

RETRIEVAL-AUGMENTED GENERATION AS NOISY IN-CONTEXT LEARNING: A UNIFIED THEORY AND RISK BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) has seen many empirical successes in recent years by aiding the LLM with external knowledge. However, its theoretical aspect has remained mostly unexplored. In this paper, we propose the first finite-sample generalization bound for RAG in in-context linear regression and derive an exact bias-variance tradeoff. Our framework views the retrieved texts as query-dependent noisy in-context examples and recovers the classical in-context learning (ICL) and standard RAG as the limit cases. Within this simplified model, our analysis suggests that an intrinsic ceiling on generalization error can exist on RAG as opposed to the ICL. Furthermore, our framework is able to model retrieval both from the training data and from external corpora by introducing uniform and non-uniform RAG noise. Finally, our theoretical insights of RAG is consistent with preliminary experimental results common QA benchmarks, such as Natural Questions and TriviaQA.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) enhances language models by appending retrieved data to the input, enabling access to information beyond pretraining. It is widely used in open-domain QA, fact-checking, and knowledge-intensive tasks (Huang et al., 2023; Lewis et al., 2020a; Ramos et al., 2022; Sarto et al., 2022; Zhao et al., 2024a). Retrieval sources typically fall into two categories: (1) *labeled dataset*, such as the training dataset itself (Liu et al., 2021; Izacard et al., 2022; Huang et al., 2024), and (2) *generic corpora without labels*, such as Wikipedia (Chen et al., 2017). Despite its promise, empirical studies show that increasing the number of retrieved passages can degrade performance, especially when irrelevant or redundant texts are included (Levy et al., 2025; 2024). However, the theoretical aspects for understanding how retrieval affects generalization remain underexplored.

To study its behavior, we frame RAG as noisy in-context learning (ICL). ICL refers to the ability of language models to adapt given the contextual information without updating model weights (Dong et al., 2024). Under this view, retrieved RAG examples can act as noisy context and their quality depends on the retrieval. This view has motivated the development of many works in in-context retrieval (Luo et al., 2024; Shi et al., 2022), where the goal is to retrieve high-quality demonstration pairs, which reduce the noise of the retrieval.

Our framework viewing RAG as noisy ICL needs to address novel challenges. Prior ICL work has analyzed structured in-context learning where the context consists of fixed format demonstration examples, assuming the examples are clean and i.i.d. (Ahn et al., 2023; Zhang et al., 2024). These assumptions do not hold in RAG, since the retrieved examples are inherently query-dependent and noisy, and their noise level tends to be inversely correlated to their relevance. Furthermore, although retrieved examples close to the query should, in principle, improve the predictive accuracy, their quantitative contribution remains unknown because RAG introduces these examples only at the test time (absent during pretraining), thus imposing a distribution shift. In this work, we take an initial step towards bridging the gap by modeling RAG as a variant of ICL that keeps the two core ingredients, *query dependence* and *noisy examples*, while abstracting away other system details for tractability. To model the noises, we propose and examine two noise regimes: uniform (noise level identical across examples) and non-uniform (noise inversely correlated with relevance), which capture key phenomena seen in practice. Indeed, our experiments on common QA benchmarks show that the two categories of RAG (using training datasets or generic corpora for retrieval) lead to experimental results that align well with our two noise regimes, respectively. This framework thus allows us to quantify the impact of retrieval noise and derive generalization bounds that depend on the number of in-context and RAG examples, and the retrieval distance from queries.

Our contributions are summarized as follows:

- We propose a simplified theoretical framework for analyzing RAG and provide the first finite sample bounds for in-context linear regression with RAG. Our bounds show that the improvement from RAG shrinks as you add more retrieved examples, and can even flip to hurt performance.
- Our framework recovers ICL and standard RAG as limit cases, and also models retrieved data under different noise regimes: uniform and non-uniform retrieval noises.
- We develop new tools for analyzing the query-dependent RAG data, e.g., a derivation of the expectation for high order Gaussian monomial (Lemma 3), which can be useful for future research on RAG.
- We conduct experiments for representative models on common QA datasets and demonstrate that the results align with our theory about the two noise regimes, providing positive support for our analysis.

2 RELATED WORK

Retrieval Augmented Generation Retrieval-augmented generation (RAG) has emerged as a widely adopted paradigm for enriching LLMs with external knowledge by prepending retrieved passages to the input context (Lewis et al., 2020a; Izacard & Grave, 2020; Borgeaud et al., 2021). From a functional perspective, RAG transforms the model’s input distribution by conditioning generation on retrieved textual evidence, often drawn from large-scale corpora via learned or heuristic retrieval mechanisms (Li et al., 2023; Meng et al., 2024; Chen et al., 2024). While much of the literature focuses on improving retrieval quality, system performance (Asai et al., 2023; Li et al., 2024; Xu et al., 2024a), and answer reliability (Xiang et al., 2024; Xu et al., 2024a). There are some recent attempts in studying the theory of RAG (Xu et al., 2024b; Kang et al., 2024) focusing on the token-level risk, and the sequence-level risk still remains underexplored. **In-context Learning (ICL)** ICL obtains its popularity from the original GPT-3 paper (Brown et al., 2020), and becomes widely used in LLM applications (Dong et al., 2024; Min et al., 2021). The recent advance in ICL theory (Ahn et al., 2023; Zhang et al., 2024; Xie et al., 2021) provides a rigorous and versatile framework to study transformers and LLMs. People have used this ICL framework to study novel settings, like out-of-distribution tasks (Wang et al., 2024b) and test-time training (Gozeten et al., 2025). People have also studied the noisy in-context learning from robustness (Cheng et al., 2025) and calibration perspectives (Zhao et al., 2024b), which are different from our setup.

In-context Retrieval In-context retrieval (Luo et al., 2024) refers to retrieving a set of query-dependent demonstrations rather than using a fixed set of demonstrations. The label of the demonstration pairs can come from various sources, such as in-domain training set (Izacard et al., 2022; Huang et al., 2024; Ye et al., 2023), cross-domain data (Cheng et al., 2023; Shi et al., 2022), automatic LLM generation (Zhang et al., 2022; Li & Qiu, 2023), pseudo-labels from unstructured data (Lyu et al., 2022; Li et al., 2022). In our theoretical analysis and experiments, we focus on the simplest in-context retrieval, in-domain retrieval from the training set, as in (Izacard et al., 2022; Huang et al., 2024). Note that in-context retrieval is a term developed later and some earlier papers discuss ICL with retrieval as retrieving relevant documents without labels (Ram et al., 2023).

3 PROBLEM SETUP

Our problem setup is similar to (Zhang et al., 2024; Garg et al., 2022), but with RAG examples to form the additional in-context examples. It is worth noting that many works focus on ICL at test (inference) time, specifically without parameter updates (Dong et al., 2022). Our work adopts the framework of *ICL with warmup*, also known as, *supervised in-context training*. Specifically, we assume that the pretraining data is also formed by in-context examples. Then, during the test time, we formed prompts with in-context examples with additional RAG examples.

Notations We denote $[n] = \{1, \dots, n\}$ for an integer $n \geq 1$. We denote the trace product of two matrices $A, B \in \mathbb{R}^{m \times n}$ as $\text{tr}(AB^\top)$. We denote $f(m, n) = \mathcal{O}_{m,n}(g(m, n))$, iff there exists $c, m_0, n_0 > 0$ such that $0 \leq f(m, n) \leq cg(m, n)$, $\forall m \geq m_0, n \geq n_0$. We denote $f(m, n) = \Theta_{m,n}(g(m, n))$, iff there exists $c_1, c_2, m_0, n_0 > 0$ such that $c_1g(m, n) \leq f(m, n) \leq c_2g(m, n)$, $\forall m \geq m_0, n \geq n_0$.

Pretraining Data We consider learning over linear regression data. The training data is a set of prompts. Each prompt is of size m : $(\mathbf{x}_1, y_1, \dots, \mathbf{x}_m, y_m, \mathbf{x}_q) \in \mathbb{R}^{d(m+1)+m}$ where $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ form the m demonstration pairs. The goal is to predict $\hat{y}_q$ for the query example $\mathbf{x}_q$ to match the true label y_q . The prompt is embedded in the following form:

$$\mathbf{p}_m^{\text{pt}} := \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_m & \mathbf{x}_q \\ y_1 & y_2 & \dots & y_m & 0 \end{pmatrix} \in \mathbb{R}^{(d+1) \times (m+1)}, \quad (1)$$

where $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m), (\mathbf{x}_q, y_q) \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_{\text{pt}}$ (pt denoting Pretraining). The output follows the linear model:

$$y_i = \mathbf{x}_i^\top \beta_{\text{pt}} + \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{under } \mathcal{D}_{\text{pt}}, \quad (2)$$

where $i \in [m] \cup \{q\}$, β_{pt} is the weight vector in pretraining, and ϵ_i is the noise for example i .

Inference Data (with RAG) During inference/test time, the test prompt $\mathbf{P}_{m,n}^{\text{tt+rag}}$ (tt denoting test-time) is formed by m in-context pairs $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$, n retrieval-augmented pairs $(\mathbf{x}_1^{\text{rag}}, y_1^{\text{rag}}), \dots, (\mathbf{x}_n^{\text{rag}}, y_n^{\text{rag}})$, and the query pair $\mathbf{x}_q, y_q$. The test prompt is embedded in the following form:

$$\mathbf{P}_{m,n}^{\text{tt+rag}} := \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_m & \mathbf{x}_1^{\text{rag}} & \dots & \mathbf{x}_n^{\text{rag}} & \mathbf{x}_q \\ y_1 & \dots & y_m & y_1^{\text{rag}} & \dots & y_n^{\text{rag}} & 0 \end{pmatrix} \in \mathbb{R}^{(d+1) \times (m+n+1)}. \quad (3)$$

The input $\mathbf{x}$ in each in-context or query pair follows the test-time distribution $\mathcal{D}_{\text{tt}}$, and the label is:

$$y_i = \mathbf{x}_i^\top \beta_{\text{tt}} + \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{under } \mathcal{D}_{\text{tt}}, \quad (4)$$

where $i \in [m] \cup \{q\}$, ϵ_i is the noise of example i , and β_{tt} is the weight vector during test time. The input $\mathbf{x}$ in each RAG pair follows the corresponding RAG distribution $\mathcal{D}_{\text{rag}}(\mathbf{x}_q)$: assume the RAG query $\mathbf{x}_i^{\text{rag}} = \mathbf{x}_q + \mathbf{r}_i$ is generated around the query example $\mathbf{x}_q$, where $\mathbf{r}_i$ is the offset. The label in the RAG example is given by:

$$y_i^{\text{rag}} = (\mathbf{x}_i^{\text{rag}})^\top \beta_{\text{tt}} + \epsilon_i^{\text{rag}}, \quad \epsilon_i^{\text{rag}} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\text{rag},i}^2) \quad \text{under } \mathcal{D}_{\text{rag}}(\mathbf{x}_q), \quad (5)$$

where $i \in [n]$, ϵ_i^{rag} is the noise of the i -th RAG example.

For the compactness of writing, we define the following matrices and vectors:

$$\begin{aligned} \mathbf{X}_{\text{icl}} &:= [\mathbf{x}_1^\top; \dots; \mathbf{x}_m^\top], \quad \mathbf{X}_{\text{rag}} := [(\mathbf{x}_1^{\text{rag}})^\top; \dots; (\mathbf{x}_n^{\text{rag}})^\top], \quad \mathbf{y}_{\text{icl}} := [y_1; \dots; y_m], \quad \mathbf{y}_{\text{rag}} := [y_1^{\text{rag}}; \dots; y_n^{\text{rag}}], \\ \boldsymbol{\epsilon}_{\text{icl}} &:= [\epsilon_1; \dots; \epsilon_m], \quad \boldsymbol{\epsilon}_{\text{rag}} := [\epsilon_1^{\text{rag}}; \dots; \epsilon_n^{\text{rag}}], \quad \mathbf{r} = [\mathbf{r}_1^\top; \dots; \mathbf{r}_n^\top] \\ \mathbf{X} &= \begin{bmatrix} \mathbf{X}_{\text{icl}} \\ \mathbf{X}_{\text{rag}} \end{bmatrix} \in \mathbb{R}^{(m+n) \times d}, \quad \mathbf{X}_{\text{rag}} = \begin{bmatrix} \mathbf{x}_q + \mathbf{r}_1 \\ \vdots \\ \mathbf{x}_q + \mathbf{r}_n \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_{\text{icl}} \\ \mathbf{y}_{\text{rag}} \end{bmatrix} \in \mathbb{R}^{m+n}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_{\text{icl}} \\ \boldsymbol{\epsilon}_{\text{rag}} \end{bmatrix} \in \mathbb{R}^{m+n}. \end{aligned}$$

Training and Testing We let $\mathbf{W}$ be the model parameters, and F be the model. Given an input prompt $\mathbf{P}_m^{\text{pt}}$ with demonstration pairs, the model predicts $\hat{y}_q := F(\mathbf{P}_m^{\text{pt}}; \mathbf{W})$. As a common practice in theoretical studies of LLM for feasible analysis, we use the MSE loss as the evaluation metrics (Zhang et al., 2024; Ahn et al., 2023; Xie et al., 2021). Then, the population loss on the pretraining data is:

$$\mathcal{L}_{\text{pt}}(\mathbf{W}) := \mathbb{E}_{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m), (\mathbf{x}_q, y_q) \sim \mathcal{D}_{\text{pt}}} \left[(y_q - F(\mathbf{P}_m^{\text{pt}}; \mathbf{W}))^2 \right]. \quad (6)$$

Its minimizer is denoted as:

$$\bar{\mathbf{W}}^* := \min_{\mathbf{W}} \mathcal{L}_{\text{pt}}(\mathbf{W}). \quad (7)$$

To apply the pretrained $\bar{\mathbf{W}}^*$ from the pretraining context size of m to the test-time context size of $m+n$, we will need to scale it properly (see Lemma 1) and use

$$\mathbf{W}^* = \frac{m}{m+n} \bar{\mathbf{W}}^*. \quad (8)$$

During the test time we evaluate the population loss over the test prompt with RAG examples $\mathbf{P}_{m,n}^{\text{tt+rag}}$:

$$\mathcal{L}_{\text{tt+rag}}(\mathbf{W}) := \mathbb{E}_{\substack{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m), (\mathbf{x}_q, y_q) \sim \mathcal{D}_{\text{tt}} \\ (\mathbf{x}_1^{\text{rag}}, y_1^{\text{rag}}), \dots, (\mathbf{x}_n^{\text{rag}}, y_n^{\text{rag}}) \sim \mathcal{D}_{\text{rag}}(\mathbf{x}_q)}} \left[(y_q - F(\mathbf{P}_{m,n}^{\text{tt+rag}}; \mathbf{W}))^2 \right]. \quad (9)$$

Model Architecture We study the single-layer linear self-attention model (LSA) as the framework for theoretical analysis, similar to many existing studies (e.g., Ahn et al., 2023; Zhang et al., 2024). The prediction of the model F on a prompt $\mathbf{P}$ with query $\mathbf{x}_q$ is:

$$\hat{y}_q := F(\mathbf{P}) = [\mathbf{P}\mathbf{W}_Q \mathbf{W}_K^\top \mathbf{P}\mathbf{W}_V]_{m+n+1, d+1} = \mathbf{x}_q^\top \mathbf{W} \mathbf{X}^\top \mathbf{y}, \quad (10)$$

where the query, key, and value matrices $\mathbf{W}_Q, \mathbf{W}, \mathbf{W}_V \in \mathbb{R}^{(d+1) \times (d+1)}$ are parameterized by $\mathbf{W}$ in the following way:

$$\mathbf{W}_Q \mathbf{W}_K^\top = \begin{bmatrix} \mathbf{W} & \mathbf{0}_{d \times 1} \\ \mathbf{0}_{1 \times d} & 0 \end{bmatrix}, \quad \mathbf{W}_V = \begin{bmatrix} \mathbf{0}_{d \times d} & \mathbf{0}_{d \times 1} \\ \mathbf{0}_{1 \times d} & 1 \end{bmatrix}.$$

We note that this parameterization is commonly used in the previous works (Ahn et al., 2023; Zhang et al., 2024), and is shown to capture the key properties of in-context learning. Furthermore, (Ahn et al., 2023) shows that the formulation is the optimum converged from pretraining on Gaussian data.

4 THEORETICAL ANALYSIS: GENERALIZATION BOUND FOR RAG

To study test-time error and sample complexity in in-context linear regression with RAG examples, we consider two noise regimes: **uniform retrieval noise** and **non-uniform retrieval noise**. Uniform retrieval noise assumes the RAG noise ϵ_i^{rag} for each example i is i.i.d. Since its variance is distance-agnostic, it can model a scenario of retrieval where the noise is similar across data points. Non-uniform retrieval noise assumes either the variance or the label-corruption probability grows with the variance of retrieval vectors — e.g. $\sigma_{\text{rag},i}^2$ increases with δ_i^2 or the probability of making mistakes increases with δ_i^2 . This captures retrieval from datasets where near neighbors often supply the right signal while far ones are potentially noisy or even misleading. Because the noise spectrum is now heavy-tailed, adding more RAG examples past a threshold could yield diminishing benefits for RAG examples and even become counter-productive. Framing RAG through these two lenses provides intuition about when extra retrieved examples will pay off, and when they will hit the intrinsic ceiling and more retrieved examples do not help anymore. These are well corroborated by our experimental results on real data (see Section 5).

First, we introduce the key data assumptions.

Assumption 1 (Gaussian Retrieval Offset). *We assume the retrieval offset $\mathbf{r}_i$, $\forall i \in [n]$ to follow a Gaussian distribution: $\mathbf{r}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \delta_i^2 \mathbf{I}_d)$.*

The key property that we want to control for RAG examples is its distance from the query points $\mathbf{x}_q$. However, modeling the queried example directly through the retrieval distance leads to complicated theoretical analysis. Here, we note that the retrieval distance $\|\mathbf{r}_i\|_2$ converges to a distribution concentrated in an $\mathcal{O}(\delta_i \sqrt{d})$ ball around the query with respect to d (Cover & Hart, 1967). Thus, controlling the variance of the retrieval offset can alternatively control the retrieval distance. And we make the following additional data assumptions.

Assumption 2 (Data Assumption). *We assume the data follows the following:*

1. **PRETRAINING EXAMPLES** ($\mathcal{D}_{\text{pt}}$). *For a pretraining prompt of length $m+1$ and for all $i \in [m] \cup \{q\}$, we assume $\mathbf{x}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, $\beta_{\text{pt}} \sim \mathcal{N}(0, I)$.*
2. **TEST TIME EXAMPLES** ($\mathcal{D}_{\text{tt}}$). *For a test-time prompt of length $m+n+1$ and for all $i \in [m] \cup \{q\}$, we assume $\mathbf{x}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, $\beta_{\text{tt}} \sim \mathcal{N}(0, I)$.*
3. **TEST-TIME RAG EXAMPLES** ($\mathcal{D}_{\text{rag}}(\mathbf{x}_q)$). *For a test-time prompt of length $m+n+1$ and for all $i \in [m+1, \dots, m+n]$, we assume $\mathbf{x}_i^{\text{rag}} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, $\epsilon_i^{\text{rag}} \sim \mathcal{N}(0, \sigma_{\text{rag},i}^2)$, and the same β_{tt} as (2).*

Here, we assume the generic Gaussian property for the input, and isotropic Gaussian property for the noise and the weight vector, a common assumption made in ICL theory (Ahn et al., 2023; Gozeten et al., 2025) for simple yet meaningful analysis.

Overview of the Key Results

- (Uniform Noise) RAG examples are as effective as ICL examples in reducing the variance-induced err but ineffective at reducing the bias-induced err, causing a loss plateau for $n \rightarrow \infty$.
- (Non-Uniform Noise) RAG could improve the variance-induced error up to a finite n at a cost of increasing bias-induced error.

Roadmap Under these assumptions and uniform retrieval noise, we will first derive the population loss of RAG, $\mathcal{L}_{\text{tt+rag}}(\mathbf{W})$, for general $\mathbf{W}$ as in Theorem 1, analyze its finite sample complexity under the optimal pretrained weight $\mathbf{W}^*$ as in Proposition 1 and derive an optimal number of RAG examples of n^* for a given number of ICL examples m

as in Proposition 2. These discussions leads to our first key result. Then, under the non-uniform retrieval noise, we will prove the sample complexity under the distance-proportional noise (Theorem 2) and distance-weighted mixture noise (Theorem 3), and obtain our second key results above.

4.1 UNIFORM RETRIEVAL NOISE

Assumption 3 (Uniform Retrieval Noise). *The RAG noise ϵ_{rag} shares the same Gaussian distribution with variance σ_{rag}^2 , i.e. $\forall i \in [m+1, \dots, m+n]$, $\sigma_{\text{rag},i}^2 = \sigma_{\text{rag}}^2$.*

First, we present the assumption for uniform retrieval noise. In other words, all RAG examples are as helpful, and its improvement on the actual prediction is determined by the retrieval distance.

Theorem 1 (Population Loss for ICL with RAG Examples). *Under Assumption 1 2 3 $\Sigma = I$, the population loss of the linear self-attention predictor $\hat{y}_q = \mathbf{x}_q^\top \mathbf{W} \mathbf{X}^\top \mathbf{y}$ satisfies*

$$\mathcal{L}_{\text{tt+rag}}(\mathbf{W}) = \underbrace{\mathbb{E}(\hat{y}_q - \mathbb{E}(\hat{y}_q))^2}_{:=\text{err}_{\text{variance}}(\mathbf{W})} + \underbrace{\mathbb{E}(\mathbb{E}(\hat{y}_q) - \mathbb{E}(y_q))^2}_{:=\text{err}_{\text{bias}}(\mathbf{W})} + \underbrace{\sigma^2}_{\text{irreducible noise}}, \text{ and specifically,} \quad (11)$$

$$\begin{aligned} \text{err}_{\text{variance}}(\mathbf{W}) &= [m\sigma^2 + (1 + \delta^2)n\sigma_{\text{rag}}^2] \text{tr}(\mathbf{W}^\top \mathbf{W}) + n\sigma_{\text{rag}}^2 \text{tr}(\mathbf{W}^2) + n\sigma_{\text{rag}}^2 \text{tr}(\mathbf{W})^2 \\ \text{err}_{\text{bias}}(\mathbf{W}) &= \beta_{\text{tt}}^\top \left[I - (n\delta^2 + 2n + m)(\mathbf{W} + \mathbf{W}^\top) - 2n \text{tr}(\mathbf{W})I + M_4 \right] \beta_{\text{tt}} \\ &= \beta_{\text{tt}}^\top \left[I - (n\delta^2 + 2n + m)(\mathbf{W} + \mathbf{W}^\top) - 2n \text{tr}(\mathbf{W})I \right. \\ &\quad + [n^2(2 + \delta^2) + n(m + \delta^2)] (\mathbf{W}^2 + (\mathbf{W}^2)^\top) + 2n(n + \delta^2) \mathbf{W} \mathbf{W}^\top \\ &\quad + [m^2 + m + mn(2 + 2\delta^2) + n^2(2 + 2\delta^2 + \delta^4) + n(2\delta^2 + \delta^4)] \mathbf{W}^\top \mathbf{W} \\ &\quad + [n^2(2 + \delta^2) + n(m + \delta^2)] (\text{tr}(\mathbf{W}) (\mathbf{W} + \mathbf{W}^\top)) \\ &\quad \left. + [n^2 + n\delta^2] (\text{tr}(\mathbf{W})^2 + \text{tr}(\mathbf{W}^2)) I + [m + n^2 + n(2\delta^2 + \delta^4)] \text{tr}(\mathbf{W}^\top \mathbf{W}) I \right] \beta_{\text{tt}}. \end{aligned}$$

Here, we derive the exact bias-variance decomposition for ICL with RAG. The first line is the variance-induced error formed by a weighted sum of noise from ICL examples and RAG examples. Because of the implicit scaling of $\mathbf{W}$ as discussed in Lemma 1, the second order term in $\mathbf{W}$ will introduce an additional weight scaling of $\frac{m^2}{(m+n)^2}$ when adapting from the weight learned on m size context to $m+n$ size context. Thus, larger n will let $\text{err}_{\text{variance}}(\mathbf{W}) \rightarrow 0$, and the convergence rate is affected by δ^2 . Larger retrieval distance leads to a slower convergence. The bias-induced error is composed of all possible monomials of $\mathbf{W}$ up to the 2nd-order with tr operation. The complex dependency on m, n, δ^2, d requires additional assumptions on $\mathbf{W}$ to further interpret. As a sanity check, when $n = 0$ (ICL-only), this decomposition can exactly recover loss as in Lemma B.2 in (Gozeten et al., 2025).

As a proof sketch, we first compute $\text{err}_{\text{variance}}(\mathbf{W}) = \mathbb{E}(\mathbf{x}_q^\top \mathbf{W} \mathbf{X}^\top \epsilon)^2$ by splitting the calculation for ICL and RAG examples based on $\mathbf{X}$. Then, we compute $\text{err}_{\text{bias}}(\mathbf{W}) = \mathbb{E}[(\mathbf{x}_q^\top (I - \mathbf{W} \mathbf{X}^\top \mathbf{X}) \beta_{\text{tt}})^2]$. The main technical challenge lies in the dependency of $\mathbf{X}_{\text{rag}}$ on $\mathbf{x}_q$, and err_{bias} has a 6th-order dependency on $\mathbf{x}_q$ (2 from $\mathbf{x}_q$ and 4 from $\mathbf{X}$). As shown in Lemma 3, $\mathbb{E}[\mathbf{x}_q \mathbf{x}_q^\top A \mathbf{x}_q \mathbf{x}_q^\top B \mathbf{x}_q \mathbf{x}_q^\top]$ gives 15 new terms that include all the second order monomials of $\mathbf{W}$ with tr . The calculation requires multiple careful applications of Isserlis' theorem (Isserlis, 1918), and the full proof can be seen in Section B. Also, see Theorem 4 in the appendix for the full theorem with generic Σ .

Here, we present the finite sample bound for pretrained $\mathbf{W}^*$ for better interpretation.

Proposition 1 (Finite Sample Generalization Bound). *Under Assumption 1 2 3 if $\delta^2 \ll 1$, $\Sigma = I$,*

$$\begin{aligned} \mathcal{L}_{\text{tt+rag}}(\mathbf{W}^*) &= \mathcal{O}_{m,n} \left(\underbrace{\sigma^2 + \frac{dm}{(m+n)^2} \sigma^2 + \frac{d^2 n}{(m+n)^2} \sigma_{\text{rag}}^2}_{\text{err}_{\text{variance}}(\mathbf{W}^*)} + \underbrace{\|\beta_{\text{tt}}\|_2^2 \left[\frac{d}{m} + d^2 \left(\frac{n}{m+n} \right)^2 \right]}_{\text{err}_{\text{bias}}(\mathbf{W}^*)} \right) \\ \text{err}_{\text{variance}}(\mathbf{W}^*) &= \begin{cases} \mathcal{O}_m(\frac{d}{m} \sigma^2 + \frac{d^2}{m^2} \sigma_{\text{rag}}^2) = \mathcal{O}_m(\frac{1}{m}) & m \rightarrow \infty, n \text{ fixed.} \\ \mathcal{O}_n(\frac{d}{n^2} \sigma^2 + \frac{d^2}{n} \sigma_{\text{rag}}^2) = \mathcal{O}_n(\frac{1}{n}) & n \rightarrow \infty, m \text{ fixed} \\ \mathcal{O}_m(\frac{d}{m} \sigma^2 + \frac{d^2}{m} \sigma_{\text{rag}}^2) = \mathcal{O}_m(\frac{1}{m}) & m \rightarrow \infty, n = \Theta_m(m) \end{cases} \quad (12) \end{aligned}$$

$$\text{err}_{\text{bias}}(\mathbf{W}^*) = \begin{cases} \mathcal{O}_m(\|\beta_{\text{tt}}\|_2^2 \frac{d}{m}) & \text{if } m \rightarrow \infty, n \text{ is fixed} \\ \mathcal{O}_n(\|\beta_{\text{tt}}\|_2^2 d^2) = C_1 & \text{if } n \rightarrow \infty, m \text{ is fixed} \\ \mathcal{O}_m(\|\beta_{\text{tt}}\|_2^2 (\frac{d}{m} + d^2)) = C_2 + \mathcal{O}_m(\|\beta_{\text{tt}}\|_2^2 \frac{d}{m}) & \text{if } m \rightarrow \infty, n = \Theta_m(m) \end{cases}. \quad (13)$$

Here, we assume $\delta^2 \ll 1$ as the test time example x_i has only a variance of I , and it is unrealistic to assume a higher retrieval variance than the input variance. On the limit case where $m \rightarrow \infty$ and n are fixed, we observe that both variance-induced and bias-induced error decay at a rate of $\mathcal{O}(1/m)$, matching the results from the existing paper (Ahn et al., 2023; Zhang et al., 2024). When $n \rightarrow \infty$, the variance-induced error decays as $\mathcal{O}(1/n)$ matching the $\mathcal{O}(1/m)$ rate. However, introducing the RAG is ineffective at reducing the bias-induced error. Even when $m \rightarrow \infty$, increasing n will cause a loss plateau.

This effect can be explained by the underlying adaptive ability of transformers. In an online learning setup, we could always use the mean of the queried data as the prediction. However, in the LSA setup, the pretrained $\mathbf{W}^*$ serves as a proxy for $\mathbb{E}^{-1}(\mathbf{X}^\top \mathbf{X})$. In order to retain the adaptivity to the entire distribution of β_{tt} , we cannot use the optimal linear classifier $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ or use the mean of the retrieved examples ad hoc. At the test stage, $\mathbf{X}_{\text{rag}}$ only appears in $\mathbf{X}^\top \mathbf{y}$ and not in $\mathbf{W}^*$. The difference between $\mathcal{D}_{\text{rag}}(x_q)$ and $\mathcal{D}_{\text{tt}}$ directly leads to the increase of variance worsened by the increase of n . See full proof in Section B. Now, a natural question is whether we can find a balance of variance and bias and obtain an optimal RAG example size n^* .

Proposition 2. Under Assumption 1, 2, 3, $\delta^2 \ll 1$, $\Sigma = I$, and reasonable choice of $\sigma^2, \sigma_{\text{rag}}^2$ ($\sigma^2, \sigma_{\text{rag}}^2 \ll \|\beta_{\text{tt}}\|_2^2$), the optimal n^* that minimizes the RAG loss follows:

$$n^* = \mathcal{O}_m \left(\frac{m(d^2 \|\beta_{\text{tt}}\|_2^2 + d\sigma^2 - d^2 \sigma_{\text{rag}}^2)}{md^2 \|\beta_{\text{tt}}\|_2^2 - d^2 \sigma_{\text{rag}}^2} \right) = \mathcal{O}_m \left(\frac{d \|\beta_{\text{tt}}\|_2^2 + \sigma^2 - d \sigma_{\text{rag}}^2}{d \|\beta_{\text{tt}}\|_2^2} \right) \quad (14)$$

and the improvement on loss from picking the optimal n^* over $n = 0$ is given as:

$$\mathcal{L}_{\text{tt+rag}}(\mathbf{W}^*)|_{n=0} - \mathcal{L}_{\text{tt+rag}}(\mathbf{W}^*)|_{n=n^*} = \mathcal{O}_m \left(\frac{1}{m^2} \right). \quad (15)$$

In fact, the optimal n^* does not scale with m omitting the lower-order terms. Note that for $\|\beta_{\text{tt}}\|_2^2 = \mathcal{O}_m(1)$, $\|\beta_{\text{tt}}\|_2^2$ will dominate the numerator for reasonable choices of σ^2 and σ_{rag}^2 . A larger ICL noise σ^2 leads to a larger n^* , i.e. requiring more RAG examples to compensate for the loss. A larger RAG noise σ_{rag}^2 leads to a smaller n^* , i.e. less efficiency on RAG examples. And the improvement converges at $\mathcal{O}(\frac{1}{m^2})$, diminishing for large m . See the full proof in Section B. Several empirical works also observe a performance drop when increasing the number of retrieved examples (Wang et al., 2024a; Levy et al., 2025).

4.2 NON-UNIFORM RETRIEVAL NOISE

The uniform-noise setup in Section 4.1 relies on a retrieval pool of data with similar noise, so we could keep the variance σ_{rag}^2 fixed. In open-domain retrieval, this assumption could collapse: many retrieved examples could contain no answer or even a wrong answer. Empirically, people have observed that passages that are closer to the query vector x_q are more likely (Yang & Seo, 2020; Yoran et al., 2023; Lewis et al., 2020b) to contain the correct label. We want to theoretically investigate if the following hypothesis still holds:

Closer to query $x_q \implies$ more likely to contain correct answer.

4.2.1 DISTANCE-PROPORTIONAL NOISE (DPN)

We first investigate the scenario where the retrieval noise is proportional to the retrieval distance. Since the ICL analysis only applies to the mean-squared error loss, we study the effect of RAG under DPN on the correctness of the predictions.

Assumption 4 (Distance-Proportional Noise). There exists a constant $\gamma_1 > 0$ such that, for every retrieved sample i , $\sigma_{\text{rag},i}^2 = \gamma_1 \sigma^2 \delta_i^2$, i.e. the RAG noise variance grows linearly with the variance δ_i^2 that governs the retrieval distance.

Under the new data assumption, we denote the corresponding RAG loss, bias-induced error, and variance-induced error for $\mathbf{W}$ to be $\hat{\mathcal{L}}_{\text{tt+rag}}(\mathbf{W})$, $\hat{\text{err}}_{\text{bias}}(\mathbf{W})$, and $\hat{\text{err}}_{\text{variance}}(\mathbf{W})$.

Theorem 2 (Finite Sample RAG Generalization Bound under DPN). *Under Assumption [1][2] $\Sigma = I$, [4] the population loss is given as:*

$$\text{err}_{\text{variance}}(\mathbf{W}) = m\sigma^2 \text{tr}(\mathbf{W}^\top \mathbf{W}) + \sum_{i=1}^n \gamma_1 \delta_i^2 [(1 + \delta_i^2) \text{tr}(\mathbf{W}^\top \mathbf{W}) + \text{tr}(\mathbf{W}^2) + \text{tr}(\mathbf{W})^2].$$

If the variance of the retrieval distance follows power law, i.e. $\exists \gamma_2 > 0, q \geq 0$ s.t. $\delta_i^2 = \gamma_2 i^q$, then

$$\text{err}_{\text{bias}}(\mathbf{W}^*) = \mathcal{O}_{m,n} \left(\text{err}_{\text{bias}}(\mathbf{W}^*) + \|\beta_{\text{tt}}\|_2^2 \left[\frac{dn^{2q+1} + n^{2q+2}}{(m+n)^2} \right] \right) \quad (16)$$

and

$$\text{err}_{\text{variance}}(\mathbf{W}^*) = \mathcal{O}_n \left(\frac{dm\sigma^2 + d(n^{2q+1})\sigma^2}{(m+n)^2} \right) = \begin{cases} \mathcal{O}_n (dn^{2q-1}\sigma^2) & \text{if } n \rightarrow \infty, q \leq 1/2 \\ \text{diverges} & \text{if } n \rightarrow \infty, q > 1/2 \end{cases}. \quad (17)$$

Here, we derive the sample complexity under DPN. A second order dependency on δ_i^2 shows up in both the variance-induced and bias-induced error (exact form seen in Section [B]). Thus, the δ_i^2 -involved constant will dominate the other constants. Specifically, it even leads to divergence for $q > 1/2$ for the variance-induced error and $q > 0$ for the bias-induced error.

4.2.2 DISTANCE-WEIGHTED MIXTURE NOISE

In this section, we discuss the scenario where further RAG examples are less likely to contain the correct answers. We use a pair of large and small noises to model the correct/incorrect examples.

Assumption 5 (Distance-Weighted Mixture Noise). *We assume that the RAG noise is formed by a mixture of small and large noise:*

$$y(\mathbf{x}_{\text{rag}}) = \begin{cases} f(\mathbf{x}_{\text{rag},i}) + \epsilon_s & \text{w.p. } p_i \\ f(\mathbf{x}_{\text{rag},i}) + \epsilon_l & \text{w.p. } 1 - p_i \end{cases},$$

where $\epsilon_s \sim \mathcal{N}(0, c_s \sigma^2)$ corresponds to the small noise and $\epsilon_l \sim \mathcal{N}(0, c_l \sigma^2)$ corresponds to the large noise, with $c_l \geq c_s \geq 0$. The probability of sampling small noise p_i follows an inverse power law of the variance of the retrieval distance, i.e. $p_i = (1 + \delta_i^2)^{-\tilde{q}}$, $\tilde{q} \geq 0$.

Here, we choose the sampling probability (of small noise) p_i to follow a polynomial decay and the constant 1 here is to ensure $p_i = 0$ when $\delta_i^2 = 0$. Under the new data assumption, we denote the corresponding RAG loss, bias-induced error, and variance-induced error for $\mathbf{W}$ to be $\tilde{\mathcal{L}}_{\text{tt+rag}}(\mathbf{W})$, $\tilde{\text{err}}_{\text{bias}}(\mathbf{W})$, and $\tilde{\text{err}}_{\text{variance}}(\mathbf{W})$.

Theorem 3 (Finite Sample RAG Bound under Distance-Weighted Mixture Noise). *Under Assumption [1][2][5] $\Sigma = I$, then $\tilde{\text{err}}_{\text{bias}}(\mathbf{W}) = \text{err}_{\text{bias}}(\mathbf{W})$, and*

$$\tilde{\text{err}}_{\text{variance}}(\mathbf{W}) = m\sigma^2 \text{tr}(\mathbf{W}^\top \mathbf{W}) + \sum_{i=1}^n (p_i \sigma_s^2 + (1 - p_i) \sigma_l^2) [(1 + \delta_i^2) \text{tr}(\mathbf{W}^\top \mathbf{W}) + \text{tr}(\mathbf{W}^2) + \text{tr}(\mathbf{W})^2].$$

If the variance of the retrieval distance follows power law, i.e. $\exists \gamma_2 > 0, q \geq 0$ s.t. $\delta_i^2 = \gamma_2 i^q$, then:

$$\tilde{\text{err}}_{\text{variance}}(\mathbf{W}^*) = \begin{cases} \mathcal{O}_n (c_l dn^{q-1} \sigma^2 - (c_l - c_s) \sigma^2 dn^{q-1-q\tilde{q}}) & \text{if } n \rightarrow \infty, q \leq 1 \\ \text{diverges} & \text{if } n \rightarrow \infty, q > 1 \end{cases}. \quad (18)$$

The bias-induced error here is the same as in DPN, since we assume a polynomial dependency for δ_i^2 on i in both settings and the bias-induced error is independent of the variance of noise. Even though the variance of small/large noise is bounded, the dependency on the retrieval distance leads to the divergence at large q ($q > 1$). The large prediction noise will dominate the variance-induced error, but a larger gap between large and small noise ($c_l - c_s$) can mitigate the error by a ratio of $\mathcal{O}_n(n^{-q\tilde{q}})$. That is, the smaller q and $\tilde{q}$ are, the lower the error.

We note that the uniform noise scenario can also admit the mixture noise model by taking a constant p_i , $\forall i$, resulting in a form similar to the standard uniform retrieval noise in Proposition [I].

5 EXPERIMENTS

We investigate the effect of RAG with three guiding questions: (Q1) Whether RAG data outperform randomly sampled in-context examples? (Q2) What are the impacts of the RAG examples from training data and RAG passages from external corpora? (Q3) With a fixed budget, what is the effect of varying the ratio between the two types of RAG data? Our experiments provide the following findings: (A1) RAG data lead to better performance than in-context ones under different data budgets. (A2) Interestingly, the first few RAG training examples significantly improve performance, but later ones are harmful, because the first few are highly relevant but later ones are noise rather than signal. In contrast, RAG passages from external corpora can slowly but monotonically improve the performance, because external corpora are large enough to provide noisy but still relevant data. These are captured by different noise models in our theory. (A3) The performance is not monotonic with the ratio, and the sweet spot depends on the data/model.

We use representative models **ATLAS** [Izacard et al. (2022)] and **RAVEN** [Huang et al. (2024)] on two standard open-domain question answering benchmarks **Natural Questions (NQ)** [Kwiatkowski et al. (2019)] and **TriviaQA** [Joshi et al. (2017)]. For evaluation, the context consists of m in-context examples, and n RAG data points (including n_1 RAG examples from the training data and n_2 RAG passages from external corpora like Wikipedia, so $n = n_1 + n_2$). We choose different m, n_1, n_2 's for our study purpose and report the standard exact match (EM) accuracy on the test set. See Section C for further experiment setup details.

RAG v.s. In-Context For a budget c , we compare using RAG only ($m = 0, n_1 = n_2 = c/2$) and in-context examples only ($m = c, n_1 = n_2 = 0$). The results in Figure 1 show that RAG consistently outperforms in-context examples, as RAG provides query-relevant data with more signals to address the query, consistent with our analysis.

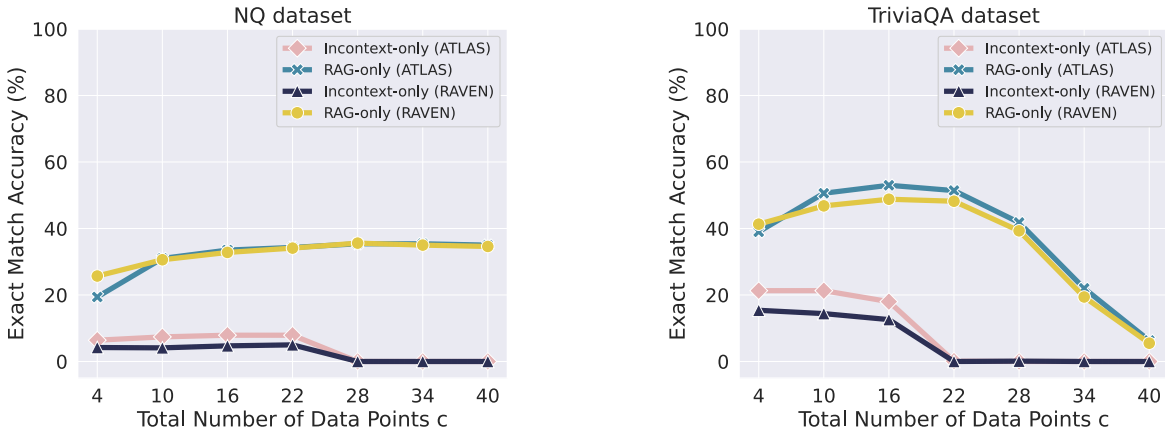


Figure 1: We compare performance between the RAG-only ($c = m$) versus in-context-only methods ($c = n_1 + n_2, n_1 = n_2$), where c is the total number of data, n_1 refers to retrieved examples and n_2 to passages.

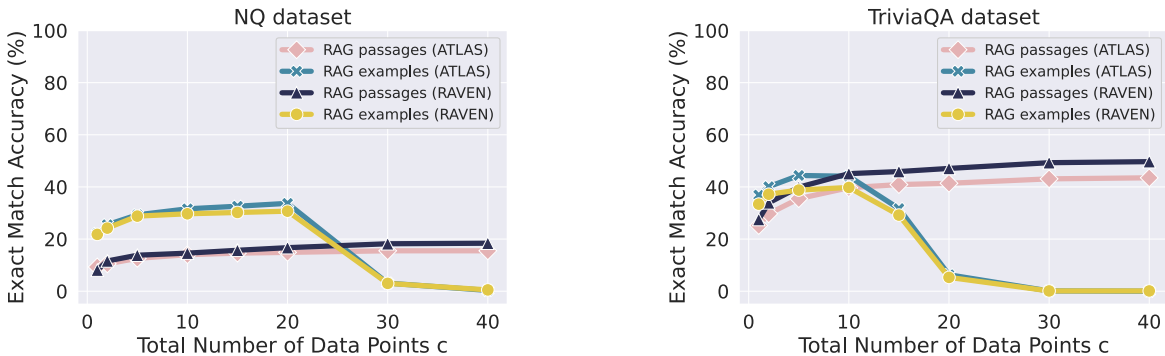
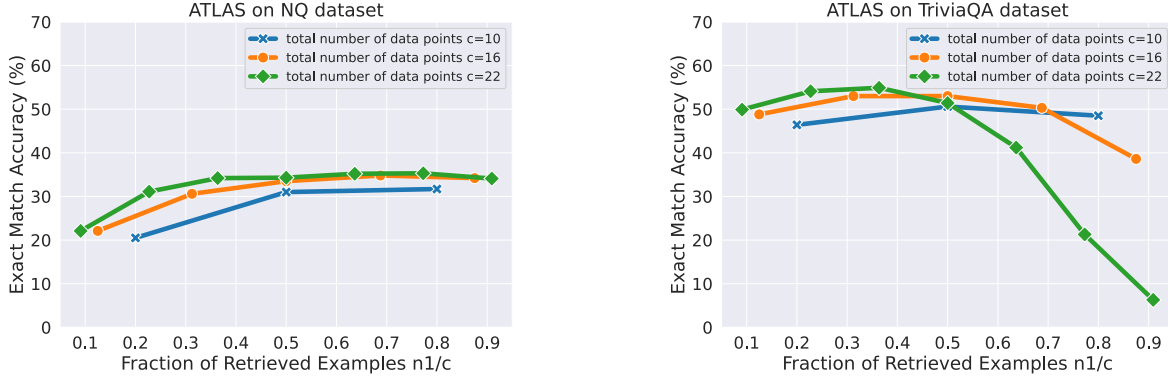


Figure 2: We compare the performance of RAG using examples ($c = n_1$) versus passages ($c = n_2$).



(a) ATLAS Performance as a function of n_1/c under different data points c on NQ.

(b) ATLAS Performance as a function of n_1/c under different data points c on TriviaQA.

Figure 3: Performance sensitivity to n_1/c under total data points c . n_1 refers to retrieved examples and n_2 to passages.

RAG Examples v.s. RAG External Passages Next, we compare using RAG examples from training data only ($m = 0, n_1 = c, n_2 = 0$) and RAG passages from external corpora only ($m = 0, n_1 = 0, n_2 = c$). The results in Figure 2 show interesting patterns. For RAG examples-only, with more examples, the performance first significantly improves but later drops. This suggests that the first few examples are highly relevant but later ones contain more noise than signal. In contrast, for RAG passages only, the performance increases more slowly but steadily for larger budgets. This suggests the passages retrieved are noisy but still have relevant signals. This aligns with our noise modeling. When n_1 is small (≤ 20 for NQ and ≤ 10 for TriviaQA), RAG examples resemble *uniform noise* due to the relevance of retrieved examples. As n_1 increases, n_1 introduces more irrelevant or conflicting examples (i.e., *non-uniform noise*). On the other hand, n_2 passages from external corpora resemble a *uniform noise* regime as the retrieval pool is broad with relevant data but also noisy.

When the retrieval budget is small, retrieval from training examples yields higher accuracy than from passages, even though both operate in the uniform-noise regime. This discrepancy follows from the mixture-noise effects: a passage judged relevant may still lack any answer-bearing text, raising its effective noise level relative to examples. Furthermore, the significant drop for the retrieval from examples as opposed to retrieval from passages can be explained by the size difference for the training data and passages pool (i.e. Wikipedia). Since the passages provide a denser coverage of the semantic space, more passages will remain relevant as opposed to examples. In all, our theory is consistent with both practical data types and matches the empirical results.

Ratio between RAG Examples and Passages The different noise properties of the two kinds of RAG data imply that we should find a proper ratio between them when the total budget c is fixed. Figure 3 (Figure 4 for RAVEN in Section C) shows that as the ratio n_1/c increases, the performance initially improves—benefiting from signal information—but eventually declines as low-quality examples dominate the context. The results demonstrate that performance initially improves as more signal (examples) is added, but eventually declines due to increasing noise from low-quality examples. This supports the theoretical perspective of balancing signal and noise in retrieval-augmented inputs.

6 CONCLUSION AND LIMITATIONS

We take an initial step by modeling RAG as query-dependent noisy in-context learning and deriving finite-sample error bounds for linear regression that help isolate the roles of retrieval signal and noise; we also explore how these bounds behave under different noise regimes and a simple test-time-training scenario. Experiments on Natural Questions and TriviaQA with RAVEN and ATLAS show trends consistent with our theoretical predictions.

Regarding limitations, our work considers the case without RAG finetuning. In practice, RAG information could also be used for finetuning, but our current work just focuses on the more common practice of using it as in-context information. Furthermore, our bounds focus on the linear setting, opening avenues for future studies on nonlinear methods like kernels and neural networks. While our framework accounts for common RAG noise models, new models may be needed for other types of RAG data. A further direction is to combine RAG with test-time training, studying how on-the-fly adaptation affects both theoretical guarantees and empirical performance. Our experiments feature representative models and datasets, but future research can explore newer retrievers and more RAG applications.

REFERENCES

- Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2021.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Annual Meeting of the Association for Computational Linguistics*, 2024.
- Chen Cheng, Xinzhi Yu, Haodong Wen, Jingsong Sun, Guanzhang Yue, Yihao Zhang, and Zeming Wei. Exploring the robustness of in-context learning with noisy labels. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.
- Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*, 2023.
- Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27, 1967.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.
- Halil Alperen Gozeten, M Emrullah Ildiz, Xuechen Zhang, Mahdi Soltanolkotabi, Marco Mondelli, and Samet Oymak. Test-time training provably improves transformers as in-context learners. *arXiv preprint arXiv:2503.11842*, 2025.
- Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. Raven: In-context learning with retrieval-augmented encoder-decoder language models, 2024. URL <https://arxiv.org/abs/2308.07922>.
- Rongjie Huang, Jia-Bin Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiaoyue Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *ArXiv*, abs/2301.12661, 2023.
- Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

- Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *ArXiv*, abs/2007.01282, 2020.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022. URL <https://arxiv.org/abs/2208.03299>.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147/>.
- Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-rag: Certified generation risks for retrieval-augmented language models. *arXiv preprint arXiv:2402.03181*, 2024.
- Tom Kwiakowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL <https://aclanthology.org/Q19-1026/>.
- Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 2024.
- Shahar Levy, Nir Mazon, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. More documents, same length: Isolating the challenge of multiple documents in rag. *arXiv preprint arXiv:2503.04388*, 2025.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020b.
- Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. Making large language models a better foundation for dense retrieval. *ArXiv*, abs/2312.15503, 2023.
- Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. Self-prompting large language models for zero-shot open-domain qa. *arXiv preprint arXiv:2212.08635*, 2022.
- Xiaonan Li and Xipeng Qiu. Mot: Pre-thinking and recalling enable chatgpt to self-improve with memory-of-thoughts. *CoRR*, 2023.
- Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2024.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.
- Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*, 2024.
- Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. Z-icl: Zero-shot in-context learning with pseudo-demonstrations. *arXiv preprint arXiv:2212.09865*, 2022.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-embedding-mistral: enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024. URL <https://blog.salesforceairesearch.com/sfr-embedded-mistral/>.

- Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.
- Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7 (15):510, 2008.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331, 2023.
- Rita Parada Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2840–2849, 2022.
- Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, 2022.
- Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. *arXiv preprint arXiv:2210.13693*, 2022.
- Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*, 2024a.
- Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. Can in-context learning really generalize to out-of-distribution tasks? *arXiv preprint arXiv:2410.09695*, 2024b.
- Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.
- Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024a.
- Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. A theory for token-level harmonization in retrieval-augmented generation. *arXiv preprint arXiv:2406.00944*, 2024b.
- Sohee Yang and Minjoon Seo. Is retriever merely an approximator of reader? *arXiv preprint arXiv:2010.10999*, 2020.
- Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.
- Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.
- Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024a.
- Yufeng Zhao, Yoshihiro Sakai, and Naoya Inoue. Noisyicl: A little noise in model parameters calibrates in-context learning. *arXiv preprint arXiv:2402.05515*, 2024b.