine-tuning runs using identical prompts, retrieval, and decoding throughout.

Three findings emerged from this empirical evaluation. Firstly, knowledge injection works in the closed-book settings as the fine-tuned LLaMA base model raised NFA from 0.148 to 0.248 over its matched control. Secondly, context gain collapses after fine-tuning across all three backbone pairs as the reduction ranged from -0.161 to -0.348 on NFA and was statistically significant for the LLaMA Instruct and Qwen Base pairs. Thirdly, instruction alignment amplified the grounding collapse effect as the Instruct-backbone fine-tune showed the largest collapse and the highest abstention rate after declining 45 of 100 queries. However, our claim is deliberately bounded because under our training recipe and experimental conditions, domain-specific fine-tuning reduced context gain. This provides evidence of a parametric override effect, but we do not claim that fine-tuning always degrades retrieval grounding but rather a standard and carefully controlled recipe did so consistently during this study.

The practical recommendation is clear as closed-book evaluation alone would call the fine-tuning a success while context-conditioned evaluation alone would have reported them as failures. Therefore, fine-tuned RAG systems must jointly be evaluated under both settings while context gain must also be reported alongside absolute scores so models that ignore supplied evidence cannot pass unnoticed. Future work should replicate the effect across multiple seeds, ablate components of the recipe to isolate the problematic mechanism, and test retrieval-aware fine-tuning to preserve injected knowledge without suppressing context use.

REFERENCES

Araci, D. T. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. https://arxiv.org/pdf/1908.10063

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... Zhu, T. (2023). Qwen Technical Report. https://arxiv.org/pdf/2309.16609

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 2020-December. https://arxiv.org/pdf/2005.14165

Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., ... Wang, W. Y. (2024). A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. http://arxiv.org/abs/2405.01769

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., & Wang, W. Y. (2022). ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 6279-6292. https://doi.org/10.18653/v1/2022.emnlp-main.421

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. *Advances in Neural Information Processing Systems*, 35. https://doi.org/10.52202/068431-2198

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Efeoglu, S., & Paschke, A. (2024). Relation Extraction with Fine-Tuned Large Language Models in Retrieval Augmented Generation Frameworks. Proceedings of the 2024 ACM Conference on Information Technology Education. https://dl.acm.org/doi/10.1145/3677619.3680654

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). Ragas: Automated Evaluation of Retrieval Augmented Generation. *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, 150-158. https://arxiv.org/pdf/2309.15217

Fatemi, S., & Hu, Y. (2023). A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis. http://arxiv.org/abs/2312.08725

Fatemi, S., Hu, Y., & Mousavi, M. (2024). A Comparative Analysis of Instruction Fine-Tuning LLMs for Financial Text Classification. http://arxiv.org/abs/2411.02476

Gehrmann, S., Clark, E., & Sellam, T. (2022). Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77, 103-166. https://doi.org/10.1613/jair.1.13715

Goyal, T., Li, J. J., & Durrett, G. (2022). News Summarization and Evaluation in the Era of GPT-3. https://arxiv.org/pdf/2209.12356

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Ma, Z. (2024). The Llama 3 Herd of Models. https://arxiv.org/pdf/2407.21783

Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic Similarity Metrics for Evaluating Source Code Summarisation. Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension (ICPC). https://doi.org/10.1145/3524610.3527882

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. *ICLR 2021 - 9th International Conference on Learning Representations*. https://arxiv.org/pdf/2009.03300

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 65-70. https://www.jstor.org/stable/4615733

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022 - 10th International Conference on Learning Representations*. https://arxiv.org/pdf/2106.09685

Jain, S., Kirk, R., Lubana, E. S., Dick, R. P., Tanaka, H., Grefenstette, E., ... Krueger, D. (2023). Mechanistically analysing the effects of fine-tuning on procedurally defined tasks. *12th International Conference on Learning Representations, ICLR 2024*. https://arxiv.org/pdf/2311.12786

Kamath, A., Jia, R., & Liang, P. (2020). Selective Question Answering under Domain Shift. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5684-5696. https://doi.org/10.18653/V1/2020.ACL-MAIN.503

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Lee, J., Jung, W., & Baek, S. (2024). In-House Knowledge Management Using a Large Language Model: Focusing on Technical Specification Documents Review. *Applied Sciences 2024, Vol. 14, Page 2096*, 14(5), 2096. https://doi.org/10.3390/APP14052096

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 2020-December. https://arxiv.org/pdf/2005.11401

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... Koreeda, Y. (2022). Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1), 140-146. https://doi.org/10.1111/nyas.15007

Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries.

Liu, X., Wang, G., Yang, H., & Zha, D. (2023). FinGPT: Democratizing Internet-scale Data for Financial Large Language Models. http://arxiv.org/abs/2307.10485

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-Based Knowledge Conflicts in Question Answering. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 7052-7063. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.565

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). https://doi.org/10.1093/BIB/BBAC409

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., & Bossan, B. (2022). PEFT: State-of-the-art Parameter-Efficient Fine-Tuning Methods.

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. http://arxiv.org/abs/2406.11903

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35. https://doi.org/10.52202/068431-2011

Pan, J. Z., Razniewski, S., Kalo, J., Singhania, S., Chen, J., Dietze, S., ... Jabeen, H. (2023). Large Language Models and Knowledge Graphs: Opportunities and Challenges. arXiv preprint. https://arxiv.org/pdf/2308.04174

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2001). BLEU. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. (2024). The Ultimate Guide to Fine-Tuning LLMs. *arXiv preprint*. https://arxiv.org/pdf/2408.13296v1

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language Models as Knowledge Bases?. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2463-2473. https://doi.org/10.18653/v1/d19-1250

Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., ... Yang, D. (2022). WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2322-2335. https://doi.org/10.18653/v1/2022.emnlp-main.148

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 3784-3803. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... Natarajan, V. (2022). Large Language Models Encode Clinical Knowledge. *Nature*, 620(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... Gray, A. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. https://arxiv.org/pdf/2206.04615

Tahmid, R. L. M., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., ... Huang, J. X. (2024). A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... Hashimoto, T. B. (2023). Alpaca: A Strong, Replicable Instruction-Following Model. https://crfm.stanford.edu/2023/03/13/alpaca.html

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., ... Stojnic, R. (2022). Galactica: A Large Language Model for Science. https://arxiv.org/pdf/2211.09085

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., ... Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. http://arxiv.org/abs/2302.13971

Wang, M., Yao, Y., Xu, Z., Qiao, S., Deng, S., Wang, P., ... Zhang, N. (2024). Knowledge Mechanisms in Large Language Models: A Survey and Perspective. https://arxiv.org/pdf/2407.15017v1

Wu, X., Yao, W., Chen, J., Pan, X., Wang, X., Liu, N., & Yu, D. (2023). From Language Modelling to Instruction Following: Understanding the Behaviour Shift in LLMs after Instruction Tuning. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, 1, 2341-2369. https://doi.org/10.18653/v1/2024.naacl-long.130

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2023). PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. http://arxiv.org/abs/2306.05443

Xu, S., Liu, S., Culhane, T., Pertseva, E., Wu, M. H., Semnani, S. J., & Lam, M. S. (2023). Fine-tuned LLMs Know More, Hallucinate Less with Few-Shot Sequence-to-Sequence Semantic Parsing over Wikidata. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 5778-5791. https://doi.org/10.18653/v1/2023.emnlp-main.353

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *8th International Conference on Learning Representations, ICLR 2020*. https://arxiv.org/pdf/1904.09675

Zhang, Y., Chen, Z., Fang, Y., Lu, Y., Li, F., Zhang, W., & Chen, H. (2023). Knowledgeable Preference Alignment for LLMs in Domain-specific Question Answering. https://arxiv.org/pdf/2311.06503

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... Irving, G. (2019). Fine-Tuning Language Models from Human Preferences. https://arxiv.org/pdf/1909.08593