C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models

Mintong Kang[1] **Nezihe Merve Gürel**[2] **Ning Yu**[3] **Dawn Song**[4] **Bo Li**[1 5]

Abstract

Despite the impressive capabilities of large language models (LLMs) across diverse applications, they still suffer from trustworthiness issues, such as hallucinations and misalignments. Retrieval-augmented language models (RAG) have been proposed to enhance the credibility of generations by grounding external knowledge, but the theoretical understandings of their generation risks remains unexplored. In this paper, we answer: *1) whether RAG can indeed lead to low generation risks, 2) how to provide provable guarantees on the generation risks of RAG and vanilla LLMs, and 3) what sufficient conditions enable RAG models to reduce generation risks.* We propose C-RAG, a novel framework to certify generation risks for RAG models. Specifically, we provide conformal risk analysis for RAG models and certify an upper confidence bound of generation risks, which we refer to as *conformal generation risk*. We also provide theoretical guarantees on conformal generation risks for general bounded risk functions under test distribution shifts. We prove that RAG achieves a lower conformal generation risk than that of a single LLM when the quality of the retrieval model and transformer is non-trivial. Our intensive empirical results demonstrate the soundness and tightness of our conformal generation risk guarantees across four widely-used NLP datasets on four state-of-the-art retrieval models.

1. Introduction

Large language models (LLMs) (Touvron et al., 2023; OpenAI et al., 2023) recently exhibit emergent abilities across different NLP tasks, such as text summarization, question answering, and machine translation. However, existing works (Wang et al., 2023a; Liang et al., 2022; Liu et al., 2023) show that the generations of LLMs can be unreliable, untrustworthy, and risky in many cases. Therefore, *certifiably controlling the generation risks of LLMs* becomes particularly important before the deployment of LLMs, especially in safety-critical domains.

Retrieval-augmented language models (RAG) (Lewis et al., 2020; Karpukhin et al., 2020; Xiong et al., 2020) have been proposed to enhance the credibility of LLMs by retrieving relevant documents from an external knowledge base and generating contents conditioned on the retrieved knowledge. RAG models are shown effective in mitigating generation risks via in-context learning from the retrieved documents (Brown et al., 2020). However, theoretical understandings of their generation risks still remain unexplored. In this work, we focus on this problem and ask:

Can RAG indeed lead to low generation risks? How can we provide provable guarantees on the generation risks of RAG and vanilla LLMs? What are the sufficient conditions that enable RAG models to reduce generation risks? Can we provably control the generation risks below a desired level?

To theoretically analyze the generation risks of RAG and answer the above questions, we propose C-RAG, a novel framework of certified generation risks for RAG models. We first propose a constrained generation protocol for RAG models to produce a controlled set of generations. The protocol operates based on specific parameter configurations, including the number of retrieved examples, the size of the generation set, and a similarity threshold for generation diversity. We then provide conformal analysis (Bates et al., 2021; Angelopoulos et al., 2021; 2022) for RAG models under the constrained generation protocol, aiming to provably control the generation risks based on test statistics from in-distribution calibration samples. To achieve this goal, we derive a high-probability upper bound of generation risks during inference time, which we call *conformal generation risk*. We show that (a) the conformal generation risk serves as a sound upper bound to the empirical generation risks given a RAG configuration in Prop. 1; (b) the generation risk can be certifiably controlled below a desired level by computing a valid set of RAG configurations via C-RAG

[1] University of Illinois at Urbana-Champaign, USA [2] Delft University of Technology, Netherlands [3] Netflix Eyeline Studios, USA [4] University of California, Berkeley, USA [5] University of Chicago, USA. Correspondence to: Mintong Kang <mintong2@illinois.edu>, Bo Li <lbo@illinois.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

in Prop. 2; (c) the conformal analysis can be extended to more complex scenarios under test-time distribution shifts in Thm. 2, which presents the *first* generation risk guarantee under test-time distribution shifts for general bounded risk functions. Based on our conformal analysis for the generation risks of RAG and vanilla LLMs, we prove that (a) the conformal generation risk of RAG is lower than that of the corresponding vanilla LLM in Thm. 1; (b) under bounded test-time distribution shifts, RAG also lowers the conformal generation risks compared to the vanilla LLM in Thm. 3.

We evaluate the conformal generation risk guarantees of C-RAG with different retrieval models on four widely-used datasets. For all retrieval methods and datasets, we empirically validate that our conformal generation risk guarantees are sound and tight even with distribution shifts, as they upper bound the empirical generation risks observed on random test sets while maintaining only a minimal gap, narrowing down to the scale of $1e-3$. We empirically show that RAG consistently achieves a lower conformal generation risk than a single LLM without retrieval, which is consistent with our theoretical findings in Secs. 5 and 6. We also evaluate the conformal generation risk for different SOTA retrieval models, such as sparse encoding metrics BM25 (Robertson et al., 2009), text-embedding-ada-002 model from OpenAI, bge model from BAAI (Zhang et al., 2023a), and supervised fine-tuned embedding model (Wang et al., 2023c) to validate our analysis on retrieval quality. We show that among these models, text-embedding-ada-002 and supervised fine-tuned embedding models outperform other baselines in achieving low conformal generation risks.

2. Related work

Retrieval augmented generation (RAG) is a framework for improving the generation quality of LLMs via retrieving relevant information from an external knowledge base and grounding the model on the retrieved knowledge for conditional generations. SOTA retrieval methods (Lewis et al., 2020; Karpukhin et al., 2020; Xiong et al., 2020) employ dual encoders to project both query and candidate texts into the embedding space and retrieve candidate texts that exhibit high similarity to the embedded query text. Although RAG is demonstrated to be effective in enhancing the generation credibility, their theoretical understanding is limited. Basu et al. conduct retrieval analysis for a constrained function and data class from a statistical perspective, but the results cannot be applicable to self-attention transformers and to arbitrary data distribution. In this work, we provide the first theoretical analysis of how RAG leads to low generation risks in self-attention transformers for arbitrary data distribution.

Conformal prediction is a statistical technique used to create prediction sets with assured coverage. (Vovk et al., 1999; 2005; Lei et al., 2013; Yang & Kuchibhotla, 2021). Broadly, conformal risk control methods (Bates et al., 2021; Angelopoulos et al., 2021; 2022; Quach et al., 2023) provide a high-confidence risk guarantee for black-box risk functions, assuming data exchangeability. However, the risk guarantee can be broken under test-time distribution shifts. While Angelopoulos et al. and Farinhas et al. offer a valid conformal risk guarantee for monotonic risk functions under distribution shifts, the monotonicity assumption is not always practical. In this work, we introduce a generally applicable conformal risk guarantee for general bounded risk functions, under distribution shifts at test time.

3. Preliminaries

Before introducing C-RAG, we first review the preliminaries of conformal controlling methods (Bates et al., 2021; Angelopoulos et al., 2021; 2022), which calibrate machine learning models to ensure their predictions meet explicit finite-sample statistical guarantees. Let $R(\lambda)$ and $\hat{R}(\lambda)$ denote the population risk and empirical risk, respectively, where $\lambda \in \Lambda$ is a parameter that induces the risk. We consider a finite parameter space Λ with N parameters, $\lambda_1, \dots, \lambda_N$. For a desired risk level α $(0 < \alpha < 1)$, we define N null hypotheses: $\mathcal{H}_j : R(\lambda_j) > \alpha$ $(j \in 1, \dots, N)$. Let p_j be the p-value of the null hypothesis $\mathcal{H}_j$. Given n calibration samples, Bates et al.; Angelopoulos et al. provide valid p-values as follows:

$$p_j = \min\Big\{\exp\Big\{-nh(\hat{R}(\lambda_j), \alpha)\Big\}, e\mathbb{P}\Big[\mathrm{Bin}(n,\alpha) \le \lceil n\hat{R}(\lambda_j)\rceil\Big]\Big\}, \tag{1}$$

where $h(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$. Let J be the index set of false null hypothesis. An algorithm $\mathcal{A} : [0,1]^N \mapsto 2^{1,\dots,N}$ is a family-wise error rate (FWER)-controlling algorithm at level δ if $\mathbb{P}\left[\mathcal{A}(p_1, \dots, p_N) \subseteq J\right] \ge 1 - \delta$. They finally show that:

$$\mathbb{P}\left[\sup_{\lambda \in \hat{\Lambda}} \{R(\lambda)\} \le \alpha\right] \ge 1 - \delta, \tag{2}$$

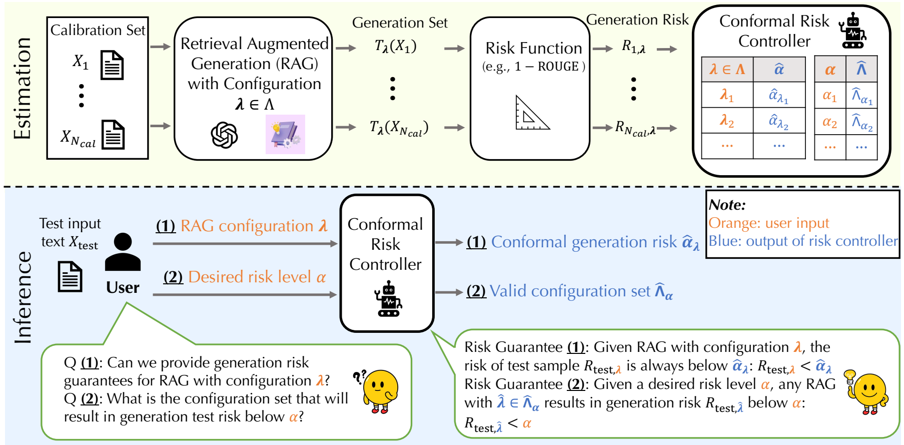
where $\hat{\Lambda}$ denotes the output of the FWER algorithm $\mathcal{A}$.

4. Conformal generation risks of RAG models

We introduce the problem setup in Sec. 4.1, our constrained generation protocol for RAG models in Sec. 4.2, and the conformal generation risks in Sec. 4.3. We prove that (1) Given a RAG configuration, C-RAG provides a high-probability generation risk upper bound in Prop. 1, and (2) Given a desired risk level α, C-RAG offers a set of configurations that can provably maintain the risk below α in Prop. 2.

4.1. Problem setup

For a pretrained language model (LM) and any user input text, we aim to provide rigorous guarantees for the genera-


Estimation
Calibration Set
X_1
$X_{N_{cal}}$
Retrieval Augmented Generation (RAG) with Configuration $\boldsymbol{\lambda} \in \Lambda$
Generation Set
$T_{\lambda}(X_1)$
$T_{\lambda}(X_{N_{cal}})$
Risk Function (e.g., 1 − ROUGE)
Generation Risk
$R_{1,\lambda}$
$R_{N_{cal},\lambda}$
Conformal Risk Controller
Inference
Test input text X_{test}
User
(1) RAG configuration $\boldsymbol{\lambda}$
(2) Desired risk level α
Conformal Risk Controller
(1) Conformal generation risk $\hat{\boldsymbol{\alpha}}_{\lambda}$
(2) Valid configuration set $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\alpha}}$
Note:
Orange: user input
Blue: output of risk controller

$\lambda \in \Lambda$	$\hat{\alpha}$
λ_1	$\hat{\alpha}_{\lambda_1}$
λ_2	$\hat{\alpha}_{\lambda_2}$
…	…

α	$\hat{\Lambda}$
α_1	$\hat{\Lambda}_{\alpha_1}$
α_2	$\hat{\Lambda}_{\alpha_2}$
…	…

Q (1): Can we provide generation risk guarantees for RAG with configuration λ? Q (2): What is the configuration set that will result in generation test risk below α?

Risk Guarantee (1): Given RAG with configuration λ, the risk of test sample $R_{\text{test},\lambda}$ is always below $\hat{\alpha}_{\lambda}$: $R_{\text{test},\lambda} < \hat{\alpha}_{\lambda}$ Risk Guarantee (2): Given a desired risk level α, any RAG with $\hat{\lambda} \in \hat{\Lambda}_{\alpha}$ results in generation risk $R_{\text{test},\hat{\lambda}}$ below α: $R_{\text{test},\hat{\lambda}} < \alpha$

Figure 1: Overview of C−RAG. In the estimation stage (upper row), the conformal risk controller computes conformal generation risks for different RAG configurations (Prop. 1), and valid configuration sets for different risk levels (Prop. 2), both based on risk statistics on the calibration set. In the inference stage (lower row), for any configuration $\boldsymbol{\lambda}$ and any desired risk level α provided by users, the conformal risk controller outputs the conformal generation risk $\hat{\alpha}_{\boldsymbol{\lambda}}$ with Risk Guarantee (1) and the configuration set $\hat{\Lambda}_{\alpha}$ with Risk Guarantee (2).

tion risks (e.g., $1 - \text{ROUGE}$). To achieve this, we develop a constrained generation protocol for RAG models. The generation protocol is governed by adjustable parameter configurations (e.g., number of retrieved examples, size of generations), which allow for more controlled RAG generations to achieve a desired risk level.

Formally, we let $\mathcal{V}$ be the vocabulary set, n_I be the maximal length of input text and n_O be the maximal length of output text. Let $\mathcal{X} := \mathcal{V}^{n_I}$ be the input text space, and $\mathcal{Y} := \mathcal{V}^{n_O}$ be the output text space. We notate $p_{\theta_l}(y|x)$ $(x \in \mathcal{X}, y \in \mathcal{Y})$ as the probability distribution of output text y given input text x, estimated by a pretrained LM parameterized by θ_l. Consider a RAG generation protocol $T_{\boldsymbol{\lambda}, p_{\theta_l}} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ with LM θ_l and parameter configuration $\boldsymbol{\lambda} \in \Lambda = \mathbb{R}^B$, where B is the maximal number of parameters to control the generation procedure. To evaluate the quality of the generation under $T_{\boldsymbol{\lambda}, p_{\theta_l}}(x)$ given input x, we define a risk function $R(T_{\boldsymbol{\lambda}, p_{\theta_l}}(x), y) : 2^{\mathcal{Y}} \times \mathcal{Y} \mapsto [0, 1]$, where y is the reference text of x. For text generation tasks, a typical selection of the risk function could be $1 - \max_{y' \in T_{\boldsymbol{\lambda}, p_{\theta_l}}(x)} \text{ROUGE}(y', y)$, where ROUGE measures the matching score between the generation y' and reference text y. Notably, our generation protocol outputs a set of generations instead of just one, allowing us to explore better generations and adjust the generation set size through a parameter for risk control.

4.2. Constrained generation protocol for RAG models

RAG models (Wang et al., 2023c; Rubin et al., 2021; Huang et al., 2023) combine a retrieval model and a generation LM. The retrieval model retrieves N_{rag} relevant examples to the query from an external knowledge base, and the LM learns in-context from these examples. The knowledge base contains N_{ext} samples in $\hat{\mathcal{D}}_{\text{ext}} = \{(X_i, Y_i)\}_{i=1}^{N_{\text{ext}}}$. The retrieval model uses an encoder to map instances into an embedding space, and then identifies the relevant examples to the query X_{test} based on similarity. This similarity, defined by $s_{\theta_r}(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and parameterized by θ_r, is used to find the nearest examples using KNN search in the embedding space.

We arrange the retrieved N_{rag} in-context examples and the test example X_{test} into augmented input text $X^{(\text{rag})}$ using a template. We then sample the generation from $p_{\theta_l}(\cdot|X^{(\text{rag})})$ repeatedly until λ_g generations are collected. To control the diversity of generations, we reject those with a similarity higher than a threshold λ_s to the previous generations. In essence, the constrained generation protocol is controlled by configuration $\boldsymbol{\lambda} = [N_{\text{rag}}, \lambda_g, \lambda_s]$ and output a generation set $T_{\boldsymbol{\lambda}, p_{\theta_l}}(x)$ based on the configuration $\boldsymbol{\lambda}$ and input x. We refer to Alg. 1 in App. C.1 for the pseudocode of the protocol.

4.3. Conformal generation risks for RAG models

We certify generation risks of the RAG models with the constrained generation protocol $T_{\boldsymbol{\lambda}, p_{\theta_l}}$ via conformal risk analysis (Bates et al., 2021; Angelopoulos et al., 2022; 2021). Conformal analysis provably controls the generation risks based on test statistics from in-distribution calibration samples. In this work, we consider a calibration set $\hat{\mathcal{D}}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{N_{\text{cal}}}$ with size N_{cal}, and compute the empirical generation risk $\hat{R}(\hat{\mathcal{D}}_{\text{cal}}) = {}^{1}/_{N_{\text{cal}}} \sum_{(x,y) \in \hat{\mathcal{D}}_{\text{cal}}} R(T_{\boldsymbol{\lambda}, p_{\theta_l}}(x), y)$.

Risk Guarantees for RAG Models For an LM θ_l, calibration set $\hat{\mathcal{D}}_{\text{cal}}$, test sample $(X_{\text{test}}, Y_{\text{test}})$, generation protocol $T_{\boldsymbol{\lambda}, p_{\theta_l}}$ with configuration $\boldsymbol{\lambda}$ and confidence level $1 - \delta$ $(\delta \in [0, 1])$, C−RAG provides two types of generation risk guarantees for RAG models:

Proposition 1 (Risk Guarantee (1), adaptation of (Bates et al., 2021) to constrained RAG generation). *Given a configuration* $\boldsymbol{\lambda}$ *in generation protocol,* C-RAG *guarantees that:*

$$\mathbb{P}\left[R(T_{\boldsymbol{\lambda},p_{\theta_l}}(x),y) \leq \hat{\alpha}_{\boldsymbol{\lambda}}\right] \geq 1-\delta, \quad (3)$$

where the high-probability risk upper bound $\hat{\alpha}_{\boldsymbol{\lambda}}$*, the so-called* ***conformal generation risk****, is given by:*

$$\hat{\alpha} = \min\left\{h^{-1}\left(\frac{1/\delta}{N_{cal}}; \hat{R}(\hat{\mathcal{D}}_{cal})\right), \Phi_{bin}^{-1}\left(\frac{\delta}{e}; N_{cal}, \hat{R}(\hat{\mathcal{D}}_{cal})\right)\right\}$$

with $h^{-1}(\cdot;\cdot)$ *as the partial inverse* $h^{-1}(h(a,b);a) = b$ *of* $h(a,b) = a\log({}^{a}/_{b}) + (1-a)\log({}^{(1-a)}/_{(1-b)})$*, and* Φ_{bin}^{-1} *as the inverse of binomial cumulative distribution function (CDF).*

Proposition 2 (Risk Guarantee **(2)**, adaptation of (Angelopoulos et al., 2021) to constrained RAG generation)**.** *Given a desired risk level* α*,* C-RAG *computes a configuration set* $\hat{\Lambda}_\alpha$ *such that each configuration in* $\hat{\Lambda}_\alpha$ *is guaranteed to keep the generation risk below* α*. Namely,*

$$\mathbb{P}\left[\sup_{\hat{\boldsymbol{\lambda}}\in\hat{\Lambda}_\alpha}\left\{R\left(T_{\hat{\boldsymbol{\lambda}},p_{\theta_l}}(x),y\right)\right\} \leq \alpha\right] \geq 1-\delta, \quad (4)$$

where the valid configuration set $\hat{\Lambda}_\alpha$ *is given by family-wise error rate controlling algorithms such as Bonferroni correction:* $\hat{\Lambda}_\alpha = \{\hat{\boldsymbol{\lambda}}_j : p_j \leq \delta/|\Lambda|\}$ *where* p_j *is the p-value of the null hypothesis:* $\mathcal{H}_j : R(T_{\boldsymbol{\lambda},p_{\theta_l}}(x),y) > \alpha \ (j \in \{1,...,|\Lambda|\})$ *and can be computed by finite-sample valid bounds as shown in App. D.2.*

Connection between Risk Guarantees (1) and (2) Risk Guarantee **(1)** computes the conformal generation risk (risk upper bound) $\hat{\alpha}_{\boldsymbol{\lambda}}$ given a configuration $\boldsymbol{\lambda}$, while Risk Guarantee **(2)** computes a configuration set $\hat{\Lambda}_\alpha$ such that any configuration in the set results in a risk below the desired level α. Risk Guarantee **(2)** can be conceptualized as accepting configurations with generation risks statistically below α with a certain error rate (p-value), such that the union of error rates over parameter space is within the uncertainty budget δ. The Bonferroni correction in Prop. 2 adopts an even partition of the uncertainty budget, while we can have a dynamic partition algorithm based on graph search (see App. D.2). Therefore, Risk Guarantee **(1)** and **(2)** are connected by the duality between p-values and confidence intervals (Bates et al., 2021). We mainly focus on the conformal analysis of Risk Guarantee **(1)** in the following, and the results can be extrapolated to Risk Guarantee **(2)** directly. We defer the proofs of Props. 1 and 2 to App. D.

Advance of C-RAG compared to conformal controlling methods (Bates et al., 2021; Angelopoulos et al., 2021; 2022) Existing conformal risk analysis assumes that test and calibration samples come from the same distribution, which allows statistical risk predictions for test samples based on the calibration data. While the conformal generation risk bounds are previously studied (Angelopoulos et al., 2022; Farinhas et al., 2023b), the scope is limited to the monotonic risk functions. In this work, we extend the scope to provide the first conformal generation risk analysis under test-time distribution shifts for general bounded risk functions in Sec. 6. In addition, we propose a constrained RAG generation protocol for enhanced effectiveness and efficiency of risk controlling for LLM generations, as illustrated in Sec. 4. We also prove that RAG achieves a lower conformal generation risk than vanilla LLMs under scenarios with or without distribution shifts in Thms. 1 and 3.

5. Theoretical analysis of C-RAG

In this section, we prove that RAG model achieves a lower conformal generation risk compared to LLMs without retrievals and its benefits are correlated with the quality of the retrieval model and transformer. We provide the structure of our theoretical analysis and conclusions in Fig. 2.

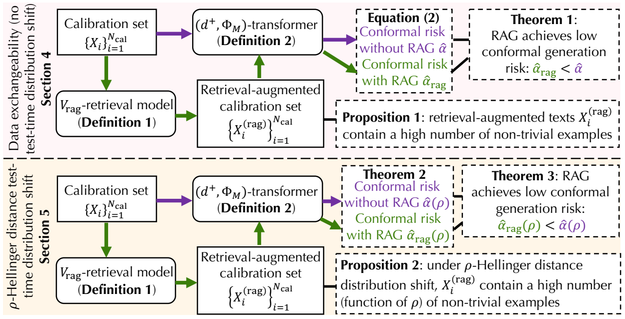

Data exchangeability (no test-time distribution shift)
Section 4
Calibration set $\{X_i\}_{i=1}^{N_{\text{cal}}}$
(d^+, Φ_M)-transformer (Definition 2)
Equation (2)
Conformal risk without RAG $\hat{\alpha}$
Conformal risk with RAG $\hat{\alpha}_{\text{rag}}$
Theorem 1: RAG achieves low conformal generation risk: $\hat{\alpha}_{\text{rag}} < \hat{\alpha}$
V_{rag}-retrieval model (Definition 1)
Retrieval-augmented calibration set $\left\{X_i^{(\text{rag})}\right\}_{i=1}^{N_{\text{cal}}}$
Proposition 1: retrieval-augmented texts $X_i^{(\text{rag})}$ contain a high number of non-trivial examples
ρ-Hellinger distance test-time distribution shift
Section 5
Calibration set $\{X_i\}_{i=1}^{N_{\text{cal}}}$
(d^+, Φ_M)-transformer (Definition 2)
Theorem 2
Conformal risk without RAG $\hat{\alpha}(\rho)$
Conformal risk with RAG $\hat{\alpha}_{\text{rag}}(\rho)$
Theorem 3: RAG achieves low conformal generation risk: $\hat{\alpha}_{\text{rag}}(\rho) < \hat{\alpha}(\rho)$
V_{rag}-retrieval model (Definition 1)
Retrieval-augmented calibration set $\left\{X_i^{(\text{rag})}\right\}_{i=1}^{N_{\text{cal}}}$
Proposition 2: under ρ-Hellinger distance distribution shift, $X_i^{(\text{rag})}$ contain a high number (function of ρ) of non-trivial examples

Figure 2: Certification framework of C-RAG. We provide theoretical results with the data exchangeability assumption in Sec. 5 (upper row) and extend the results to more complex scenarios under test-time distribution shifts in Sec. 6 (lower row).

5.1. Analysis setup

For our analysis, paralleling the previous transformer studies by (Von Oswald et al., 2023; Zhang et al., 2023b; Han et al., 2023), we consider a one-layer self-attention transformer parameterized with the embedding matrix $W_E : \mathcal{V} \mapsto \mathbb{R}^{d_1}$, query matrix $W_Q : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$, key matrix $W_K : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$, value matrix $W_V : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$, projection matrix $W_P : \mathbb{R}^{d_2} \mapsto \Delta^{|\mathcal{V}|}$, and treat each instance approximately as a single token. The retrieval-augmented input text then consists of N_{rag} retrieved examples and 1 query example. We denote the augmented input text by $\boldsymbol{q} \in \mathcal{V}^{N_{\text{rag}}+1}$. We categorize pairs of queries $\boldsymbol{q}_i, \boldsymbol{q}_j$ $(i, j \in [N_{\text{rag}}+1])$ as positive if they convey identical semantic meanings, indicated by $g(\boldsymbol{q}_i) = g(\boldsymbol{q}_j)$. In this context, $\boldsymbol{q}_i$ can be referred to as a **positive example** of $\boldsymbol{q}_j$. Conversely, pairs $\boldsymbol{q}_i, \boldsymbol{q}_j$ $(i, j \in [N_{\text{rag}}+1])$ are considered negative if they are semantically different. We use such a definition for clear interpretation of our findings, but our analysis can be extended to include broader definitions of positive pairs, as addressed in the remarks of Thm. 1.

]

Following the single-layer transformer formulation in (Von Oswald et al., 2023), given an input text $\boldsymbol{q}$, we consider the single-token output $O^{(\text{rag})}(\boldsymbol{q}) \in \Delta^{|\mathcal{V}|}$ corresponding to the query example $\boldsymbol{q}_{N_{\text{rag}+1}}$ at the last position, formulated as:

$$O^{(\text{rag})}(\boldsymbol{q}) = \sigma\big(W_P W_V \{ \underbrace{W_E \boldsymbol{q}_{N_{\text{rag}+1}}}_{\text{residual}} + \\ (W_E \boldsymbol{q}) \underbrace{\sigma((W_K W_E \boldsymbol{q})^T (W_Q W_E \boldsymbol{q}_{N_{\text{rag}+1}}))}_{\text{attention scores to } \boldsymbol{q}_{N_{\text{rag}+1}}} \}\big), \tag{5}$$

where $\sigma(\cdot)$ is the Softmax function. Note that without RAG, the output probability vector $O(\boldsymbol{q})$ is formulated as $O(\boldsymbol{q}) = \sigma\left(W_P W_V W_E \boldsymbol{q}_{N_{\text{rag}+1}}\right)$.

5.2. Retrieval quality analysis

To quantify the quality of retrieval models, we introduce the concept of V_{rag}-retrieval model, where V_{rag} measures the variance of the contrastive loss of the retrieval model. A small V_{rag} implies a well-trained low-variance retrieval model and can be theoretically linked to the retrieval quality, which is measured by the number of retrieved positive examples with respect to the query text.

Definition 1 (V_{rag}-retrieval model)**.** Consider a retrieval model with similarity measurement $s_{\theta_r}(\cdot,\cdot)$ parameterized with θ_r and trained with contrastive loss $\mathcal{L}_{\text{cont}}$. Let x^+, x^- be positive and negative samples to sample x. Consider common contrastive loss $\mathcal{L}_{\text{cont}} = -\log\left(\sigma_{\text{sig}}(\exp\{s_\theta(x,x^-) - \exp\{s_\theta(x,x^+)\}\right)$, where $\sigma_{\text{sig}}(\cdot)$ is the sigmoid function. We define a V_{rag}-retrieval model as the retrieval model with (a) a non-trivial utility such that the expected contrastive loss L_τ is better than random: $L_\tau := \mathbb{E}[\mathcal{L}_{\text{cont}}] < \ln 2$ (i.e., $\mathbb{E}[s_\theta(x,x^+) - s_\theta(x,x^-)] > 0$); and (b) bounded variance such that the training is stable and converges well: $V_{\text{rag}} := \mathbb{V}[s_\theta(x,x^+) - s_\theta(x,x^-)]^{1/2}\log(\exp\{L_\tau\} - 1) < 1$

Remarks. (R1) Note that a retrieval model with random initialization can achieve $\mathbb{E}[s_\theta(x,x^+)] = \mathbb{E}[s_\theta(x,x^-)]$ asymptotically. We merely assume a V_{rag}-retrieval model that can non-trivially differentiate the positive from negative examples. (R2) We also assume a moderate stability with bounded variance, which implicitly assumes a moderate generalization of the retrieval model based on the variance-generalization link (Lam, 2016; Gotoh et al., 2018; Namkoong & Duchi, 2017). This is essential for the analysis as the knowledge base distribution is non-identical to the calibration/test distribution. (R3) We define V_{rag}-retrieval model using a standard contrastive loss in (Wang et al., 2023c; Rubin et al., 2021), but it can be adapted for other contrastive loss such as triplet loss (Hermans et al., 2017), by altering the logarithmic factor in the V_{rag} formula. We defer the proof sketches as well as the detailed proofs and remarks to App. F.

With a V_{rag}-retrieval model, we show that C-RAG can retrieve a high number of positive examples as in-context demonstrations as follows.

Proposition 3 (lower bound of the number of retrieved positive examples)**.** *Consider the V_{rag}-retrieval model in Def. 1 and RAG generation protocol in Sec. 4.2. Let $r_{cal}^{(c)}$ and $r_{ext}^{(c)}$ be the portion of data with groundtruth output $c \in \mathcal{Y}$ in the calibration data distribution and external knowledge data distribution, respectively. We have:*

$$\mathbb{E}\left[N_{pos}\right] \geq \frac{9}{10} N_{rag} \Bigg(1 - \sum_{c \in \mathcal{Y}} r_{cal}^{(c)} \Big(N_{ext} - r_{ext}^{(c)} N_{ext} \\ + \sqrt{2 \ln 10}\Big) V_{rag}^{0.5\left(r_{ext}^{(c)} N_{ext} - \sqrt{2\ln 10}\right)}\Bigg) \tag{6}$$

where N_{pos} is the number of retrieved positive examples, N_{rag} is the total number of retrieved examples, and N_{ext} is the number of examples in the external knowledge base.

Remarks. Prop. 3 offers a guarantee on the minimum number of positive examples retrieved by the V_{rag}-retrieval model. (R1) The ratio of the retrieved examples that are positive increases at an exponential rate with respect to N_{ext}, which suggests that expanding the external knowledge base could enhance the retrieval quality and therefore benefit in-context learning of LLMs, as shown in (Min et al., 2022; Wang et al., 2022a). For a sufficiently large N_{ext} (a common scenario in practice), the lower bound approximately scales with $0.9N_{\text{rag}}$. (R2) If the knowledge base is highly long-tailed such that samples of certain reference texts are rare (i.e., small $r_{\text{ext}}^{(c)}$), we require a larger sample size of knowledge base N_{ext} to compensate for the long-tail distribution and achieve comparable retrieval quality. (R3) A low-variance retrieval model is expected to generalize well to test distribution and increase retrieval quality. The above guarantee we provide for $\mathbb{E}[N_{\text{pos}}]$ in relation to V_{rag} is a rigorous demonstration of this.

5.3. RAG achieves provably lower conformal generation risk than a single LLM without retrieval

Besides retrieval quality, the generation risk in RAG models is also affected by LLM quality, and to measure transformer quality, we define a (d^+, Φ_M)-transformer as follows.

Definition 2 ((d^+, Φ_M)-transformer)**.** We assume that each in-context example (X_i, Y_i) $(i \in [N_{\text{rag}} + 1])$ is encoded with a single token $\boldsymbol{q}_i$. Let $\boldsymbol{q}$ be the retrieval-augmented input, consisting of N_{rag} retrieved in-context examples and 1 query example. We define a random variable to represent the negative prediction margin: $M = \max_{c \neq g(\boldsymbol{q}_{N_{\text{rag}}+1})} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_{N_{\text{rag}}+1})}(\boldsymbol{q})$, where $O(\boldsymbol{q})$ is the output probability vector without RAG. Let $\Phi_M(\cdot)$ be the CDF of random variable M. We define a (d^+, Φ_M)-transformer as a single-layer self-attention transformer with (a) non-trivial self-attention layer with $\sigma\left((W_K W_E \boldsymbol{q}_i)^T (W_Q W_E \boldsymbol{q}_j)\right) \geq d^+ > 0$ for semantically identical examples with $g(\boldsymbol{q}_i) = g(\boldsymbol{q}_j)$; and (b) the prediction utility that is better than random: $\int_{-1}^{1} \Phi_M(v) dv > 1$.

Remarks. (R1) d^+ measures the minimal attention scores for positive pairs and reflects the effectiveness of the transformer's embedding, key, and query matrices. Since we always have $d^+ \geq 0$ due to the Softmax activation, the condition $d^+ > 0$ only assumes a non-trivial self-attention layer. (R2) The integral $\int_{-1}^{1} \Phi_M(v)dv$ measures the quality of the embedding, value, and projection matrices. Note that a random prediction margin M_{rand} over a uniform distribution $[-1, 1]$ results in $\int_{-1}^{1} \Phi_{M_{\text{rand}}}(v)dv = 1$. Thus, $\int_{-1}^{1} \Phi_M(v)dv > 1 = \int_{-1}^{1} \Phi_{M_{\text{rand}}}(v)dv$ only indicates better-than-random prediction utility.

Next, we prove that RAG in C-RAG achieves a lower conformal generation risk than a single LLM without retrieval with high probability.

Theorem 1 (RAG reduces the conformal generation risk)**.** *Consider the setup in Sec. 5.1 as well as the V_{rag}-retrieval model in Def. 1 and (d^+, Φ_M)-transformer in Def. 2. Let $r_{cal}^{(c)}$ and $r_{ext}^{(c)}$ be as defined in Prop. 3. We show that the conformal generation risk of RAG $\hat{\alpha}_{rag}$ is smaller than that of a single LLM $\hat{\alpha}$ with high probability:*

$$\mathbb{P}\left[\hat{\alpha}_{rag} < \hat{\alpha}\right] \geq 1 - p_t - p_r, \quad \textit{where}$$

$$p_t = \exp\{-2N_{cal}\underbrace{[\Phi_M(\frac{1}{2}\overbrace{d^+(\int_{-1}^{1}\Phi_M(v)dv - 1)}^{\textit{quality of transformers}}N_{rag}) - \Phi_M(0)]^2}_{\textit{improvement of generation quality with RAG}}\}$$

$$p_r = \frac{25}{N_{rag}}(4 - 9\underbrace{\sum_{c=1}^{C} r_{cal}^{(c)}(1.5N_{ext} - r_{ext}^{(c)}N_{ext})V_{rag}^{0.25 r_{ext}^{(c)} N_{ext}}}_{\textit{number of retrieved negative examples}})^{-2} \tag{7}$$

provided that $N_{ext} > 2\sqrt{2\ln 10}/\min_c r_{ext}^{(c)}$, $N_{rag} > 2/d^+$ and $N_{ext}V_{rag}^{0.25 \min_c r_{ext}^{(c)} N_{ext}} < 4/9$. p_t, p_r are the uncertainty induced by the quality of transformer and retrieval model.

Remarks. (R1) The probability of reduced risk with RAG ($\mathbb{P}\left[\hat{\alpha}_{\text{rag}} < \hat{\alpha}\right]$) increases with both N_{cal}, which improves risk approximation via enhanced calibration, and N_{rag} and N_{ext}, which expand the scope of retrieved knowledge. (R2) The reduced risk probability also increases with the transformer's quality induced by attention scores and prediction capability. (R3) A low-variance retrieval model (small V_{rag}) enhances generalization and reduces retrieval model uncertainty p_r. (R4) For certification simplicity, we define positive pairs as semantically identical examples, but this can be expanded to pairs similar in the embedding space for boosting attention scores, in-context learning, and generation quality. (R5) These result can readily extend to various conformal risks such as (Angelopoulos et al., 2022). (R6) For sufficiently large sample size N_{ext} in the external knowledge base, we further have $\mathbb{P}\left[\hat{\alpha}_{\text{rag}} < \hat{\alpha}\right] \geq 1 - p_t - 25/16N_{\text{rag}}$ (Cor. 1 in App. G). In contrast to Thm. 1, the bound has no dependency on the external knowledge distribution $r_{\text{ext}}^{(c)}$ and calibration distribution $r_{\text{cal}}^{(c)}$.

6. C-RAG under distribution shifts

Here, we present a valid distribution-drift conformal generation risk and prove the benefit of RAG compared to vanilla LLM under test-time distribution shifts.

6.1. Analysis setup

Conformal risk guarantees often assume that calibration and testing samples come from the same distribution (Bates et al., 2021; Angelopoulos et al., 2022; 2021). Next, building on Sec. 5.1, we extend these guarantees to distribution shifts between calibration and testing.

6.2. Conformal generation risk under distribution shifts

Under test-time distribution shifts, the certification guarantees of prior work (Angelopoulos et al., 2022; Farinhas et al., 2023a) are limited to the monotonic risk functions and to distribution shifts caused by changes in sample weights. Here, we provide generation risk certification for any bounded risk function and any distribution shift, which is as follows.

Theorem 2 (Conformal generation risk under distribution shifts)**.** *Suppose that the test instance (X_{test}, Y_{test}) is sampled from a shifted distribution $\mathcal{Q}$ with bounded Hellinger distance from the calibration distribution $\mathcal{D}$: $H(\mathcal{D}, \mathcal{Q}) \leq \rho$. Let $\hat{R} = \sum_{i=1}^{N_{cal}} R(Z_i)/N_{cal}$ be the empirical risk on calibration samples Z_i $(i \in \{1, ..., N_{cal}\})$ and $\hat{V} = 1/N_{cal}(N_{cal} - 1)\sum_{1 \leq i < j \leq N_{cal}}(R(Z_i) - R(Z_j))^2$ be the unbiased estimator of the risk variance on the calibration set. Then we have the following guarantee of conformal generation risk on the shifted distribution $\mathcal{Q}$:*

$$\mathbb{P}_{(X_{test}, Y_{test}) \sim \mathcal{Q}}\left[R(T_{\boldsymbol{\lambda}}(X_{test}), Y_{test}) \leq \hat{\alpha}(\rho)\right] \geq 1 - \delta, \quad \textit{where}$$
$$\hat{\alpha}(\rho) := \min\left\{h^{-1}\left(\frac{8/\delta}{N_{cal}}; \overline{\hat{R}_\rho}\right), \Phi_{bin}^{-1}\left(\frac{\delta}{8e}; N_{cal}, \overline{\hat{R}_\rho}\right)\right\}. \tag{8}$$

where $h^{-1}(\cdot;\cdot)$ is the partial inverse function as defined in Prop. 1 and $\overline{\hat{R}_\rho}$ is formulated as:

$$\overline{\hat{R}_\rho} = \underbrace{\hat{R} + \rho^2(2-\rho^2)(1-\hat{R})}_{\textit{empirical mean scaled by } \rho} + \underbrace{2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\hat{V}}}_{\textit{estimated variance scaled by } \rho} +$$
$$\underbrace{(1-\rho^2)\left(\frac{1-\rho^2}{\sqrt{2N_{cal}}} + \frac{2\sqrt{2}\rho\sqrt{2-\rho^2}}{\sqrt{N_{cal}-1}}\right)\sqrt{\ln(4/\delta)} + \sqrt{\frac{\ln(8/\delta)}{2N_{cal}}}}_{\textit{finite-sample error}}$$

where the radius ρ satisfies the following: $\rho^2 \leq 1 - \left[1 + \left(\hat{R} - 1 + \sqrt{\ln(4/\delta)/2N_{cal}}\right)^2 / \left(\sqrt{\hat{V}} + \sqrt{2\ln(2/\delta)/(N_{cal}-1)}\right)^2\right]^{-2}$.

Remarks. Thm. 2 offers a distribution-drift conformal generation risk $\hat{\alpha}(\rho)$, under the distribution shift with radius ρ. (R1) We adopt the Hellinger distance for measuring distribution distances due to its f-divergence properties and direct applicability to total variation distance between $\mathcal{D}$ and $\mathcal{Q}$ ((Steerneman, 1983), Equation 1). (R2) For conformal guarantees under distribution shifts, we derive an empirical risk upper bound considering worst-case shifts in Thm. 2, which is efficiently calculable with empirical statistics on calibration distribution $\mathcal{D}$. (R3) We recover the empirical mean $\hat{R}$ from $\hat{\overline{R}}_\rho$ when $\rho \to 0$ and $N_{\text{cal}} \to \infty$ in Thm. 2.

6.3. RAG achieves provably lower conformal generation risk than a single LLM under distribution shifts

Next, we prove that RAG mitigates conformal generation risk better than a single LLM under distribution shifts.[1]

Theorem 3 (RAG reduces conformal generation risk even under distribution shifts)**.** *Suppose that the shifted test distribution $\mathcal{Q}$ is within bounded Hellinger distance $\rho > 0$ to the calibration distribution $\mathcal{D}$. Consider the same setup and assumptions as Thm. 1. Consider also a V_{rag}-retrieval model in Def. 1 and (d^+, Φ_M)-transformer in Def. 2. Under the condition that $N_{ext} > 2\sqrt{2\ln 10}/r^m_{ext}$, $N_{ext} V_{rag}(\rho)^{0.25 r^m_{ext} N_{ext}} < 8/17$, and $N_{rag} > 2/d^+$, we have:*

$$\mathbb{P}\left[\hat{\alpha}_{rag}(\rho) < \alpha(\rho)\right] \geq 1 - p_t - p_r(\rho), \quad \textit{where}$$

$$p_r(\rho) = \frac{100}{N_{rag}} \left(8 - 17 N_{ext} V_{rag}(\rho)^{0.25\left(\min_c r^{(c)}_{ext} N_{ext}\right)} \right)^{-2}, \tag{9}$$

where p_t is the uncertainty induced by the transformer quality as Eq. (7) and $p_r(\rho)$ is the uncertainty induced by the retrieval model. Moreover, $V_{rag}(\rho) = m(\rho) V_{rag}$ quantifies the quality of retrieval models under distance ρ, where

$$m(\rho) = \underbrace{\left(\frac{\sqrt{-6\rho^4 + 12\rho^2 + 1} - 4\rho(1-\rho^2)\sqrt{2-\rho^2}}{1 - 16\rho^2 + 8\rho^4} \right)^{-2}}_{\textit{retrieval model quality decay factor by distribution shifts}}.$$

Remarks. Our result rigorously characterizes the effect of distribution shift on the reduced risk guarantee of RAG. (R1) Compared to Thm. 1, only the uncertainty of retrieval model $p_r(\rho)$ is affected by the distribution shift ρ. This affect is reflected on the the retrieval quality $V_{\text{rag}}(\rho)$. In particular, a large distance radius ρ will downgrade the retrieval quality $V_{\text{rag}}(\rho)$ and thus lead to a higher uncertainty $p_r(\rho)$. However, the influence of ρ on $p_r(\rho)$ can be reduced by N_{rag} inverse proportionally and by N_{ext} exponentially, demonstrating the robustness of RAG with more retrieval knowledge. (R2) Since $V_{\text{rag}}(\rho)$ is proportional to model variance V_{rag}, a low-variance retrieval model demonstrates better robustness against distribution drifts, aligning with existing empirical observations (Lam, 2016; Gotoh et al., 2018; Namkoong & Duchi, 2017), which evaluate the generalization ability of low-variance retrieval models under distribution shifts. (R3) Compared to Thm. 1, Thm. 3 has no dependence on varying label portions $r^{(c)}_{\text{cal}}$ during distribution shifts, as long as the size of external knowledge base N_{ext} is moderately large, a condition often met in practice with large knowledge bases.

7. Evaluation

We evaluate C-RAG on four datasets using different retrieval models. We find that **(1)** our conformal generation risks in Prop. 1 is empirically sound and tight in Sec. 7.2, **(2)** RAG reduces the conformal generation risks for different retrieval models, which empirically validates Thm. 1 in Sec. 7.2, **(3)** the conformal generation risk under distribution shifts in Thm. 2 is empirically sound and tight for varying distances in Sec. 7.3, **(4)** multi-dimensional RAG configurations maintain sound and tight conformal generation risks in Sec. 7.4, and **(5)** C-RAG computes valid configurations with empirical risks always below the desired risk level in Sec. 7.5.

The codes are publicly available at https://github.com/kangmintong/C-RAG.

7.1. Evaluation setup

Datasets & knowledge base We evaluate C-RAG on four widely used NLP datasets, including AESLC (Zhang & Tetreault, 2019), CommonGen (Lin et al., 2019), DART (Nan et al., 2020), and E2E (Novikova et al., 2017). Following (Wang et al., 2023c; Cheng et al., 2023), we construct the knowledge base as a collection of 30 public datasets from 9 distinct categories with over 6 million documents.

Retrieval models We consider four retrieval models: (1) BM25 (Robertson et al., 2009) with token-level matching scores, (2) BAAI/bge (Zhang et al., 2023a) as SOTA embedding model in MTEB benchmark (Muennighoff et al., 2022), (3) OpenAI/ada as SOTA close source text embedding model, and (4) Biencoder-SFT (Wang et al., 2023c) as a biencoder retriever trained with in-domain data samples.

RAG Generation protocol We use our generation protocol (Alg. 1 in App. C.1) controlled by the number of retrieved examples N_{rag}, generation set size λ_g, and diversity threshold λ_s. We use Llama-2-7b for inference and perform conformal calibration on validation sets with uncertainty $\delta = 0.1$. We use $1 -$ ROUGE-L as the risk function. See App. J.1 for more details of evaluation setup.

7.2. Evaluation of conformal generation risks

Soundness and tightness of conformal generation risks To achieve generation risk guarantee in Eq. (3), C-RAG computes the conformal generation risk using Prop. 1. We evaluate the conformal generation risks of RAG models $\hat{\alpha}_{\text{rag}}$ under different numbers of retrieved examples N_{rag} by calibration statistics on the validation set. To validate the soundness and tightness of the conformal generation risk guarantee, we evaluate the empirical risks on randomly sampled test instances. The sampling protocol is detailed in Alg. 3 in App. J.2. We provide the results using OpenAI/ada retrieval model in Fig. 3 and results for BM25, BAAI/bge, Biencoder-SFT in Figs. 8 to 10 in App. J.2. The results show

[1] Additionally, we examine retrieval quality and prove a lower bound of retrieved positive examples under test-time distribution shifts. We leave the analysis to App. H for interested readers.

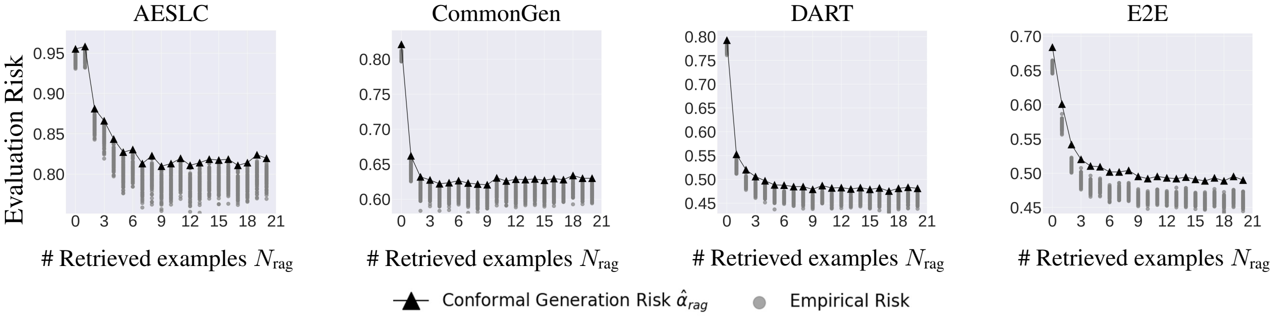


Figure 3: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and empirical risk based on retrieval model OpenAI/ada taking different N_{rag} ($\lambda_g = 1, \lambda_s = 1.0$). We observe that our conformal generation risk (Prop. 1) is valid and tight; larger N_{rag} reduces risk $\hat{\alpha}_{\text{rag}}$ (empirically validating Thm. 1).

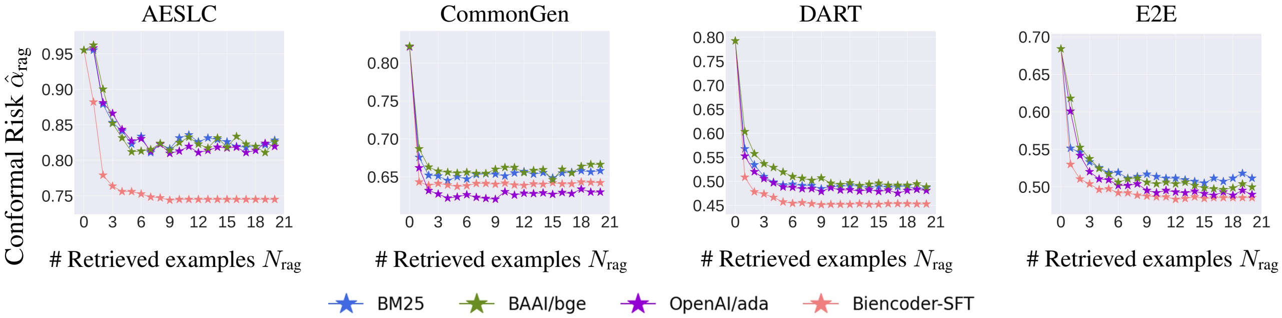


Figure 4: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ with different N_{rag} using different retrieval models ($\lambda_g = 1, \lambda_s = 1.0$). We observe that large N_{rag} effectively reduces $\hat{\alpha}_{\text{rag}}$ for different models; the trained Biencoder-SFT usually leads to the lowest conformal generation risk.

that (1) the conformal generation risks $\hat{\alpha}_{\text{rag}}$ upper bound the empirical risks of the sampled test instances, (2) for some test instances, the empirical risks nearly reach the conformal generation risk, demonstrating the soundness and tightness of our generation risk guarantees, (3) the conformal generation risk decreases as the number of retrieved examples N_{rag} increases, which shows the effectiveness of RAG models and aligns with our theoretical analysis in Thm. 1.

Comparisons of different SOTA retrieval models We compare the conformal generation risks for token-level BM25 scores, and SOTA embedding models BAAI/bge, OpenAI/ada, and Biencoder-SFT. The results in Fig. 4 show that RAG achieves lower conformal generation risks than LLM without retrieval (i.e., $N_{\text{rag}} = 0$) for different retrieval models. Biencoder-SFT, trained with in-domain data samples, leads to lower conformal generation risk in general compared with other retrieval models. OpenAI/ada, which is known of high quality and trained on large open corpus, also demonstrates low conformal generation risks.

7.3. Conformal generation risk under distribution shifts

Soundness and tightness of conformal generation risk under distribution shifts In practice, user input text may deviate from the calibration distribution. In Thm. 2, we provide the first conformal generation risk under distribution shifts for general bounded risk functions. We evaluate the conformal generation risk $\hat{\alpha}(\rho)$ in Eq. (8). To empirically verify the soundness, we create test sets with covariate shifts by varying sample weights. The Hellinger distance is computed using original and shifted sample weights, with details in Alg. 4 in App. J.3. We compare the conformal generation risk and empirical risks with $N_{\text{rag}} = 15$ using OpenAI/ada in Fig. 5, and using BM25, BAAI/bge and Biencoder-SFT in Figs. 11 to 13 in App. J.3. The results show that (1) our conformal generation risks under distribution shifts are sound and tight across various models, and (2) conformal generation risks increase linearly with Hellinger distance ρ, remaining non-trivial up to $\rho = 0.2$.

Comparison of SOTA retrieval models under distribution shifts We compare conformal generation risks for different retrieval models under distribution shifts in Fig. 14 in App. J.3. All models show a linear rise in risk with increasing Hellinger distance, with BiEncoder-SFT and OpenAI/ada showing lower risks at varying distances.

7.4. C-RAG with multi-dimensional RAG configurations

So far, we demonstrate the effectiveness of retrieved in-context examples quantified by N_{rag}. To further improve the conformal generation risk, we can adjust the RAG configurations, such as the number of generations λ_g and the diversity-controlling similarity threshold λ_s. We follow RAG generation protocol in Alg. 1 and define the risk function as the minimal risk among λ_g candidate generations. Our tests on AESLC, CommonGen, DART, and E2E datasets (see Fig. 6 and Fig. 15 in App. J.4) show that the multi-dimensional RAG configurations maintain sound and tight conformal

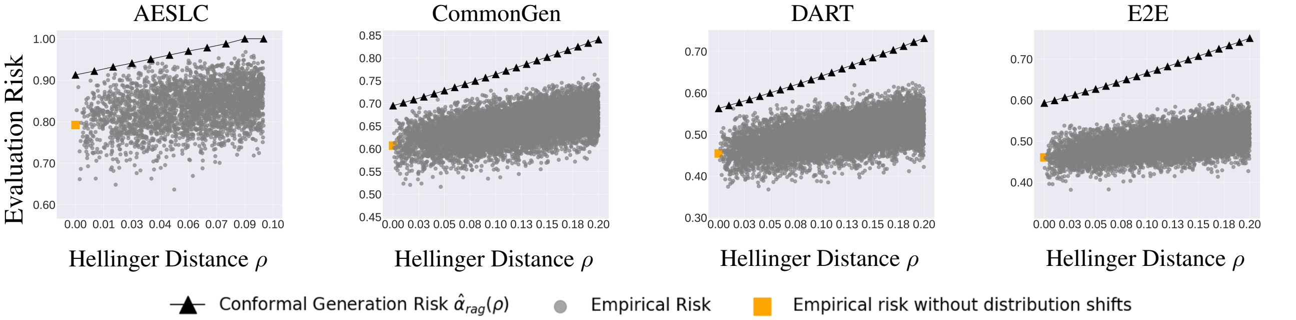


Figure 5: Conformal generation risk $\hat{\alpha}_{\text{rag}}(\rho)$ and empirical risks based on retrieval model OpenAI/ada under distribution shifts ($N_{\text{rag}} = 15, \lambda_g = 1, \lambda_s = 1.0$). We observe that our distribution-drift conformal generation risk (Thm. 2) is empirically valid and tight.

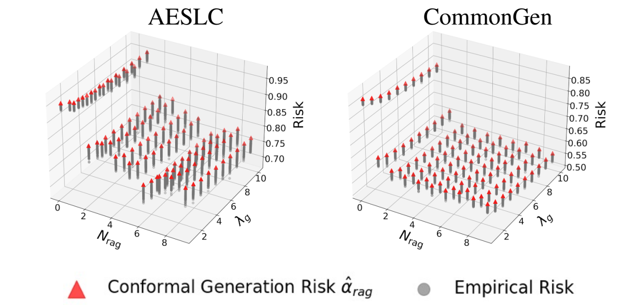


Figure 6: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and empirical risks with different λ_g and N_{rag} for OpenAI/ada.

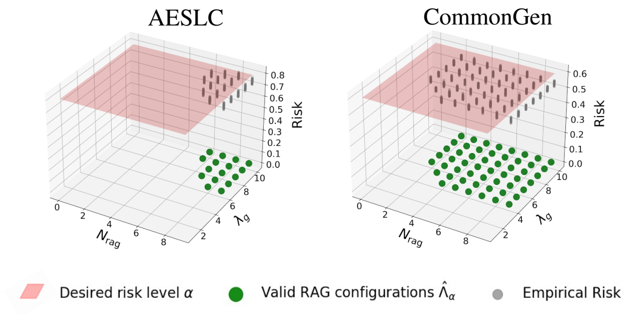


Figure 7: Valid configurations $\hat{\Lambda}_\alpha$ given a desired risk level α and the empirical risks with different λ_g and N_{rag} for OpenAI/ada.

generation risks. Notably, a higher N_{rag} reduces generation risks more effectively than adjusting λ_g.

7.5. Valid configurations given desired risk levels

In risk guarantee **(2)** outlined in Sec. 4.3, given a desired risk level α, C-RAG computes a valid RAG configuration set $\hat{\Lambda}_\alpha$, such that configurations within this set will lead to generation risks below α. We apply the Bonferroni correction in Prop. 2 for rigorous family-wise error rate control and assess empirical risks on random test sets with the identified valid configurations. We provide the results on AESLC and CommenGen in Fig. 7 and results on DART and ECE in Fig. 15 in App. J.5. These results validate our certification, as the empirical risks of generated configurations $\hat{\Lambda}_\alpha$ are consistently below the given conformal generation risk α.

The results also show that a high number of retrieved examples N_{rag} and a larger generation set size λ_g contribute to reducing conformal generation risk. The impact of the diversity threshold λ_s is explored in App. J.5.

In App. J.6, we provide comparisons of the conformal generation risks for different LLMs as inference models. In App. J.7, we also provide a quantitative example to show how the constrained generation protocol benefits in reducing LLM hallucination risks.

8. Conclusion

In this paper, we propose C-RAG to provide conformal generation risk guarantees for RAG models. C-RAG certifies **(1)** a conformal generation risk for a given RAG configuration, and **(2)** a valid configuration set for a given desired risk level. We theoretically show that RAG reduces conformal generation risks of a single LLM. We empirically validate the soundness and tightness of our risk guarantees.

Acknowledgements

This work is partially supported by the National Science Foundation under grant No. 1910100, No. 2046726, No. 2229876, DARPA GARD, the National Aeronautics and Space Administration (NASA) under grant No. 80NSSC20M0229, the Alfred P. Sloan Fellowship, the Amazon research award, and the eBay research award.

Impact Statement

The C-RAG framework is able to safeguard LLMs' practical deployment and applications against ethical and societal concerns. Existing research shows that the responses of LLMs can be biased towards some demographic groups and not be aligned with human ethics. With C-RAG, we can define a bias/ethics risk function and control the generation risk below a desired level. The risk guarantee provided by C-RAG enhances the use of LLMs, addressing societal issues and regulatory infringements. We do not expect any negative societal consequences for our work.

References

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Basu, S., Rawat, A. S., and Zaheer, M. A statistical perspective on retrieval-based models. In *International Conference on Machine Learning*, pp. 1852–1886. PMLR, 2023.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Bentkus, V. On hoeffding's inequalities. 2004.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Chen, F.-L., Zhang, D.-Z., Han, M.-L., Chen, X.-Y., Shi, J., Xu, S., and Xu, B. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

Chen, S., Li, Z., Liu, C., and Yang, W. Decix: Explain deep learning based code generation applications. *arXiv preprint arXiv*, 2024a.

Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., Shen, Y., Gu, J., and Chen, H. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*, 2024b.

Cheng, D., Huang, S., Bi, J., Zhan, Y., Liu, J., Wang, Y., Sun, H., Wei, F., Deng, D., and Zhang, Q. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*, 2023.

Farinhas, A., Zerva, C., Ulmer, D., and Martins, A. F. Non-exchangeable conformal risk control. *arXiv preprint arXiv:2310.01262*, 2023a.

Farinhas, A., Zerva, C., Ulmer, D., and Martins, A. F. T. Non-exchangeable conformal risk control, 2023b.

Gatt, A. and Krahmer, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

Gotoh, J.-y., Kim, M. J., and Lim, A. E. Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters*, 46(4):448–452, 2018.

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., and Chen, W. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

Han, C., Wang, Z., Zhao, H., and Ji, H. In-context learning of large language models explained as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

Hermans, A., Beyer, L., and Leibe, B. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pp. 409–426, 1994.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Huang, J., Ping, W., Xu, P., Shoeybi, M., Chang, K. C.-C., and Catanzaro, B. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*, 2023.

Jiang, L., Wu, Y., Xiong, J., Ruan, J., Ding, Y., Guo, Q., Wen, Z., Zhou, J., and Deng, X. Hummer: Towards limited competitive preference dataset. *arXiv preprint arXiv:2405.11647*, 2024.

Kang, M., Lin, Z., Sun, J., Xiao, C., and Li, B. Certifiably byzantine-robust federated conformal prediction. 2023.

Kang, M., Gürel, N. M., Li, L., and Li, B. Colep: Certifiably robust learning-reasoning conformal prediction via probabilistic circuits. *arXiv preprint arXiv:2403.11348*, 2024a.

Kang, M., Song, D., and Li, B. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 36, 2024b.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Kishore, V., Wan, C., Lovelace, J., Artzi, Y., and Weinberger, K. Q. Incdsi: incrementally updatable document retrieval. In *International Conference on Machine Learning*, pp. 17122–17134. PMLR, 2023.

Lam, H. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Li, Y., Dong, B., Lin, C., and Guerin, F. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201*, 2023.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., and Ren, X. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*, 2019.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., and Li, H. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2023.

Luo, Z., Xu, C., Zhao, P., Geng, X., Tao, C., Ma, J., Lin, Q., and Jiang, D. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*, 2023.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30, 2017.

Nan, L., Radev, D., Zhang, R., Rau, A., Sivaprasad, A., Hsieh, C., Tang, X., Vyas, A., Verma, N., Krishna, P., et al. Dart: Open-domain structured data record to text generation. *arXiv preprint arXiv:2007.02871*, 2020.

Novikova, J., Dušek, O., and Rieser, V. The e2e dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*, 2017.

OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2023.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Robertson, S., Zaragoza, H., et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.

Saw, J. G., Yang, M. C., and Mo, T. C. Chebyshev inequality with estimated mean and variance. *The American Statistician*, 38(2):130–132, 1984.

Stankeviciute, K., M Alaa, A., and van der Schaar, M. Conformal time-series forecasting. *Advances in neural information processing systems*, 34:6216–6228, 2021.

Steerneman, T. On the total variation and hellinger distance between signed measures; an application to product measures. *Proceedings of the American Mathematical Society*, 88(4):684–688, 1983.

Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35: 21831–21843, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Vovk, V., Gammerman, A., and Saunders, C. Machine-learning applications of algorithmic randomness. 1999.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022a.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S., Arora, S., Mazeika, M., Hendrycks, D., Liu, Z., Cheng, Y., Keyejo, S., Song, D., and Li, B. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *NeurIPS*, 2023a.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023b.

Wang, L., Yang, N., and Wei, F. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*, 2023c.

Wang, Y., Hou, Y., Wang, H., Miao, Z., Wu, S., Chen, Q., Xia, Y., Chi, C., Zhao, G., Liu, Z., et al. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35:25600–25614, 2022b.

Weber, M. G., Li, L., Wang, B., Zhao, Z., Li, B., and Zhang, C. Certifying out-of-domain generalization for blackbox functions. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., and Overwijk, A. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Yang, Y. and Kuchibhotla, A. K. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., and Yu, D. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

Zhang, P., Xiao, S., Liu, Z., Dou, Z., and Nie, J.-Y. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*, 2023a.

Zhang, R. and Tetreault, J. This email could save your life: Introducing the task of email subject line generation. *arXiv preprint arXiv:1906.03497*, 2019.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023b.

Appendices

Contents

A Discussions of limitations and future work 15

B Additional related work 15

C Conformal generation risks for RAG models 16

C.1 Constrained generation protocol for RAG models 16

D Risk Guarantees 16

D.1 Risk guarantee (1) (Prop. 1) 16

D.2 Risk guarantee (2) (Prop. 2) 18

E Grammian generalization bound 19

F Proofs and detailed remarks in Sec. 5 19

F.1 Detailed remark of Def. 1 19

F.2 Detailed remark of Def. 2 19

F.3 Proof and detailed remark of Prop. 3 20

F.4 Proof and detailed remark of Thm. 1 21

G Cor. 1: Asymtotic result of Thm. 1 25

H Prop. 4: Retrieval quality analysis under distribution shifts 26

I Proofs and detailed remarks in Sec. 6 29

I.1 Proof and detailed remark of Thm. 2 29

I.2 Proof and detailed remark of Thm. 3 30

J Additional evaluation results 32

J.1 Experiment setup 32

J.2 Conformal risk bounds evaluation 34

J.3 Conformal risk bounds evaluations under distribution shifts 35

J.4 Conformal risk bounds with multi-dimensional RAG configurations 36

J.5 Valid configurations identification with specified risk levels 37

J.6 Conformal generation risks with different inference models 37

J.7 Qualitative example 38

A. Discussions of limitations and future work

Limitations One potential challenge of applying C-RAG practically could be the collection of calibration data. In practice, the user input texts are sampled from a time-series data distribution. Therefore, accessing in-distribution calibration samples requires collecting real-time query samples, which could pose the challenge of computational resources and system latency. Another potential limitation may lie in the probability of the guarantee. Since C-RAG can only provide a high-confidence risk bound via conformal risk analysis, generations with excessive risks can exist. Therefore, we may need more calibration samples to counter for a higher confidence level, and thus mitigate the appearance of outliers to a large extent. Also, although the analysis in C-RAG shows the benefits of a large external knowledge base to a low conformal generation risk, the large knowledge base may induce a larger time complexity of KNN searching and space complexity of storing the examples, leading to a trade-off between the generalization/utility and inference efficiency.

Future work One interesting future work is to provide conformal risk analysis for time-series data. Conformal prediction for time series (Zaffran et al., 2022; Xu & Xie, 2021; Stankeviciute et al., 2021) adaptively adjusts the prediction coverage for sequential data for the regression and classification task. However, the adaptive risk calibration for time series is unexplored but important to practical deployments. Therefore, conformal risk analysis for time series can further motivate the application of conformal risk analysis for LLMs.

Further discussions on calibration data collection. In principle, if test and calibration instances are from the same distribution, randomly sampling from this distribution with a sufficient sample size N_{cal} already provides competitive generation risk guarantees (Proposition 1 and Theorem 1). Otherwise, if sampling from the test distribution is impractical, we should sample instances from a proposal distribution with a small distribution distance (ρ) to the test distribution and sample variance ($\hat{V}$) so the distribution of the calibration set mimics that of test data (Theorem 2). We can further use the following techniques for calibration data selection: (1) rejection sampling: drawing samples from a proposal distribution and then rejecting some of these samples based on the known criterion of the test distribution, (2) importance sampling: adjusting the sample weights to closely match the target distribution, and (3) variance reduction such as input normalization. To exemplify, consider a composite domain with medical support, wiki question answering, and service assistance fields, where only a broad proposal distribution is available. We can leverage the strategies mentioned above as follows. We can (1) reject out-of-scope samples, (2) perform importance sampling by adjusting the sample weights based on the proposal distribution and the test distribution, and (3) normalize samples to minimize the distribution gap and variance, for instance, through a unified prompt reformulation. To improve the probability of the risk guarantee given fixed sample sizes, one can seek advanced concentration analysis with additional constraints on data distribution, which may lead to tighter risk bounds in practice.

B. Additional related work

Retrieval augmented generation (RAG) is a framework for improving the generation quality of LLMs via retrieving relevant information from the external knowledge base and grounding the model on the information for conditional generations. Biencoder retrieval methods (Lewis et al., 2020; Karpukhin et al., 2020; Xiong et al., 2020) leverage two encoders to map the query text and candidate texts into the embedding space and retrieve candidate texts with high embedding similarity to the query text embedding. End-to-end retrieval methods (Tay et al., 2022; Wang et al., 2022b; Kishore et al., 2023) train a model to map the query text to the id of relevant candidate documents directly. Another line of work (Luo et al., 2023; Gou et al., 2023) leverages external tools such as LLMs to retrieve relevant documents via prompting design. Although RAG demonstrates impressive capacities, the theoretical analysis of retrieval models for LLM generations is limited. Basu et al. analyze the retrieval model of constrained function class from a statistical perspective, but the results cannot be generalized to the self-attention transformers. In this work, we provide the first analysis of how RAG enhances the generation quality and mitigate generation risks of self-attention transformers.

Conformal prediction is a statistical tool to construct the prediction set with guaranteed prediction coverage (Vovk et al., 1999; 2005; Lei et al., 2013; Yang & Kuchibhotla, 2021; Kang et al., 2023; 2024a), assuming that the data is exchangeable. However, conformal prediction can only provide guarantees for the regression and classification tasks and is not directly applicable to the generation tasks, which are more relevant for LLMs. Conformal risk controlling methods (Bates et al., 2021; Angelopoulos et al., 2021; 2022; Quach et al., 2023) provide a high-confidence risk guarantee with the data exchangeability assumption for any black-box risk functions. We can define a specific risk function for a RAG model and certify a risk upper bound of generations based on statistics on in-distribution calibration set. However, the risk guarantee is violated under

Algorithm 1 Constrained generation protocol for RAG

```
Input: input prompt $X_{\text{test}}$, LM $p_{\theta_l}(y|x)$, generation set size $\lambda_s$, retrieved example size $N_{\text{rag}}$, external knowledge base $\hat{\mathcal{D}}_{\text{ext}}$, similarity measurement function $s_{\theta_r}(\cdot,\cdot)$ with embedding model parameterized by $\theta_r$, generation similarity threshold $\lambda_g$, parameter configuration $\boldsymbol{\lambda} = [N_{\text{rag}}, \lambda_g, \lambda_s]$
Output: Generation set $\mathcal{G}_{\boldsymbol{\lambda}}(X_{\text{test}})$
$\mathcal{G}_{\boldsymbol{\lambda}}(X_{\text{test}}) \leftarrow \Phi$
$\mathcal{Z} \leftarrow \text{KNN}(X_{\text{test}}, N_{\text{rag}}; \hat{\mathcal{D}}_{\text{ext}}, s_{\theta_r})$ {Retrieve $N_{\text{rag}}$ examples from $\hat{\mathcal{D}}_{\text{ext}}$ via KNN search with similarity measurement $s_{\theta_r}(\cdot,\cdot)$}
$X^{(\text{rag})} \leftarrow \text{Template}(X_{\text{test}}, \mathcal{Z})$ {Augmented prompt with $X_{\text{test}}$ and retrieved examples $\mathcal{Z}$ with a template}
while $|\mathcal{G}_{\boldsymbol{\lambda}}(X_{\text{test}})| < \lambda_s$ do
    $y \sim p_{\theta_l}(\cdot|X^{(\text{rag})})$
    while $\exists g \in \mathcal{G}_{\boldsymbol{\lambda}}, s_{\theta_r}(y, g) > \lambda_g$ do
        $y \sim p_{\theta_l}(\cdot|X^{(\text{rag})})$ {Reject sampling}
    end while
    $\mathcal{G}_{\boldsymbol{\lambda}}(X_{\text{test}}) = \mathcal{G}_{\boldsymbol{\lambda}}(X_{\text{test}}) \cup \{y\}$
end while
Return $\mathcal{G}_{\boldsymbol{\lambda}}(X_{\text{test}})$
```

distribution shifts at test time. Angelopoulos et al.; Farinhas et al. offer a valid conformal risk for monotonic risk functions under distribution shifts, but the monotonicity assumption may not always hold in practice. In this work, we introduce the first conformal risk bound for general bounded risk functions under test-time distribution shifts.

C. Conformal generation risks for RAG models

C.1. Constrained generation protocol for RAG models

To safeguard diverse foundation model-based applications (Chen et al., 2024b;a; Jiang et al., 2024; Li et al., 2023; Chen et al., 2023; Zhang et al., 2024), we typically leverage RAG to enhance the trustworthiness of generations (Wang et al., 2023b; Kang et al., 2024b). RAG models (Wang et al., 2023c; Rubin et al., 2021; Huang et al., 2023) combine a retrieval model and a generation LM. The retrieval model retrieves N_{rag} relevant examples to the query from an external knowledge base, and the LM learns in-context from these examples. The knowledge base contains N_{ext} samples in $\hat{\mathcal{D}}_{\text{ext}} = \{(X_i, Y_i)\}_{i\in[N_{\text{ext}}]}$. The retrieval model uses an encoder to map instances into an embedding space, and then identifies the relevant examples to the query X_{test} based on similarity. This similarity, defined by $s_{\theta_r}(\cdot,\cdot): \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and parameterized by θ_r, is used to find the nearest examples using KNN search in the embedding space.

We arrange the retrieved N_{rag} in-context examples and the test example X_{test} into augmented input text $X^{(\text{rag})}$ using a template. We then sample the generation from $p_{\theta_l}(\cdot|X^{(\text{rag})})$ repeatedly until λ_g generations are collected. To control the diversity of generations, we reject those with a similarity higher than a threshold λ_s to the previous generations. In essence, the constrained generation protocol is controlled by configuration $\boldsymbol{\lambda} = [N_{\text{rag}}, \lambda_g, \lambda_s]$ and output a generation set $T_{\boldsymbol{\lambda},p_{\theta_l}}(x)$ based on the configuration $\boldsymbol{\lambda}$ and input x. We refer to Alg. 1 for the pseudocode of the protocol.

D. Risk Guarantees

D.1. Risk guarantee (1) (Prop. 1)

Proof of Prop. 1. The proof sketch follows (Angelopoulos et al., 2021). Since the risk function $R(\cdot,\cdot)$ is upper bounded by 1, we can apply a tighter version of Hoeffding's inequality (Hoeffding, 1994) for $\hat{\alpha} > \mathbb{E}[R(T_{\boldsymbol{\lambda},p_{\theta_l}}(x), y)]$:

$$\mathbb{P}\left[R(T_{\boldsymbol{\lambda},p_{\theta_l}}(x), y) \geq \hat{\alpha}\right] \leq \exp\left\{-N_{\text{cal}} h(\hat{R}(\hat{\mathcal{D}}_{\text{cal}}), \hat{\alpha})\right\} \tag{10}$$

Also, applying Bentkus inequality (Bentkus, 2004), we have:

$$\mathbb{P}\left[R(T_{\boldsymbol{\lambda},p_{\theta_l}}(x), y) \geq \hat{\alpha}\right] \leq e\mathbb{P}\left[\text{Bin}(N_{\text{cal}}, \hat{\alpha}) \leq \left\lceil N_{\text{cal}} \hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right\rceil\right] \tag{11}$$

Combining Eqs. (10) and (11), we have:

$$\mathbb{P}\left[R(T_{\boldsymbol{\lambda},p_{\theta_l}}(x),y)\geq\hat{\alpha}\right]\leq\min\left(\exp\left\{-N_{\text{cal}}h\left(\hat{R}(\hat{\mathcal{D}}_{\text{cal}}),\hat{\alpha}\right)\right\},e\mathbb{P}\left[\text{Bin}(N_{\text{cal}},\hat{\alpha})\leq\left\lceil N_{\text{cal}}\hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right\rceil\right]\right)\tag{12}$$

Or equivalently, given uncertainty $1-\delta$, we have:

$$\delta=\min\left(\exp\left\{-N_{\text{cal}}h\left(\hat{R}(\hat{\mathcal{D}}_{\text{cal}}),\hat{\alpha}\right)\right\},e\mathbb{P}\left[\text{Bin}(N_{\text{cal}},\hat{\alpha})\leq\left\lceil N_{\text{cal}}\hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right\rceil\right]\right),\tag{13}$$

which leads to the following by formulating an inverse function:

$$\hat{\alpha}=\min\left\{h^{-1}\left(\frac{1/\delta}{N_{\text{cal}}};\hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right),\Phi^{-1}_{\text{bin}}\left(\frac{\delta}{e};N_{\text{cal}},\hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right)\right\}\tag{14}$$

□

Remarks. Given the constrained generation protocol $T_{\boldsymbol{\lambda},p_{\theta_l}}$, RAG generation parameter $\boldsymbol{\lambda}$, a calibration set $\hat{\mathcal{D}}_{\text{cal}}$, and a risk function $R(\cdot,\cdot):2^{\mathcal{Y}}\times\mathcal{Y}\mapsto\mathbb{R}$, we aim to provide a risk guarantee of the test sample $(X_{\text{test}},Y_{\text{test}})$:

$$\mathbb{P}\left[R(T_{\boldsymbol{\lambda},p_{\theta_l}}(X_{\text{test}}),Y_{\text{test}})\leq\hat{\alpha}\right]\geq 1-\delta,\tag{15}$$

where $\hat{\alpha}$ is the conformal risk upper bound, and the confidence level $1-\delta$ can be computed by Hoeffding-Bentkus inequalities (Bates et al., 2021):

$$\delta=\min\left(\exp\left\{-N_{\text{cal}}h\left(\hat{R}(\hat{\mathcal{D}}_{\text{cal}}),\hat{\alpha}\right)\right\},e\mathbb{P}\left[\text{Bin}(N_{\text{cal}},\hat{\alpha})\leq\left\lceil N_{\text{cal}}\hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right\rceil\right]\right),\tag{16}$$

where $h(a,b)=a\log(a/b)+(1-a)\log((1-a)/(1-b))$, $\text{Bin}(\cdot,\cdot)$ denotes the binomial distribution, N_{cal} is the number of samples in the calibration set, and $\hat{R}(\cdot)$ computes the empirical risk on the calibration set.

Given the confidence level $1-\delta$, we can also inversely compute the conformal risk upper bound $\hat{\alpha}$ as the following:

$$\hat{\alpha}=\min\left\{h^{-1}\left(\frac{1/\delta}{N_{\text{cal}}};\hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right),\Phi^{-1}_{\text{bin}}\left(\frac{\delta}{e};N_{\text{cal}},\hat{R}(\hat{\mathcal{D}}_{\text{cal}})\right)\right\}\tag{17}$$

where $h^{-1}(\cdot;\cdot)$ is the partial inverse function such that $h^{-1}(h(a,b);a)=b$ with $h(a,b)=a\log(a/b)+(1-a)\log((1-a)/(1-b))$, and Φ^{-1}_{bin} denotes the inverse of CDF of binomial distribution. The HB bound uses the empirical risk on the calibration set as test statistics and provides finite-sample statistical results with surprising empirical effectiveness.

Alternative approach for Risk Guarantee Prop. 1 We can obtain a tighter guarantee of the conformal risk if we assume that the given configuration vector $\boldsymbol{\lambda}$ has dimension 1 (i.e., $B=1$) and the risk function $R(\cdot,\cdot)$ monotonically increases in parameter $\boldsymbol{\lambda}$ and is upper bounded by C ($R:2^{\mathcal{Y}}\times\mathcal{Y}\mapsto(-\infty,C]$). Then, we can have the following conformal risk guarantee according to (Angelopoulos et al., 2022):

$$\frac{N_{\text{cal}}}{N_{\text{cal}}+1}\hat{R}(\hat{\mathcal{D}}_{\text{cal}})-\frac{C}{N_{\text{cal}}+1}\leq\mathbb{E}\left[R(T_{\hat{\boldsymbol{\lambda}},p_{\theta_l}}(X_{\text{test}}),Y_{\text{test}})\right]\leq\frac{N_{\text{cal}}}{N_{\text{cal}}+1}\hat{R}(\hat{\mathcal{D}}_{\text{cal}})+\frac{C}{N_{\text{cal}}+1}.\tag{18}$$

Although the guarantee in Eq. (18) is tighter with the guarantee of the upper bound and the lower bound, the additional assumption of single dimensionality and monotonicity does not hold for many practical generation protocols. Therefore, we mainly consider the conformal risk bound in Eq. (17) across the analysis. We also add discussions that C-RAG is flexible in considering different types of conformal risk bounds, which basically presents an explicit function of the controlled risk with respect to the empirical risk. Since we build the connection between the empirical risk $\hat{R}$ to the retrieval model in the risk in the analysis, we only need to directly connect the empirical risk to the controlled risk via the explicit function to achieve end-to-end certification.

Algorithm 2 Graph-based valid configurations search

```
Input: confidence error level $\delta$, parameter configurations $\Lambda = \{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_N\}$, $p-$values $(p_1, ..., p_N)$, graph $\mathcal{G}$, initial error budget $\delta_i$ such that $\sum_i \delta_i = \delta$
Output: valid configurations set with certified conformal risk $\hat{\Lambda}$
$\hat{\Lambda} \leftarrow \Phi$
while $\exists i : p_i \leq \delta_i$ do
    Select any $i$ such that $p_i \leq \delta_i$
    $\hat{\Lambda} \leftarrow \hat{\Lambda} \cup \{\boldsymbol{\lambda}_i\}$
    Update the error level and the graph:
        $\delta_j \leftarrow \begin{cases} \delta_j + \delta_i g_{i,j}, & \lambda_j \in \Lambda \backslash \hat{\Lambda} \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad g_{k,j} \leftarrow \begin{cases} \dfrac{g_{k,j} + g_{k,i} g_{i,j}}{1 - g_{k,i} g_{i,k}}, & \lambda_k, \lambda_j \in \Lambda \backslash \hat{\Lambda}, k \neq j \\ 0, & \text{otherwise} \end{cases} \qquad (21)$
end while
Return $\hat{\Lambda}$
```

D.2. Risk guarantee (2) (Prop. 2)

Proof of Prop. 2. The proof follows (Holm, 1979). We consider $|\Lambda|$ independent hypothesis test corresponding to the $|\Lambda|$ Null hypothesis. By the Bonferroni method, each test is performed at a significance level of $\dfrac{\delta}{|\Lambda|}$. Therefore, The probability of not making a Type I error in a single test is $1 - \dfrac{\delta}{|\Lambda|}$. The probability of making no Type I error in all $|\Lambda|$ tests is $(1 - \dfrac{\delta}{|\Lambda|})^{|\Lambda|}$. The probability of making at least one Type I error (i.e., FWER) is the complement of making no Type I errors, which is $1 - (1 - \dfrac{\delta}{|\Lambda|})^{|\Lambda|} \leq \delta$. Therefore, we prove that the familywise error rate is δ for Bonferroni correction. Thus, going back to the risk guarantee, we have:

$$\mathbb{P}\left[\sup_{\hat{\boldsymbol{\lambda}} \in \hat{\Lambda}_\alpha} \left\{ R\left(T_{\hat{\boldsymbol{\lambda}}, p_{\theta_l}}(x), y\right)\right\} \leq \alpha \right] \geq 1 - \delta \tag{19}$$

□

Remarks. To achieve conformal analysis (2), we follow the procedure in (Angelopoulos et al., 2021): (a) for each parameter configuration $\boldsymbol{\lambda}$ in the feasible region Λ, associate the null hypothesis: $\mathcal{H}_j : R(T_{\boldsymbol{\lambda}, p_{\theta_l}}) > \alpha$ (rejecting the null hypothesis implies controlling the risk below α with hyperparameter $\boldsymbol{\lambda}$), (b) for each null hypothesis, compute a finite-sample valid p-value p_j using Hoeffding-Bentkus inequality, and (c) return $\hat{\Lambda}_\alpha = \mathcal{A}(p_1, ..., p_{|\Lambda|})$, where $\mathcal{A}$ is an algorithm that controls the family-wise error rate (FWER). Essentially, FWER controls the error by union bounds over the hyperparameter space. The Bonferroni correction yields $\hat{\Lambda}_\alpha = \{\hat{\boldsymbol{\lambda}}_j : \delta_j \leq \delta / |\Lambda|\}$. The graph-based search in Alg. 2 dynamically assigns the error levels and yields a tighter certification. Specifically, we maintain a directed graph with nodes denoting the error rate of the parameter configuration and edges denoting the correlations between two parameters. The correlations can be instantiated randomly. We first randomly assign error rates to all feasible parameter configurations, and then once we search for one valid parameter with a smaller p-value than the assigned error rate, we will add the parameter to the valid set and propagate the excessive error rate to other nodes. The procedure repeats until no valid parameter can be found.

Computation of p-values. Due to the duality between p-values and confidence intervals (Bates et al., 2021), we compute the p-value by applying the Hoeffding-Bentkus inequality as the following:

$$p_j = \min\left(\exp\left\{-N_{\text{cal}} h\left(\hat{\mathbb{E}}[R(T_{\boldsymbol{\lambda}_j, p_{\theta_l}}(x), y)], \hat{\alpha}\right)\right\}, e\mathbb{P}\left[\text{Bin}(N_{\text{cal}}, \hat{\alpha}) \leq \left\lceil N_{\text{cal}} \hat{\mathbb{E}}[R(T_{\boldsymbol{\lambda}_j, p_{\theta_l}}(x), y)] \right\rceil\right]\right), \tag{20}$$

where $\hat{\mathbb{E}}[R(T_{\boldsymbol{\lambda}_j, p_{\theta_l}}(x), y)]$ denotes the empirical mean risk with configuration $\boldsymbol{\lambda}_j$ $(j \in \{1, 2, .., |\Lambda|\})$.

E. Grammian generalization bound

Lemma E.1 ((Weber et al., 2022)). *Let $\mathcal{D}$ and $\mathcal{Q}$ denote two distributions supported on $\mathcal{X} \times \mathcal{Y}$. Let $h_\theta : \mathcal{X} \mapsto \mathcal{Y}$ be any black-box pretrained model. Consider any risk/loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ such that $0 \leq \ell(h_\theta(X), Y) \leq T$, then*

$$\begin{aligned}
&\max_{\mathcal{Q},\theta} \mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\ell(h_\theta(X), Y)] \quad \text{s.t.} \quad H(\mathcal{D}, \mathcal{Q}) \leq \rho \\
\leq &\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h_\theta(X), Y)] + 2C_\rho \sqrt{\mathbb{V}_{(X,Y)\sim\mathcal{D}}[\ell(h_\theta(X), Y)]} + \\
&\rho^2(2-\rho^2)\left(T - \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h_\theta(X), Y)] - \frac{\mathbb{V}_{(X,Y)\sim\mathcal{D}}[\ell(h_\theta(X), Y)]}{T - \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h_\theta(X), Y)]}\right),
\end{aligned} \tag{22}$$

where $C_\rho = \sqrt{\rho^2(1-\rho^2)^2(2-\rho^2)}$, for any given distance bound $\rho > 0$ that satisfies

$$\rho^2 \leq 1 - \left(1 + \frac{(T - \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h_\theta(X), Y)])^2}{\mathbb{V}_{(X,Y)\sim\mathcal{D}}[\ell(h_\theta(X), Y)]}\right)^{-1/2}. \tag{23}$$

This theorem provides a closed-form expression that upper bounds the risk of $h_\theta(\cdot)$ on shifted distribution (namely $\mathbb{E}_{\mathcal{Q}}[\ell(h_\theta(X), Y)]$), given bounded Hellinger distance $H(\mathcal{D}, \mathcal{Q})$ and the mean E and variance V of loss on $\mathcal{D}$ under two mild conditions: (1) the function is positive and bounded (denote the upper bound by T); and (2) the distance $H(\mathcal{D}, \mathcal{Q})$ is not too large (specifically, $H(\mathcal{D}, \mathcal{Q})^2 \leq \bar{\gamma}^2 := 1 - (1 + (T-E)^2/V)^{-\frac{1}{2}}$). Since Lem. E.1 holds for arbitrary models and risk functions $\ell(h_\theta(\cdot), \cdot)$ as long as the function value is bounded by $[0, T]$, using Lem. E.1 allows us to provide a generic and succinct retrieval analysis and conformal risk certificate in Thm. 2 and Thm. 3 that holds for generic models including DNNs without engaging complex model architectures. Indeed, we only need to query the mean and variance under $\mathcal{D}$ for the retrieval model to compute the certificate in Lem. E.1.

F. Proofs and detailed remarks in Sec. 5

F.1. Detailed remark of Def. 1

Remarks. (R1) Note that a retrieval model with random initialization can achieve $\mathbb{E}[s_\theta(x, x^+)] = \mathbb{E}[s_\theta(x, x^-)]$ asymptotically. We only assume a V_{rag}-retrieval model that differentiates positive examples from negative examples slightly better than random with the condition $\mathbb{E}[s_\theta(x, x^+) - s_\theta(x, x^-)] > 0$. (R2) We also only assume a moderate stability characterized with bounded variance $\mathbb{V}[s_\theta(x, x^+) - s_\theta(x, x^-)]^{1/2} < \ln(\exp\{-L_\tau\}/(1 - \exp\{-L_\tau\}))$, which implicitly assumes a moderate generalization of the retrieval model by the variance-generalization connection (Lam, 2016; Gotoh et al., 2018; Namkoong & Duchi, 2017). The moderate generalization ability of the retrieval model is essential since we do not assume that the knowledge base distribution is identical to the calibration/test distribution. Since the frequently adopted cosine similarity is bounded in $[-1, 1]$, the variance of difference in similarities $\mathbb{V}[s_\theta(x, x^+) - s_\theta(x, x^-)]$ is upper bounded by 1 (derived by the variance bound $\mathbb{V}[X] \leq (b-a)^2/4$ for random variable X bounded in $[a, b]$). Therefore, as long as L_τ is small such that $\ln(\exp\{-L_\tau\}/(1 - \exp\{-L_\tau\})) > 1$, the variance requirement can be automatically satisfied. (R3) We define the V_{rag}-retrieval model with a commonly used contrastive loss for retrieval model training (Wang et al., 2023c; Rubin et al., 2021), but we also allow for flexibility in considering other types of contrastive loss such as the triplet loss (Hermans et al., 2017). Towards that, we only need to connect a different loss formulation to the denominator of the formulation of V_{rag} (i.e., $\ln(\exp\{-L_\tau\}/(1 - \exp\{-L_\tau\}))$).

F.2. Detailed remark of Def. 2

Remarks. (R1) d^+ represents the attention score of positive pairs and quantifies the utility of embedding matrix W_E, key matrix W_K, and query matrix W_Q of the transformer. Note that the attention scores are always non-negative after the Softmax activation, so $d^+ > 0$ usually holds. (R2) $\int_{-1}^{1} \Phi_M(v)dv$ characterizes the quality of the embedding matrix W_E, value matrix W_V, and projection matrix W_P. We have a Softmax normalization for output probability vectors and M is bounded in $[-1, 1]$, so the integral over $[-1, 1]$ traverses the support of M. Predictions aligning well with the reference text induce a generally small value of M and thus a large integral of CDF $\int_{-1}^{1} \Phi_M(v)dv$. Also, note that a random prediction margin M_{rand} with a uniform distribution over $[-1, 1]$ satisfies $\int_{-1}^{1} \Phi_{M_{\text{rand}}}(v)dv = 1$. Therefore, $\int_{-1}^{1} \Phi_M(v)dv > 1 = \int_{-1}^{1} \Phi_{M_{\text{rand}}}(v)dv$ only assumes a better-than-random prediction margin.

F.3. Proof and detailed remark of Prop. 3

Remarks. Eq. (6) shows the lower bound of the expectation of the retrieved positive examples by the retrieval model. (R1) For a sufficiently large number of instances in the external knowledge base N_{ext} (a typical scenario in practice), the lower bound approximately scales with $0.9N_{\text{rag}}$ and this scaling occurs at an exponential rate with respect to N_{ext}. These findings imply that with a large external knowledge base, a large ratio of the retrieved examples is positive (i.e., with the same groundtruth output), which is valuable as the retrieved positive examples which share similar semantic meanings as the query examples can improve in-context learning of LLMs (Min et al., 2022; Wang et al., 2022a). To formulate this observation in a rigorous way, we theoretically show the benefits of retrieved positive in-context examples in achieving low conformal generation risks in Thm. 1. (R2) The lower bound of the expected retrieved positive examples also correlates with the balance in the external knowledge base (i.e., $r^{(c)}_{\text{ext}}$). The correlation implies that if the knowledge base is highly long-tailed such that samples of certain reference texts are rare (i.e., $r^{(c)}_{\text{ext}}$ is small), we require a larger sample size of knowledge base N_{ext} to compensate for the long-tail distribution and achieve comparable retrieval quality. (R3) The bound also shows that a low-variance retrieval model (i.e., a small V_{rag}) can generalize well to test distribution and induce a better retrieval quality.

Proof sketch. We first formulate the expectation of similarity difference between positive pairs and negative pairs $\mathbb{E}[s_{\theta_r}(x,x^+) - s_{\theta_r}(x,x^-)]$ as a function of the contrastive loss of the retrieval model L_τ. Then, we apply Chebyshev's inequality to upper bound the failure probability $\mathbb{P}[s_\theta(x,x^+) < s_\theta(x,x^-)]$ as a function of V_{rag}. We then derive a lower bound of the number of retrieved positive examples, which follows a binomial distribution with N_{rag} trials and the failure rate as a function of V_{rag}. We finally correct the bound with the finite-sample errors of the knowledge base by the tail bound of categorical distribution.

Proof of Prop. 3. Let $\mathcal{D}$ be the data distribution, which is also the training distribution of the retrieval model, conformal calibration distribution, and test distribution. For a sample $(x,y) \sim \mathcal{D}$, we denote $\mathcal{D}^+_{\text{ext}}(x)$, $\mathcal{D}^-_{\text{ext}}(x)$ be the distribution of positive examples (with the same groundtruth output y) and negative examples (with different groundtruth output to y) of sample x in the external knowledge base. Then we can formulate the expectation of contrastive loss of the retrieval model as:

$$\begin{aligned} L_\tau &= \mathbb{E}_{x\sim\mathcal{D}, x^+\sim\mathcal{D}^+_{\text{ext}}(x), x^-\sim\mathcal{D}^-_{\text{ext}}(x)} \left[\mathcal{L}_{\text{contrastive}}(x,x^+,x^-)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}, x^+\sim\mathcal{D}^+_{\text{ext}}(x), x^-\sim\mathcal{D}^-_{\text{ext}}(x)} \left[-\log \frac{\exp\left\{s_\theta(x,x^+)\right\}}{\exp\left\{s_\theta(x,x^+)\right\} + \exp\left\{s_\theta(x,x^-)\right\}}\right], \end{aligned} \tag{24}$$

which is equivalent to

$$\mathbb{E}_{x\sim\mathcal{D}, x^+\sim\mathcal{D}^+_{\text{ext}}(x), x^-\sim\mathcal{D}^-_{\text{ext}}(x)} \left[s_\theta(x,x^+) - s_\theta(x,x^-)\right] = \ln \frac{\exp\left\{-L_\tau\right\}}{1-\exp\left\{-L_\tau\right\}} > 0. \tag{25}$$

Then we can apply Chebyshev's inequality (Saw et al., 1984) to the random variable $s_\theta(x,x^+) - s_\theta(x,x^-)$ and get the following:

$$\mathbb{P}\left[s_\theta(x,x^+) < s_\theta(x,x^-)\right] = \mathbb{P}_{x\sim\mathcal{D}, x^+\sim\mathcal{D}^+_{\text{ext}}(x), x^-\sim\mathcal{D}^-_{\text{ext}}(x)} \left[s_\theta(x,x^+) - s_\theta(x,x^-) < 0\right] \tag{26}$$

$$\leq \frac{\mathbb{V}\left[s_\theta(x,x^+) - s_\theta(x,x^-)\right]}{\mathbb{E}\left[s_\theta(x,x^+) - s_\theta(x,x^-)\right]^2} \tag{27}$$

$$\leq \left(\frac{\sqrt{\mathbb{V}\left[s_\theta(x,x^+) - s_\theta(x,x^-)\right]}}{\mathbb{E}\left[s_\theta(x,x^+) - s_\theta(x,x^-)\right]}\right)^2 = V^2_{\text{rag}} < 1. \tag{28}$$

We first focus on one demonstration example Z_r retrieved by $s_\theta(\cdot,\cdot)$. According to the retrieval mechanism, the example Z_r has the highest similarity (measured by $s_\theta(\cdot,\cdot)$) to the query sample Z_q. Recall that $r^{(c)}_{\text{cal}}$ and $r^{(c)}_{\text{ext}}$ be the event probability of the c-th category in the categorical calibration distribution and categorical external knowledge distribution. Let $\boldsymbol{r}_{\text{ext}} = \left[r^{(1)}_{\text{ext}}, r^{(2)}_{\text{ext}}, ..., r^{(C)}_{\text{ext}}\right]$. Since we only have N_{ext} finite sample drawn from $\mathcal{D}_{\text{ext}}$ in the external knowledge base in practice, we notate the empirical categorical portions as $\hat{\boldsymbol{q}}_{\text{ext}} = \left[\hat{r}^{(1)}_{\text{ext}}, \hat{r}^{(2)}_{\text{ext}}, ..., \hat{r}^{(C)}_{\text{ext}}\right]$, where $\hat{r}^{(c)}_{\text{ext}}$ $(c \in \{1,..,C\})$ represents the portion of samples with grountruth text $c \in \mathcal{Y}$ in the external knowledge base. Then we can apply the concentration bound of categorical distribution as (Agrawal & Jia, 2017):

$$\mathbb{P}\left[\|\hat{\boldsymbol{r}}_{\text{ext}} - \boldsymbol{r}_{\text{ext}}\|_1 \geq \frac{\sqrt{2\ln(1/\delta_{\text{ext}})}}{N_{\text{ext}}} \right] \leq \delta_{\text{ext}}, \tag{29}$$

where $\delta_{\text{ext}} > 0$ represents the confidence level of the tail bound. Then we can upper bound the probability that the groundtruth output $g(Z_r)$ of the retrieved sample Z_r is not equal to that of the query sample $g(Z_q)$ as the following:

$$\begin{aligned}
\mathbb{P}\left[g(Z_r) \notin g(Z_q)\right] &= \mathbb{P}_{Z_q\sim\mathcal{D}}\left[g(Z_r) \notin g(Z_q) \,\middle|\, s_\theta(Z_q, Z_r) \geq \max_{z\in\mathcal{D}_{\text{ext}}} s_\theta(z, Z_q) \right] &(30)\\
&= \mathbb{P}_{Z_q\sim\mathcal{D}}\left[\max_{Z^-\in\mathcal{D}^-_{\text{ext}}(Z_q)} s_\theta(Z^-, Z_q) \geq \max_{Z^+\in\mathcal{D}^+_{\text{ext}}(Z_q)} s_\theta(Z^+, Z_q) \right] &(31)\\
&= \mathbb{P}_{Z_q\sim\mathcal{D}}\left[s_\theta(Z^-, Z_q) \geq s_\theta(Z^+, Z_q),\ \forall Z^+ \in \mathcal{D}^+_{\text{ext}}(Z_q),\ \exists Z^- \in \mathcal{D}^-_{\text{ext}}(Z_q) \right] &(32)\\
&\leq \sum_{c=1}^{C} r^{(c)}_{\text{cal}} \left(1 - r^{(c)}_{\text{ext}}\right) N_{\text{ext}} \mathbb{P}\left[s_\theta(x, x^+) < s_\theta(x, x^-)\right]^{N_{\text{ext}} r^{(c)}_{\text{ext}}}, &(33)
\end{aligned}$$

where Eq. (33) is derived by applying the union bound. Considering finite-sample error of categorical distribution in Eq. (29) and combining Eq. (28), we finally have:

$$\begin{aligned}
\mathbb{P}\left[g(Z_r) \notin g(Z_q)\right] &\leq \sum_{c=1}^{C} r^{(c)}_{\text{cal}} \left(1 - r^{(c)}_{\text{ext}}\right) N_{\text{ext}} \mathbb{P}\left[s_\theta(x, x^+) < s_\theta(x, x^-)\right]^{N_{\text{ext}} r^{(c)}_{\text{ext}}} &(34)\\
&\leq \sum_{c=1}^{C} r^{(c)}_{\text{cal}} \left(1 - r^{(c)}_{\text{ext}} + \frac{\sqrt{2\ln(1/\delta_{\text{ext}})}}{N_{\text{ext}}}\right) N_{\text{ext}} V_{\text{rag}}^{0.5 N_{\text{ext}}\left(r^{(c)}_{\text{ext}} - \sqrt{2\ln(1/\delta_{\text{ext}})/N_{\text{ext}}}\right)}. &(35)
\end{aligned}$$

Since we assume that the retrieval model retrieves samples identically from the external knowledge base, the number of retrieved positive examples N_{pos} follows a Binomial distribution with N_{rag} trials and failure rate in Eq. (35). Therefore, we can lower bound the expectation of N_{pos} as the following:

$$\mathbb{E}\left[N_{\text{pos}}\right] \geq N_{\text{rag}}\left(1 - \delta_{\text{ext}}\right)\left(1 - \sum_{c=1}^{C} r^{(c)}_{\text{cal}} \left(1 - r^{(c)}_{\text{ext}} + \frac{\sqrt{2\ln(1/\delta_{\text{ext}})}}{N_{\text{ext}}}\right) N_{\text{ext}} V_{\text{rag}}^{0.5 N_{\text{ext}}\left(r^{(c)}_{\text{ext}} - \sqrt{2\ln(1/\delta_{\text{ext}})/N_{\text{ext}}}\right)}\right), \tag{36}$$

which holds for any $\delta_{\text{ext}} > 0$. Therefore, letting $\delta_{\text{ext}} = 0.1$, we can finally conclude that:

$$\mathbb{E}\left[N_{\text{pos}}\right] \geq \frac{9}{10} N_{\text{rag}}\left(1 - \sum_{c=1}^{C} r^{(c)}_{\text{cal}} \left(N_{\text{ext}} - r^{(c)}_{\text{ext}} N_{\text{ext}} + \sqrt{2\ln 10}\right) V_{\text{rag}}^{0.5\left(r^{(c)}_{\text{ext}} N_{\text{ext}} - \sqrt{2\ln 10}\right)}\right). \tag{37}$$

□

F.4. Proof and detailed remark of Thm. 1

Remarks. In Thm. 1, we theoretically show that the conformal generation risk with RAG $\hat{\alpha}_{\text{rag}}$ is smaller than the risk without RAG $\hat{\alpha}$ with a high probability. (R1) We can observe that the probability monotonically increases in the sample size of the calibration set N_{cal}, the size of retrieved examples N_{rag}, and the number of instances in the external knowledge base N_{ext}. In particular, a large N_{cal} reduces the finite-sample error during the calibration and induces a better approximation of the true generation risk with the empirical risk on the calibration set. A large N_{rag} and N_{ext} brings in related background information from a more knowledge-intensive knowledge base, which enhances the quality of generations augmented by retrieval. (R2) Furthermore, the probability $\mathbb{P}\left[\hat{\alpha}_{\text{rag}} < \hat{\alpha}\right]$ increases with the increase in transformer's quality, which is quantified by the attention scores for a positive pair (i.e., d^+) and the prediction capability (without RAG) (i.e., $\int_{-1}^{1} \Phi_M(v)dv - 1$). Since $1 - \Phi_M(0)$ represents the population risk without RAG, the difference of the prediction margin CDF $\Phi_M(\cdot)$ (monotonically increasing) directly characterizes the benefit of generation quality with RAG. The quality improvement provided by RAG also exponentially induces a larger probability of reducing the conformal generation risk of a single LLM. The transformer uncertainty p_t also decreases exponentially with a large number of retrieved examples, indicating that more examples retrieved by a good retrieval model benefit a lower conformal generation risk. (R3) The retrieval model uncertainty p_r decreases with a low-variance retrieval model (small V_{rag}), which can generalize well to test distribution. (R4) We focus on the conformal generation risk formulated in Prop. 1, but we can easily adapt the results to any other forms of conformal risks. Since we build the connection between the empirical risk $\hat{R}$ to the retrieval model quality, we only need to directly connect the empirical risk to the certified generation risk via the explicit function to achieve end-to-end certification. (R5) We define the positive pairs as examples sharing the same semantic meanings of reference texts for simplicity of the certification results. In the certification framework, we can also consider a relaxed definition of positive pairs by the similarity of reference texts in the embedding space. Similarly, the examples with high similarity of reference texts to the query example will induce high attention scores and benefit the generation with the attention mechanism.

Proof sketch. We decompose the one-layer self-attention mapping as the combinations of attention with positive examples and attention with negative examples. Based on the explicit formulation, we then derive a lower bound of the logit difference of the ground truth token by taking a lower bound of the number of positive examples (derived from Prop. 3) and the attention scores with positive examples. Next, we get a lower bound of the risk difference between the transformer without RAG and the RAG model. Finally, we apply Hoeffding's inequality to derive a lower bound of the difference in empirical risks, and accordingly, conformal risk bounds. Applying union bounds over all uncertainty levels concludes the proof.

Proof of Thm. 1. From Prop. 3, we prove that:

$$\mathbb{E}\left[N_{\text{pos}}\right] \geq \underline{N}_{\text{pos}} := \frac{9}{10} N_{\text{rag}} \left(1 - \sum_{c=1}^{C} r_{\text{cal}}^{(c)} \left(N_{\text{ext}} - r_{\text{ext}}^{(c)} N_{\text{ext}} + \sqrt{2 \ln 10} \right) V_{\text{rag}}^{0.5\left(r_{\text{ext}}^{(c)} N_{\text{ext}} - \sqrt{2 \ln 10} \right)} \right). \tag{38}$$

Since N_{pos} is a binomial random variable with N_{rag} trials, we have the upper bound of the variance $\mathbb{V}[N_{\text{pos}}] \leq \dfrac{N_{\text{rag}}}{4}$. Applying Chebyshev's inequality to the random variable N_{pos}, the following holds $\forall n_{\text{pos}} < \underline{N}_{\text{pos}}$:

$$\mathbb{P}\left[N_{\text{pos}} \geq n_{\text{pos}}\right] \geq 1 - \frac{\mathbb{V}[N_{\text{rag}}]}{(\underline{N}_{\text{pos}} - n_{\text{pos}})^2} \geq 1 - \frac{N_{\text{rag}}}{4(\underline{N}_{\text{pos}} - n_{\text{pos}})^2}, \tag{39}$$

which implicates that we can do the analysis with $N_{\text{pos}} \geq n_{\text{pos}}$ with probability $1 - \dfrac{\mathbb{V}[N_{\text{rag}}]}{(\underline{N}_{\text{pos}} - n_{\text{pos}})^2}$.

Since the query example is encoded at the last position of the sequence (i.e., $N_{\text{rag}} + 1$-th position), we let $N := N_{\text{rag}} + 1$ for ease of notation. We denote the probability vector at the position as $O^{(\text{rag})}(\boldsymbol{q})$ with RAG and $O(\boldsymbol{q}_N)$ without RAG (without RAG, the input text is only the query sample $\boldsymbol{q}_N$). Recall the single-layer self-attention transformer:

$$O^{(\text{rag})}(\boldsymbol{q}) = W_P \left\{ W_V W_E \boldsymbol{q}_N + (W_V W_E \boldsymbol{q}) \sigma \left((W_K W_E \boldsymbol{q})^T W_Q W_E \boldsymbol{q}_N \right) \right\}. \tag{40}$$

Note that each raw vector of the linear projection matrix (fully connected layer) represents the prototype embedding denoted as $\boldsymbol{p}_c$ of the corresponding groundtruth output c. Formally, we have $W_P := [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_C]^T$. Recall that we denote $g(\boldsymbol{q}_N)$ as the groundtruth output of example $\boldsymbol{q}_N$. Then we can reformulate Eq. (40) as the following:

$$O^{(\text{rag})}(\boldsymbol{q}) = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_C]^T \left\{ W_V W_E \boldsymbol{q}_N + (W_V W_E \boldsymbol{q}) \sigma \left((W_K W_E \boldsymbol{q})^T W_Q W_E \boldsymbol{q}_N \right) \right\}, \tag{41}$$

which indicates the formulation of $O_c^{(\text{rag})}(\boldsymbol{q})$ denoting the probability of query sample $\boldsymbol{q}_N$ being with groundtruth output c:

$$O_c^{(\text{rag})}(\boldsymbol{q}) = \boldsymbol{p}_c^T \left\{ W_V W_E \boldsymbol{q}_N + (W_V W_E \boldsymbol{q}) \sigma \left((W_K W_E \boldsymbol{q})^T W_Q W_E \boldsymbol{q}_N \right) \right\}. \tag{42}$$

We can also similarly formulate the prediction without RAG:

$$O_c(\boldsymbol{q}) = \boldsymbol{p}_c^T W_V W_E \boldsymbol{q}_N. \tag{43}$$

Then we will focus on analyzing $O_c^{(\text{rag})}(\boldsymbol{q})$ and connect it with the quantities of characterizing the quality of the transformer (i.e., d^+, d^-, t^+, t^-) and the quality of retrieved examples. Towards that, we let $\mathcal{I}^+(\boldsymbol{q}_N)$ be the index set of retrieved examples with the same groundtruth output as $\boldsymbol{q}_N$ (i.e., positive examples), and $\mathcal{I}^-(\boldsymbol{q}_N)$ be the index set of retrieved examples with the different groundtruth output to $\boldsymbol{q}_N$ (i.e., negative examples). Then we can reformulate Eq. (42) as the following:

$$\begin{aligned}
O_c^{(\text{rag})}(\boldsymbol{q}) =& \boldsymbol{p}_c^T W_V W_E \boldsymbol{q}_N + \boldsymbol{p}_c^T (W_V W_E \boldsymbol{q}) \sigma\left((W_K W_E \boldsymbol{q})^T W_Q W_E \boldsymbol{q}_N \right) \\
=& \boldsymbol{p}_c^T W_V W_E \boldsymbol{q}_N + \sum_{i^+ \in \mathcal{I}^+(\boldsymbol{q}_N)} \sigma\left((W_K W_E \boldsymbol{q}_{i^+}) W_Q W_E \boldsymbol{q}_N \right) \boldsymbol{p}_c^T (W_V W_E \boldsymbol{q}_{i^+}) \\
&+ \sum_{i^- \in \mathcal{I}^-(\boldsymbol{q}_N)} \sigma\left((W_K W_E \boldsymbol{q}_{i^-}) W_Q W_E \boldsymbol{q}_N \right) \boldsymbol{p}_c^T (W_V W_E \boldsymbol{q}_{i^-})
\end{aligned} \tag{44}$$

Recall that we have the following assumption:

$$\sigma\left((W_K W_E \boldsymbol{q}_i)^T (W_Q W_E \boldsymbol{q}_j) \right) \geq d^+ > 0, \quad \text{for } g(\boldsymbol{q}_i) = g(\boldsymbol{q}_j), \tag{45}$$

We denote N_{pos} as the number of positive retrieved examples to the query sample $\boldsymbol{q}_N$ and the lower bound of it n_{pos} with probability $1 - \dfrac{\mathbb{V}[N_{\text{rag}}]}{(\underline{N}_{\text{pos}} - n_{\text{pos}})^2}$ according to Eq. (39). Note that the attention scores are normalized by Softmax with the summation of them being 1. By Eq. (44), $\forall c \neq g(\boldsymbol{q}_N)$, we have:

$$\begin{aligned}
&\mathbb{E}\left[O_{g(\boldsymbol{q}_N)}^{(\text{rag})}(\boldsymbol{q}) - O_c^{(\text{rag})}(\boldsymbol{q}) - (\boldsymbol{p}_{g(\boldsymbol{q}_N)}^T W_V W_E \boldsymbol{q}_N - \boldsymbol{p}_c^T W_V W_E \boldsymbol{q}_N) \right] \\
=& \mathbb{E}\left[\sum_{i^+ \in \mathcal{I}^+(\boldsymbol{q}_N)} \sigma\left((W_K W_E \boldsymbol{q}_{i^+}) W_Q W_E \boldsymbol{q}_N \right) (\boldsymbol{p}_{g(\boldsymbol{q}_N)} - \boldsymbol{p}_c)^T (W_V W_E \boldsymbol{q}_{i^+}) \right. \\
&\left. + \sum_{i^- \in \mathcal{I}^-(\boldsymbol{q}_N)} \sigma\left((W_K W_E \boldsymbol{q}_{i^-}) W_Q W_E \boldsymbol{q}_N \right) (\boldsymbol{p}_{g(\boldsymbol{q}_N)} - \boldsymbol{p}_c)^T (W_V W_E \boldsymbol{q}_{i^-}) \right] \\
\geq& \mathbb{E}\left[d^+ \sum_{i^+ \in \mathcal{I}^+(\boldsymbol{q}_N)} (\boldsymbol{p}_{g(\boldsymbol{q}_N)} - \boldsymbol{p}_c)^T (W_V W_E \boldsymbol{q}_{i^+}) + (1 - n_{\text{pos}} d^+) \left(- \sum_{i^+ \in \mathcal{I}^+(\boldsymbol{q}_N)} (\boldsymbol{p}_{g(\boldsymbol{q}_N)} - \boldsymbol{p}_c)^T (W_V W_E \boldsymbol{q}_{i^+}) \right) \right] \\
\geq& \left((n_{\text{pos}} + 1) d^+ - 1 \right) \mathbb{E}\left[\sum_{i^+ \in \mathcal{I}^+(\boldsymbol{q}_N)} (\boldsymbol{p}_{g(\boldsymbol{q}_N)} - \boldsymbol{p}_c)^T (W_V W_E \boldsymbol{q}_{i^+}) \right] \\
\geq& \left((n_{\text{pos}} + 1) d^+ - 1 \right) n_{\text{pos}} \mathbb{E}\left[\boldsymbol{p}_{g(\boldsymbol{q}_N)}^T W_V W_E \boldsymbol{q}_N - \boldsymbol{p}_c^T W_V W_E \boldsymbol{q}_N \right] \\
\geq& \left((n_{\text{pos}} + 1) d^+ - 1 \right) n_{\text{pos}} \mathbb{E}\left[\boldsymbol{p}_{g(\boldsymbol{q}_N)}^T W_V W_E \boldsymbol{q}_N - \max_{c \neq g(\boldsymbol{q}_N)} \boldsymbol{p}_c^T W_V W_E \boldsymbol{q}_N \right]
\end{aligned} \tag{46}$$

Recall that $\Phi_M(\cdot)$ is the CDF function of the random variable of prediction marigin $\max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_N)}(\boldsymbol{q})$ such that $\Phi_M(v) = \mathbb{P}[\max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) < v]$. Since the output probability of the transformer is bounded in $[0, 1]$, we define a new random variable $X = \max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) + 1$ with $P[0 \leq X \leq 2] = 1$. Then we have the following:

$$\mathbb{E}\left[O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - \max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) \right] = 1 - \mathbb{E}\left[X \right] \tag{47}$$

$$= 1 - \int_0^2 \left(1 - \Phi_X(x) \right) dx \tag{48}$$

$$= 1 - \int_{-1}^1 \left(1 - \Phi_M(v) \right) dv \tag{49}$$

$$= \int_{-1}^1 \Phi_M(v) dv - 1 \tag{50}$$

Note that from Eq. (43), we have $O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - O_c(\boldsymbol{q}) = \boldsymbol{p}_{g(\boldsymbol{q}_N)}^T W_V W_E \boldsymbol{q}_N - \boldsymbol{p}_c^T W_V W_E \boldsymbol{q}_N$. Combining Eq. (46) and Eq. (50), the following holds $\forall c \neq g(\boldsymbol{q}_N)$:

$$\mathbb{E}\left[O^{(\text{rag})}_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - O^{(\text{rag})}_c(\boldsymbol{q})\right] \geq \mathbb{E}\left[O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - O_c(\boldsymbol{q})\right] + \left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right). \tag{51}$$

Letting $c^* = \arg\max_{c \neq g(\boldsymbol{q}_N)} O^{(\text{rag})}_c(\boldsymbol{q})$, we have:

$$\begin{aligned}
\mathbb{E}\left[O^{(\text{rag})}_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - \max_{c \neq g(\boldsymbol{q}_N)} O^{(\text{rag})}_{c*}(\boldsymbol{q})\right] &= \mathbb{E}\left[O^{(\text{rag})}_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - O^{(\text{rag})}_{c*}(\boldsymbol{q})\right] \\
&\geq \mathbb{E}\left[O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - O_{c^*}(\boldsymbol{q})\right] + \left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right) \\
&\geq \mathbb{E}\left[O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - \max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q})\right] + \left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right),
\end{aligned} \tag{52}$$

which implies the following:

$$\mathbb{E}\left[O^{(\text{rag})}_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - \max_{c \neq g(\boldsymbol{q}_N)} O^{(\text{rag})}_c(\boldsymbol{q})\right] - \mathbb{E}\left[O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) - \max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q})\right] \geq \left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right), \tag{53}$$

which is equivalent to the following:

$$\mathbb{E}\left[\max_{c \neq g(\boldsymbol{q}_N)} O^{(\text{rag})}_c(\boldsymbol{q}) - O^{(\text{rag})}_{g(\boldsymbol{q}_N)}(\boldsymbol{q})\right] \leq \mathbb{E}\left[\max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_N)}(\boldsymbol{q})\right] - \left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right), \tag{54}$$

Recall that we define the risk as $1 -$ Accuracy and notate R and R_{rag} as the risk without RAG and with RAG, respectively. Then we have:

$$\begin{aligned}
\mathbb{E}[R] &= 1 - \mathbb{E}\left[\mathbb{I}\left[\max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) < 0\right]\right] && (55)\\
&= 1 - \mathbb{P}\left[\max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) < 0\right] && (56)\\
&= 1 - \Phi_M(0). && (57)
\end{aligned}$$

Similarly for the risk with RAG R_{rag}, we have:

$$\begin{aligned}
\mathbb{E}[R_{\text{rag}}] &= 1 - \mathbb{E}\left[\mathbb{I}\left[\max_{c \neq g(\boldsymbol{q}_N)} O^{(\text{rag})}_c(\boldsymbol{q}) - O^{(\text{rag})}_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) < 0\right]\right] && (58)\\
&= 1 - \mathbb{P}\left[\max_{c \neq g(\boldsymbol{q}_N)} O^{(\text{rag})}_c(\boldsymbol{q}) - O^{(\text{rag})}_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) < 0\right] && (59)\\
&\leq 1 - \mathbb{P}\left[\max_{c \neq g(\boldsymbol{q}_N)} O_c(\boldsymbol{q}) - O_{g(\boldsymbol{q}_N)}(\boldsymbol{q}) < \left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right)\right] && (60)\\
&\leq 1 - \Phi_M\left(\left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right)\right), && (61)
\end{aligned}$$

where Eq. (60) holds by applying Eq. (54). Therefore, combining Eq. (57) and Eq. (61), the following holds:

$$\mathbb{E}[R - R_{\text{rag}}] \geq \Phi_M\left(\left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v)dv - 1\right)\right) - \Phi_M(0). \tag{62}$$

Let $n_{\text{rag}} = N_{\text{rag}}/2$. Combining Eq. (62) and Eq. (39), we get that if $\underline{N}_{\text{pos}} > N_{\text{rag}}/2 > 1/d^+$, with probability $1 -$

$\frac{N_{\text{rag}}}{4(\underline{N}_{\text{pos}}-N_{\text{rag}}/2)^2}$, we have:

$$\mathbb{E}\left[R-R_{\text{rag}}\right] \geq \Phi_M\left(\frac{d^+\left(\int_{-1}^{1}\Phi_M(v)dv-1\right)N_{\text{rag}}}{2}\right)-\Phi_M(0). \tag{63}$$

Let $R(Z)$ be the risk of Z sampled from the distribution $\mathcal{D}$. Define the empirical risk $\hat{R}$ and $\hat{R}_{\text{rag}}$ as the following:

$$\hat{R}=\frac{1}{N_{\text{cal}}}\sum_{i=1}^{N_{\text{cal}}}R(Z_i), \quad \hat{R}_{\text{rag}}=\frac{1}{N_{\text{cal}}}\sum_{i=1}^{N_{\text{cal}}}R_{\text{rag}}(Z_i) \tag{64}$$

According to Eq. (16), the statistical guarantee of conformal risk $\hat{\alpha}$ and $\hat{\alpha}_{\text{rag}}$ with confidence $1-\delta$ can be formulated as:

$$\hat{\alpha}=\min\left\{h^{-1}\left(\frac{1/\delta}{N_{\text{cal}}};\hat{R}\right),\Phi_{\text{bin}}^{-1}\left(\frac{\delta}{e};N_{\text{cal}},\hat{R}\right)\right\}, \tag{65}$$

$$\hat{\alpha}_{\text{rag}}=\min\left\{h^{-1}\left(\frac{1/\delta}{N_{\text{cal}}};\hat{R}_{\text{rag}}\right),\Phi_{\text{bin}}^{-1}\left(\frac{\delta}{e};N_{\text{cal}},\hat{R}_{\text{rag}}\right)\right\}, \tag{66}$$

where $\Phi_{\text{bin}}^{-1}(\cdot)$ is the inverse function of CDF of binomial distribution. Noting that $\hat{\alpha}$ is monotonically increasing in $\hat{R}$, the following holds by Hoeffding's inequality:

$$\mathbb{P}\left[\hat{\alpha}_{\text{rag}}<\hat{\alpha}\right]\geq\mathbb{P}\left[\hat{R}_{\text{rag}}<\hat{R}\right]\geq 1-\exp\left\{-2N_{\text{cal}}\mathbb{E}\left[R-R_{\text{rag}}\right]^2\right\}. \tag{67}$$

Combining Eq. (63) and Eq. (67) and using the union bound, under the condition that $\underline{N}_{\text{pos}}>N_{\text{rag}}/2>1/d^+$, we have:

$$\mathbb{P}\left[\hat{\alpha}_{\text{rag}}<\hat{\alpha}\right]\geq 1-\exp\left\{-2N_{\text{cal}}\left[\Phi_M\left(\frac{d^+\left(\int_{-1}^{1}\Phi_M(v)dv-1\right)N_{\text{rag}}}{2}\right)-\Phi_M(0)\right]^2\right\}-\frac{N_{\text{rag}}}{4(\underline{N}_{\text{pos}}-N_{\text{rag}}/2)^2}. \tag{68}$$

Let $r_{\text{ext}}^m := \min_{c\in\{1,..,C\}} r_{\text{ext}}^{(c)}$. Then we can show that one sufficient condition of $\underline{N}_{\text{pos}}>N_{\text{rag}}/2>1/d^+$ is that $N_{\text{ext}}>\frac{2\sqrt{2\ln 10}}{r_{\text{ext}}^m}$, $N_{\text{ext}}V_{\text{rag}}^{0.25 r_{\text{ext}}^m N_{\text{ext}}}<\frac{4}{9}$, and $N_{\text{rag}}>\frac{2}{d^+}$. Rearranging the terms and considering the sufficient condition in Eq. (68), we can finally conclude that, under the condition that $N_{\text{ext}}>\frac{2\sqrt{2\ln 10}}{r_{\text{ext}}^m}$, $N_{\text{ext}}V_{\text{rag}}^{0.25 r_{\text{ext}}^m N_{\text{ext}}}<\frac{4}{9}$, and $N_{\text{rag}}>\frac{2}{d^+}$, the following holds:

$$\begin{aligned}\mathbb{P}\left[\hat{\alpha}_{\text{rag}}<\hat{\alpha}\right]\geq & 1-\exp\left\{-2N_{\text{cal}}\left[\Phi_M\left(\frac{d^+\left(\int_{-1}^{1}\Phi_M(v)dv-1\right)N_{\text{rag}}}{2}\right)-\Phi_M(0)\right]^2\right\} \\ & -\frac{25}{N_{\text{rag}}}\left(4-9\sum_{c=1}^{C}r_{\text{cal}}^{(c)}\left(N_{\text{ext}}-r_{\text{ext}}^{(c)}N_{\text{ext}}+\sqrt{2\ln 10}\right)V_{\text{rag}}^{0.5\left(r_{\text{ext}}^{(c)}N_{\text{ext}}-\sqrt{2\ln 10}\right)}\right)^{-2}.\end{aligned} \tag{69}$$

□

G. Cor. 1: Asymtotic result of Thm. 1

Corollary 1 (Thm. 1 with a sufficiently large external knwoledge base)**.** *Under the same conditions as Thm. 1, suppose that we have a* ***sufficiently large sample size*** *N_{ext} in the external knowledge base. We then have the following guarantee:*

$$\mathbb{P}\left[\hat{\alpha}_{rag} < \hat{\alpha}\right] \geq 1 - p_t - \frac{25}{16 N_{rag}}, \tag{70}$$

where p_t is the uncertainty induced by the quality of the transformer as formulated in Eq. (7), and $\hat{\alpha}_{rag}$ and $\hat{\alpha}$ are the conformal generation risks with and without RAG, respectively.

Proof of Cor. 1. The proof directly follows that of Thm. 1. Considering Eq. (7) in the asymptoptic limit (i.e., $N_{\text{ext}} \to +\infty$), we obtain the formulation in Eq. (70). □

Remarks. (R1) Cor. 1 shows that the conformal generation risk of transformer with RAG $\hat{\alpha}_{\text{rag}}$ is smaller than that of without RAG $\hat{\alpha}$ with high probability (RHS of Eq. (70)), which asymptotically approaches 1 with a sufficiently large sizes of the calibration set N_{cal} and retrieved augmented examples N_{rag}. (R2) In contrast to Thm. 1, the bound has no dependency on distributions in the external knowledge base $r^{(c)}_{\text{ext}}$ since a sufficiently large knowledge base can cover also rare examples. (R3) Compared to Thm. 1, the bound also has no dependency on the distribution of the calibration/test distribution $r^{(c)}_{\text{cal}}$, showing that a sufficiently large external knowledge base can better generalize to unknown test distributions. Additionally, the lower bound in Eq. (70) is tighter than that in Eq. (7), which demonstrates the benefit of the large external knowledge base.

H. Prop. 4: Retrieval quality analysis under distribution shifts

Under test-time distribution shifts, retrieval model quality declines. To derive conformal generation risk, we first examine the lower bound of retrieved positive examples.

Proposition 4 (Lower bound to the retrieved positive examples under test-time distribution shifts)**.** *Suppose that the potential test distribution $\mathcal{Q}$ is shifted from the original test distribution $\mathcal{D}$ with bounded Hellinger distance $\rho > 0$. Consider the same setup as Prop. 3 and a large external knowledge base where $N_{ext} > 2\sqrt{2 \ln 10} / \min_c r^{(c)}_{ext}$. We have:*

$$\mathbb{E}\left[N_{pos}\right] \geq \frac{9}{10} N_{rag} \left(1 - 1.5 N_{ext} V_{rag}(\rho)^{0.25\left(\min_c r^{(c)}_{ext} N_{ext}\right)}\right), \tag{71}$$

where $V_{rag}(\rho) := m(\rho) V_{rag}$ and

$$m(\rho) = \underbrace{\left(\frac{\sqrt{-6\rho^4 + 12\rho^2 + 1} - 4\rho(1-\rho^2)\sqrt{2-\rho^2}}{1 - 16\rho^2 + 8\rho^4}\right)^{-2}}_{\textit{retrieval model quality decay factor by distribution shifts}}.$$

Proof sketch. We first formulate the expectation of similarity difference between positive pairs and negative pairs $\mathbb{E}[s_{\theta_r}(x, x^+) - s_{\theta_r}(x, x^-)]$ as a function of the contrastive loss of the retrieval model L_τ. Then, we apply Chebyshev's inequality to upper bound the failure probability $\mathbb{P}[s_\theta(x, x^+) < s_\theta(x, x^-)]$ as a function of the variance and expectation of the similarity difference. Considering the distribution shifts, the variance and expectation bound can be derived via Grammian bound as (Weber et al., 2022). We then plug in the failure rate corrected by distribution shifts and follow the proof structure of Prop. 3.

Remarks. (R1) Different from Prop. 3, the quality of retrieval models under distribution shifts is decreased from V_{rag} to $V_{\text{rag}}(\rho)$ with a linear decay factor $m(\rho)$. As we require $V_{\text{rag}}(\rho) < 1$ to ensure high retrieval quality, large distribution shift radius ρ must be compensated by small V_{rag}. This is consistent with the existing observations that low-variance models can generalize better under distribution shifts (Lam, 2016; Gotoh et al., 2018; Namkoong & Duchi, 2017). (R2) Compared to Prop. 3, Prop. 4 removes the dependency on varying label portions $r^{(c)}_{\text{cal}}$ during distribution shifts, as long as the size of external knowledge base is sufficiently large N_{ext} to offset the worst-case long-tail distributions, a condition often met in practice with large knowledge bases.

Proof of Prop. 4. Let $\mathcal{D}$ be the data distribution, which is also the training distribution of the retrieval model and conformal calibration distribution. Let $\mathcal{Q}$ be the test distribution where the test samples are drawn from. $\mathcal{Q}$ is within Hellinger distance ρ from the distribution $\mathcal{D}$: $H(\mathcal{D}, \mathcal{Q}) \leq \rho$.

For a sample $(x, y) \sim \mathcal{D}$, we denote $\mathcal{D}^+_{\text{ext}}(x)$, $\mathcal{D}^-_{\text{ext}}(x)$ be the distribution of positive examples (with the same groundtruth output y) and negative examples (with different groundtruth output to y) of sample x in the external knowledge base. Then we can formulate the expectation of contrastive loss of the retrieval model as:

$$\begin{aligned} L_\tau &= \mathbb{E}_{x\sim\mathcal{D}, x^+\sim\mathcal{D}^+_{\text{ext}}(x), x^-\sim\mathcal{D}^-_{\text{ext}}(x)}\left[\mathcal{L}_{\text{contrastive}}(x, x^+, x^-)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}, x^+\sim\mathcal{D}^+_{\text{ext}}(x), x^-\sim\mathcal{D}^-_{\text{ext}}(x)}\left[-\log\frac{\exp\{s_\theta(x,x^+)\}}{\exp\{s_\theta(x,x^+)\}+\exp\{s_\theta(x,x^-)\}}\right]. \end{aligned} \tag{72}$$

We can apply Chebyshev's inequality (Saw et al., 1984) to the random variable $s_\theta(x,x^+) - s_\theta(x,x^-)$ and get the following:

$$\mathbb{P}\left[s_\theta(x,x^+) < s_\theta(x,x^-)\right] = \mathbb{P}_{x\sim\mathcal{Q}, x^+\sim\mathcal{D}^+_{\text{ext}}(x), x^-\sim\mathcal{D}^-_{\text{ext}}(x)}\left[s_\theta(x,x^+) - s_\theta(x,x^-) < 0\right] \tag{73}$$

$$\leq \frac{\mathbb{V}\left[s_\theta(x,x^+) - s_\theta(x,x^-)\right]}{\mathbb{E}\left[s_\theta(x,x^+) - s_\theta(x,x^-)\right]^2}. \tag{74}$$

For ease of notation, we notate random variable $S = s_\theta(x,x^+) - s_\theta(x,x^-)$. Then we have $\mathbb{P}[-2 \leq S \leq 2] = 1$, and $\mathbb{E}[S] \geq \underline{\mathbb{E}[S]} := \dfrac{\exp\{-L_\tau(\rho)\}}{1-\exp\{-L_\tau(\rho)\}}$. Note that $\mathbb{P}[0 \leq S^2 \leq 4] = 1$, we have $\mathbb{V}[S^2] \leq (4-0)^2/4 = 4$. Then by Lem. E.1, we have the following:

$$\mathbb{E}_\mathcal{Q}\left[S^2\right] \leq \mathbb{E}_\mathcal{D}\left[S^2\right] + \rho^2(2-\rho^2)\left(1-\mathbb{E}_\mathcal{D}\left[S^2\right]\right) + 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\mathbb{V}_\mathcal{D}\left[S^2\right]} \tag{75}$$

$$\leq (1-\rho^2)^2\mathbb{E}_\mathcal{D}\left[S^2\right] + \rho^2(2-\rho^2) + 4\rho(1-\rho^2)\sqrt{2-\rho^2}. \tag{76}$$

By applying the lower bound of expectation values in Theorem A.2 in (Weber et al., 2022),which is a straightforward variation of Lem. E.1, we have the following:

$$\mathbb{E}_\mathcal{Q}\left[S\right] \geq (1-\rho^2)^2\mathbb{E}_\mathcal{D}\left[S\right] - 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\mathbb{V}_\mathcal{D}\left[S\right]} + \rho^2(2-\rho^2)\frac{\mathbb{V}_\mathcal{D}\left[S\right]}{\mathbb{E}_\mathcal{D}\left[S\right]}. \tag{77}$$

Since we assume that $\mathbb{E}_\mathcal{D}\left[S\right] = \mathbb{E}_\mathcal{D}\left[s_\theta(x,x^+) - s_\theta(x,x^-)\right] > 0$, we have the following:

$$\mathbb{E}_\mathcal{Q}\left[S\right] \geq (1-\rho^2)^2\mathbb{E}_\mathcal{D}\left[S\right] - 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\mathbb{V}_\mathcal{D}\left[S\right]}. \tag{78}$$

Combining Eqs. (76) and (78), we have the following:

$$\frac{\mathbb{V}_\mathcal{Q}\left[S\right]}{\mathbb{E}_\mathcal{Q}\left[S\right]^2} = \frac{\mathbb{E}_\mathcal{Q}\left[S^2\right] - \mathbb{E}_\mathcal{Q}\left[S\right]^2}{\mathbb{E}_\mathcal{Q}\left[S\right]^2} \tag{79}$$

$$\leq \frac{(1-\rho^2)^2\left(\mathbb{V}_\mathcal{D}\left[S\right] + \mathbb{E}_\mathcal{D}\left[S\right]^2\right) + \rho^2(2-\rho^2) + 4\rho(1-\rho^2)\sqrt{2-\rho^2}}{\mathbb{E}_\mathcal{Q}\left[S\right]^2} - 1 \tag{80}$$

$$\leq \frac{(1-\rho^2)^2\left(\mathbb{V}_\mathcal{D}\left[S\right] + \mathbb{E}_\mathcal{D}\left[S\right]^2\right) + \rho^2(2-\rho^2) + 4\rho(1-\rho^2)\sqrt{2-\rho^2}}{\left((1-\rho^2)^2\mathbb{E}_\mathcal{D}\left[S\right] - 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\mathbb{V}_\mathcal{D}\left[S\right]}\right)^2} - 1. \tag{81}$$

Through some algebraic rearrangement, we can show that:

$$\frac{\mathbb{V}_\mathcal{Q}\left[S\right]}{\mathbb{E}_\mathcal{Q}\left[S\right]^2} \leq V_{\text{rag}}(\rho) := \frac{\mathbb{V}_\mathcal{D}\left[S\right]}{\mathbb{E}_\mathcal{D}\left[S\right]^2}\left(\frac{\sqrt{-6\rho^4 + 12\rho^2 + 1} - 4\rho(1-\rho^2)\sqrt{2-\rho^2}}{1 - 16\rho^2 + 8\rho^4}\right)^{-2}. \tag{82}$$

Therefore, one sufficient condition of $V_{\text{rag}}(\rho) < 1$ is that:

$$\sqrt{\mathbb{V}_\mathcal{D}\left[S\right]} \leq \left(\frac{\sqrt{-6\rho^4 + 12\rho^2 + 1} - 4\rho(1-\rho^2)\sqrt{2-\rho^2}}{1 - 16\rho^2 + 8\rho^4}\right)\mathbb{E}_\mathcal{D}\left[S\right] \tag{83}$$

$$= \left(\frac{\sqrt{-6\rho^4 + 12\rho^2 + 1} - 4\rho(1-\rho^2)\sqrt{2-\rho^2}}{1 - 16\rho^2 + 8\rho^4}\right)\ln\frac{\exp\{-L_\tau\}}{1-\exp\{-L_\tau\}} \tag{84}$$

Then, we follow a similar procedure as the proof of Prop. 3 to analyze the expected retrieved positive examples. We first focus on one demonstration example Z_r retrieved by $s_\theta(\cdot,\cdot)$. According to the retrieval mechanism, the example Z_r has the highest similarity (measured by $s_\theta(\cdot,\cdot)$) to the query sample Z_q. Recall that $r^{(c)}_{\text{test}}(\mathcal{Q})$ and $r^{(c)}_{\text{ext}}$ be the event probability of the c-th category in the categorical test distribution (i.e., $\mathcal{Q}$) and categorical external knowledge distribution. Let $\boldsymbol{r}_{\text{ext}} = \left[r^{(1)}_{\text{ext}}, r^{(2)}_{\text{ext}}, ..., r^{(C)}_{\text{ext}}\right]$. Since we only have N_{ext} finite sample drawn from $\mathcal{D}_{\text{ext}}$ in the external knowledge base in practice, we notate the empirical categorical portions as $\hat{\boldsymbol{q}}_{\text{ext}} = \left[\hat{r}^{(1)}_{\text{ext}}, \hat{r}^{(2)}_{\text{ext}}, ..., \hat{r}^{(C)}_{\text{ext}}\right]$, where $\hat{r}^{(c)}_{\text{ext}}$ ($c \in \{1, .., C\}$) represents the portion of samples with grountruth output c in the external knowledge base. Then we can apply the concentration bound of categorical distribution as (Agrawal & Jia, 2017):

$$\mathbb{P}\left[\|\hat{\boldsymbol{r}}_{\text{ext}} - \boldsymbol{r}_{\text{ext}}\|_1 \geq \frac{\sqrt{2\ln(1/\delta_{\text{ext}})}}{N_{\text{ext}}}\right] \leq \delta_{\text{ext}}, \tag{85}$$

where $\delta_{\text{ext}} > 0$ represents the confidence level of the tail bound. Then we can upper bound the probability that the groundtruth output $g(Z_r)$ of the retrieved sample Z_r is not equal to the groundtruth output of the query sample $g(Z_q)$ as the following:

$$\begin{aligned}
\mathbb{P}\left[g(Z_r) \notin g(Z_q)\right] &= \mathbb{P}_{Z_q\sim\mathcal{D}}\left[g(Z_r) \notin g(Z_q) \middle| s_\theta(Z_q, Z_r) \geq \max_{z\in\mathcal{D}_{\text{ext}}} s_\theta(z, Z_q)\right] && (86)\\
&= \mathbb{P}_{Z_q\sim\mathcal{D}}\left[\max_{Z^-\in\mathcal{D}^-_{\text{ext}}(Z_q)} s_\theta(Z^-, Z_q) \geq \max_{Z^+\in\mathcal{D}^+_{\text{ext}}(Z_q)} s_\theta(Z^+, Z_q)\right] && (87)\\
&= \mathbb{P}_{Z_q\sim\mathcal{D}}\left[s_\theta(Z^-, Z_q) \geq s_\theta(Z^+, Z_q),\ \forall Z^+ \in \mathcal{D}^+_{\text{ext}}(Z_q),\ \exists Z^- \in \mathcal{D}^-_{\text{ext}}(Z_q)\right] && (88)\\
&\leq \max_{\mathcal{Q}} \sum_{c=1}^{C} r^{(c)}_{\text{test}}(\mathcal{Q})\left(1 - r^{(c)}_{\text{ext}}\right) N_{\text{ext}}\mathbb{P}\left[s_\theta(x, x^+) < s_\theta(x, x^-)\right]^{N_{\text{ext}} r^{(c)}_{\text{ext}}}, && (89)
\end{aligned}$$

where Eq. (89) is derived by applying the union bound. Considering finite-sample error of categorical distribution in Eq. (85) and combining Eqs. (74) and (82), we finally have:

$$\begin{aligned}
\mathbb{P}\left[g(Z_r) \notin g(Z_q)\right] &\leq \max_{\mathcal{Q}} \sum_{c=1}^{C} r^{(c)}_{\text{test}}(\mathcal{Q})\left(1 - r^{(c)}_{\text{ext}}\right) N_{\text{ext}}\mathbb{P}\left[s_\theta(x, x^+) < s_\theta(x, x^-)\right]^{N_{\text{ext}} r^{(c)}_{\text{ext}}} && (90)\\
&\leq \max_{\mathcal{Q}} \sum_{c=1}^{C} r^{(c)}_{\text{test}}(\mathcal{Q})\left(1 - r^{(c)}_{\text{ext}} + \frac{\sqrt{2\ln(1/\delta_{\text{ext}})}}{N_{\text{ext}}}\right) N_{\text{ext}} V_{\text{rag}}(\rho)^{0.5 N_{\text{ext}}\left(r^{(c)}_{\text{ext}} - \sqrt{2\ln(1/\delta_{\text{ext}})}/N_{\text{ext}}\right)} && (91)
\end{aligned}$$

Since we assume that the retrieval model retrieves samples identically from the external knowledge base, the number of retrieved positive examples N_{pos} follows a Binomial distribution with N_{rag} trials and failure rate in Eq. (91). Therefore, we can lower bound the expectation of N_{pos} as the following:

$$\mathbb{E}\left[N_{\text{pos}}\right] \geq \min_{\mathcal{Q}} N_{\text{rag}}\left(1 - \delta_{\text{ext}}\right)\left(1 - \sum_{c=1}^{C} r^{(c)}_{\text{test}}(\mathcal{Q})\left(1 - r^{(c)}_{\text{ext}} + \frac{\sqrt{2\ln(1/\delta_{\text{ext}})}}{N_{\text{ext}}}\right) N_{\text{ext}} V_{\text{rag}}(\rho)^{0.5 N_{\text{ext}}\left(r^{(c)}_{\text{ext}} - \sqrt{2\ln(1/\delta_{\text{ext}})}/N_{\text{ext}}\right)}\right) \tag{92}$$

which holds for any $\delta_{\text{ext}} > 0$. Therefore, letting $\delta_{\text{ext}} = 0.1$, we can finally derive the following:

$$\begin{aligned}
\mathbb{E}\left[N_{\text{pos}}\right] &\geq \min_{\mathcal{Q}} \frac{9}{10} N_{\text{rag}}\left(1 - \sum_{c=1}^{C} r^{(c)}_{\text{test}}(\mathcal{Q})\left(N_{\text{ext}} - r^{(c)}_{\text{ext}} N_{\text{ext}} + \sqrt{2\ln 10}\right) V_{\text{rag}}(\rho)^{0.5\left(r^{(c)}_{\text{ext}} N_{\text{ext}} - \sqrt{2\ln 10}\right)}\right) && (93)\\
&\geq \frac{9}{10} N_{\text{rag}}\left(1 - 1.5 N_{\text{ext}} V_{\text{rag}}(\rho)^{0.25\left(\min_c r^{(c)}_{\text{ext}} N_{\text{ext}}\right)}\right), && (94)
\end{aligned}$$

with a sample size in the external knowledge base such that $N_{\text{ext}} > 2\sqrt{2\ln 10}/\min_c r^{(c)}_{\text{ext}}$.

□

I. Proofs and detailed remarks in Sec. 6

I.1. Proof and detailed remark of Thm. 2

Remarks. (R1) Different from Prop. 3, the quality of retrieval models under distribution shifts is decreased from V_{rag} to $V_{\text{rag}}(\rho)$ with a linear decay factor $m(\rho)$. As we require $V_{\text{rag}}(\rho) < 1$ to ensure high retrieval quality, large distribution shift radius ρ must be compensated by small V_{rag}. This is consistent with the existing observations that low-variance models can generalize better under distribution shifts (Lam, 2016; Gotoh et al., 2018; Namkoong & Duchi, 2017). (R2) Compared to Prop. 3, Prop. 4 removes the dependency on varying label portions $r^{(c)}_{\text{cal}}$ during distribution shifts, as long as the size of external knowledge base is sufficiently large N_{ext} to offset the worst-case long-tail distributions, a condition often met in practice with large knowledge bases.

Proof sketch. We first formulate the expectation of similarity difference between positive pairs and negative pairs $\mathbb{E}[s_{\theta_r}(x,x^+) - s_{\theta_r}(x,x^-)]$ as a function of the contrastive loss of the retrieval model L_τ. Then, we apply Chebyshev's inequality to upper bound the failure probability $\mathbb{P}[s_\theta(x,x^+) < s_\theta(x,x^-)]$ as a function of the variance and expectation of the similarity difference. Considering the distribution shifts, the variance and expectation bound can be derived via Grammian bound as (Weber et al., 2022). We then plug in the failure rate corrected by distribution shifts and follow the proof structure of Prop. 3.

Proof of Thm. 2. Recall that the calibration samples $(Z_1, Z_2, ..., Z_{N_{\text{cal}}})$ are drawn from the distribution $\mathcal{D}$ and the test sample $Z_{\text{test}} = (X_{\text{test}}, Y_{\text{test}})$ is sampled from the distribution $\mathcal{Q}$. The distribution $\mathcal{Q}$ and $\mathcal{D}$ have bounded Hellinger distance $H(\mathcal{Q},\mathcal{D}) \leq \rho$ and the risk is upper bounded by 1. Let $R_\mathcal{Q}$, $R_\mathcal{D}$ be the population risk on distribution $\mathcal{Q}$, $\mathcal{D}$ and $V_\mathcal{D}$ be the variance of risk on distribution $\mathcal{D}$. We can apply Lem. E.1 and get the following:

$$R_\mathcal{Q} \leq R_\mathcal{D} + 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{V_\mathcal{D}} + \rho^2(2-\rho^2)\left(1 - R_\mathcal{D} - \frac{V_\mathcal{D}}{1-R_\mathcal{D}}\right), \tag{95}$$

with any given distance bound $\rho > 0$ that satisfies:

$$\rho^2 \leq 1 - \left(1 + \frac{(1-R_\mathcal{D})^2}{V_\mathcal{D}}\right)^{-1/2}. \tag{96}$$

Since the variance is non-negative, Eq. (95) further implies that:

$$R_\mathcal{Q} \leq R_\mathcal{D} + 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{V_\mathcal{D}} + \rho^2(2-\rho^2)\left(1-R_\mathcal{D}\right). \tag{97}$$

Then we will consider the finite-sample error of the calibration set. From Hoeffding's inequality (Hoeffding, 1994), with probability $1-\delta/4$, we have:

$$R_\mathcal{D} \leq \hat{R} + \sqrt{\frac{\ln(4/\delta)}{2N_{\text{cal}}}}. \tag{98}$$

From sample variance bound in (Maurer & Pontil, 2009), with probability $1-\delta/4$, we have:

$$\sqrt{V_\mathcal{D}} \leq \sqrt{\hat{V}} + \sqrt{\frac{2\ln(4/\delta)}{N_{\text{cal}}-1}}. \tag{99}$$

Note that the RHS of Eq. (97) monotonically increases in $R_\mathcal{D}$ and $V_\mathcal{D}$, combining Eqs. (97) to (99) and applying the union bound, with probability $1-\delta/2$, we have:

$$R_\mathcal{Q} \leq \hat{R} + \rho^2(2-\rho^2)\left(1-\hat{R}\right) + 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\hat{V}} + (1-\rho^2)(\frac{1-\rho^2}{\sqrt{2N_{\text{cal}}}} + \frac{2\sqrt{2}\rho\sqrt{2-\rho^2}}{\sqrt{N_{\text{cal}}-1}})\sqrt{\ln(4/\delta)}. \tag{100}$$

By two-sided Hoeffding's inequality, with probability $1-\delta/4$, we have:

$$\hat{R}_\rho \leq R_\mathcal{Q} + \sqrt{\frac{\ln(8/\delta)}{2N_{\text{cal}}}}. \tag{101}$$

Combining Eqs. (100) and (101) and applying the union bound, with probability $1-3\delta/4$, we have:

$$\hat{R}_\rho \leq \hat{R} + \rho^2(2-\rho^2)\left(1-\hat{R}\right) + 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\hat{V}} + (1-\rho^2)\left(\frac{1-\rho^2}{\sqrt{2N_{\text{cal}}}} + \frac{2\sqrt{2}\rho\sqrt{2-\rho^2}}{\sqrt{N_{\text{cal}}-1}}\right)\sqrt{\ln(4/\delta)} + \sqrt{\frac{\ln(8/\delta)}{2N_{\text{cal}}}}. \tag{102}$$

Recall that the controlled conformal risk $\hat{\alpha}$ can be formulated as a function of the empirical risk $\hat{R}$ as Eq. (16). When $(X_{\text{test}}, Y_{\text{test}})$ is sampled from $\mathcal{D}$, the guarantee of conformal risk is as follows:

$$\mathbb{P}\left[R(T_{\hat{\boldsymbol{\lambda}}}; X_{\text{test}}, Y_{\text{test}}) \leq \hat{\alpha} := \min\left\{h^{-1}\left(\frac{8/\delta}{N_{\text{cal}}}; \hat{R}\right), \Phi^{-1}_{\text{bin}}\left(\frac{\delta}{8e}; N_{\text{cal}}, \hat{R}\right)\right\}\right] \geq 1-\delta/4, \tag{103}$$

where $h^{-1}(\cdot;\cdot)$ is the partial inverse function such that $h^{-1}(h(a,b);a) = b$ with $h_1(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$. Note that $\hat{\alpha}$ monotonically increases in $\hat{R}$. By Eq. (102), with probability $1-3\delta/4$, the following holds about the conformal risk on distribution $\mathcal{Q}$ (denoted by $\hat{\alpha}_\rho$), which is within bounded Hellinger distance ρ to the original distribution $\mathcal{D}$:

$$\hat{\alpha}_\rho \leq \min\left\{h^{-1}\left(\frac{8/\delta}{N_{\text{cal}}}; \overline{\hat{R}_\rho}\right), \Phi^{-1}_{\text{bin}}\left(\frac{\delta}{8e}; N_{\text{cal}}, \overline{\hat{R}_\rho}\right)\right\}, \tag{104}$$

where $\overline{\hat{R}_\rho}$ is formulated as:

$$\overline{\hat{R}_\rho} = \hat{R} + \rho^2(2-\rho^2)\left(1-\hat{R}\right) + 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\hat{V}} + (1-\rho^2)\left(\frac{1-\rho^2}{\sqrt{2N_{\text{cal}}}} + \frac{2\sqrt{2}\rho\sqrt{2-\rho^2}}{\sqrt{N_{\text{cal}}-1}}\right)\sqrt{\ln(4/\delta)} + \sqrt{\frac{\ln(8/\delta)}{2N_{\text{cal}}}}. \tag{105}$$

Combining Eqs. (103) and (104) and applying the union bound, we can finally conclude that:

$$\mathbb{P}\left[R(T_{\hat{\boldsymbol{\lambda}}}; X_{\text{test}}, Y_{\text{test}}) \leq \hat{\alpha}_\rho := \min\left\{h^{-1}\left(\frac{8/\delta}{N_{\text{cal}}}; \overline{\hat{R}_\rho}\right), \Phi^{-1}_{\text{bin}}\left(\frac{\delta}{8e}; N_{\text{cal}}, \overline{\hat{R}_\rho}\right)\right\}\right] \geq 1-\delta, \tag{106}$$

where $\overline{\hat{R}_\rho}$ is formulated as:

$$\overline{\hat{R}_\rho} = \hat{R} + \rho^2(2-\rho^2)\left(1-\hat{R}\right) + 2\rho(1-\rho^2)\sqrt{2-\rho^2}\sqrt{\hat{V}} + (1-\rho^2)\left(\frac{1-\rho^2}{\sqrt{2N_{\text{cal}}}} + \frac{2\sqrt{2}\rho\sqrt{2-\rho^2}}{\sqrt{N_{\text{cal}}-1}}\right)\sqrt{\ln(4/\delta)} + \sqrt{\frac{\ln(8/\delta)}{2N_{\text{cal}}}}. \tag{107}$$

□

I.2. Proof and detailed remark of Thm. 3

Remarks. Our result rigorously characterizes the effect of distribution shift on the reduced risk guarantee of RAG. (R1) Compared to Thm. 1, only the uncertainty of retrieval model $p_r(\rho)$ is affected by the distribution shift ρ. This affect is reflected on the the retrieval quality $V_{\text{rag}}(\rho)$. In particular, a large distance radius ρ will downgrade the retrieval quality $V_{\text{rag}}(\rho)$ and thus lead to a higher uncertainty $p_r(\rho)$. However, the influence of ρ on $p_r(\rho)$ can be reduced by N_{rag} inverse proportionally and by N_{ext} exponentially, demonstrating the robustness of RAG with more retrieval knowledge. (R2) Since $V_{\text{rag}}(\rho)$ is proportional to model variance V_{rag}, a low-variance retrieval model demonstrates better robustness against distribution drifts, aligning with existing empirical observations (Lam, 2016; Gotoh et al., 2018; Namkoong & Duchi, 2017), which evaluate the generalization ability of low-variance retrieval models under distribution shifts. Different from Prop. 3, the quality of retrieval models under distribution shifts is decreased from V_{rag} to $V_{\text{rag}}(\rho)$ with a linear decay factor $m(\rho)$. As we require $V_{\text{rag}}(\rho) < 1$ to ensure high retrieval quality, large distribution shift radius ρ must be compensated by small V_{rag}. This is consistent with the existing observations that low-variance models can generalize better under distribution shifts (Lam, 2016; Gotoh et al., 2018; Namkoong & Duchi, 2017). (R3) Compared to Thm. 1, Thm. 3 has no dependence on varying label portions $r^{(c)}_{\text{cal}}$ during distribution shifts, as long as the size of external knowledge base N_{ext} is moderately large, ($N_{\text{ext}} > 2\sqrt{2\ln 10}/\min_c r^{(c)}_{\text{ext}}$) to offset the worst-case long-tail distributions, a condition often met in practice with large knowledge bases.

Proof sketch. We apply Prop. 4 to provide the lower bound of the retrieved positive examples under distribution shifts. We plug in the term and analyze the functionality (logit difference statistics) of the self-attention transformer as proof of Thm. 1. Connecting the logit difference statistics to the empirical risks and the distribution-shifted conformal risk bound in Thm. 2 finally concludes the proof.

Proof of Thm. 3. From Prop. 4, we prove that:

$$\mathbb{E}\left[N_{\text{pos}}\right] \geq \underline{N}_{\text{pos}} := \frac{9}{10} N_{\text{rag}} \left(1 - 1.5 N_{\text{ext}} V_{\text{rag}}(\rho)^{0.25\left(\min_c r_{\text{ext}}^{(c)} N_{\text{ext}}\right)}\right). \tag{108}$$

Since N_{pos} is a binomial random variable with N_{rag} trials, we have the upper bound of the variance $\mathbb{V}[N_{\text{pos}}] \leq \dfrac{N_{\text{rag}}}{4}$. Applying Chebyshev's inequality to the random variable N_{pos}, the following holds $\forall n_{\text{pos}} < \underline{N}_{\text{pos}}$:

$$\mathbb{P}\left[N_{\text{pos}} \geq n_{\text{pos}}\right] \geq 1 - \frac{\mathbb{V}[N_{\text{rag}}]}{(\underline{N}_{\text{pos}} - n_{\text{pos}})^2} \geq 1 - \frac{N_{\text{rag}}}{4(\underline{N}_{\text{pos}} - n_{\text{pos}})^2}, \tag{109}$$

which implicates that we can do the analysis with $N_{\text{pos}} \geq n_{\text{pos}}$ with probability $1 - \dfrac{\mathbb{V}[N_{\text{rag}}]}{(\underline{N}_{\text{pos}} - n_{\text{pos}})^2}$.

According to the proof of Thm. 1, by Eq. (62), we have the following:

$$\mathbb{E}\left[R - R_{\text{rag}}\right] \geq \Phi_M\left(\left((n_{\text{pos}}+1)d^+ - 1\right) n_{\text{pos}} \left(\int_{-1}^{1} \Phi_M(v) dv - 1\right)\right) - \Phi_M(0). \tag{110}$$

.

Let $n_{\text{rag}} = N_{\text{rag}}/2$. Combining Eq. (110) and Eq. (109), we get that if $\underline{N}_{\text{pos}} > N_{\text{rag}}/2 > 1/d^+$, with probability $1 - \dfrac{N_{\text{rag}}}{4(\underline{N}_{\text{pos}} - N_{\text{rag}}/2)^2}$, we have:

$$\mathbb{E}\left[R - R_{\text{rag}}\right] \geq \Phi_M\left(\frac{d^+\left(\int_{-1}^{1} \Phi_M(v) dv - 1\right) N_{\text{rag}}}{2}\right) - \Phi_M(0). \tag{111}$$

Let $R(Z)$ be the risk of Z sampled from the distribution $\mathcal{Q}$. Define the empirical risk $\hat{R}$ and $\hat{R}_{\text{rag}}$ as the following:

$$\hat{R} = \frac{1}{N_{\text{cal}}} \sum_{i=1}^{N_{\text{cal}}} R(Z_i), \quad \hat{R}_{\text{rag}} = \frac{1}{N_{\text{cal}}} \sum_{i=1}^{N_{\text{cal}}} R_{\text{rag}}(Z_i) \tag{112}$$

Note that since the risk $R(\cdot)$ is bounded in $[0,1]$, the variance estimator $\hat{V} = \dfrac{1}{N_{\text{cal}}(N_{\text{cal}}-1)} \sum_{1 \leq i < j \leq N_{\text{cal}}} (R(Z_i) - R(Z_j))^2$ is bounded in $[0,1]$. Leveraging the fact and according to Thm. 2, the statistical guarantee of conformal risk $\hat{\alpha}_\rho$ and $\hat{\alpha}_\rho^{\text{rag}}$ with confidence $1-\delta$ can be formulated as:

$$\begin{gathered}\hat{\alpha}_\rho := \min\left\{h^{-1}\left(\frac{8/\delta}{N_{\text{cal}}}; \overline{\hat{R}_\rho}\right), \Phi_{\text{bin}}^{-1}\left(\frac{\delta}{8e}; N_{\text{cal}}, \overline{\hat{R}_\rho}\right)\right\} \\ \text{where } \overline{\hat{R}_\rho} = \hat{R} + \rho^2(2-\rho^2)(1-\hat{R}) + 2\rho(1-\rho^2)\sqrt{2-\rho^2} + C(\rho, N_{\text{cal}}),\end{gathered} \tag{113}$$

$$\begin{gathered}\hat{\alpha}_\rho^{\text{rag}} := \min\left\{h^{-1}\left(\frac{8/\delta}{N_{\text{cal}}}; \overline{\hat{R}_\rho^{\text{rag}}}\right), \Phi_{\text{bin}}^{-1}\left(\frac{\delta}{8e}; N_{\text{cal}}, \overline{\hat{R}_\rho^{\text{rag}}}\right)\right\} \\ \text{where } \overline{\hat{R}_\rho^{\text{rag}}} = \hat{R}_{\text{rag}} + \rho^2(2-\rho^2)(1-\hat{R}_{\text{rag}}) + 2\rho(1-\rho^2)\sqrt{2-\rho^2} + C(\rho, N_{\text{cal}}),\end{gathered} \tag{114}$$

where $C(\rho, N_{\text{cal}}) = (1-\rho^2)\left(\dfrac{1-\rho^2}{\sqrt{2N_{\text{cal}}}} + \dfrac{2\sqrt{2}\rho\sqrt{2-\rho^2}}{\sqrt{N_{\text{cal}}-1}}\right)\sqrt{\ln(4/\delta)} + \sqrt{\dfrac{\ln(8/\delta)}{2N_{\text{cal}}}}$.

Noting that $\hat{\alpha}_\rho$ is monotonically increasing in $\overline{\hat{R}_\rho}$ and $\overline{\hat{R}_\rho}$ is monotonically increasing in $\hat{R}$, $\hat{\alpha}_\rho$ is monotonically increasing in $\hat{R}$. Then, the following holds by Hoeffding's inequality:

$$\mathbb{P}\left[\hat{\alpha}_\rho^{\text{rag}} < \hat{\alpha}_\rho\right] \geq \mathbb{P}\left[\hat{R}_{\text{rag}} < \hat{R}\right] \geq 1 - \exp\left\{-2N_{\text{cal}}\mathbb{E}\left[R - R_{\text{rag}}\right]^2\right\}. \tag{115}$$

Algorithm 3 Test distribution sampling protocol.

```
Input: original test set $\mathcal{D}$, test sample pool $\mathcal{D}_{\text{pool}}$, Risk function $R(\cdot,\cdot) : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$
Output: empirical risk on the sampled test set $\mathcal{Q}$ $\hat{R}_{\mathcal{Q}}$
Randomly sample $\mathcal{Q}$ from $\mathcal{D}_{\text{pool}}$ with equalized set size as $\mathcal{D}$: $|\mathcal{Q}| = |\mathcal{D}|$
Evaluate empirical risks for all samples in $\mathcal{Q}$: $\hat{R}_{\mathcal{Q}} = \frac{1}{|\mathcal{Q}|} \sum_{(x,y)\in\mathcal{Q}} R(x,y)$
Return $\hat{R}_{\mathcal{Q}}$
```

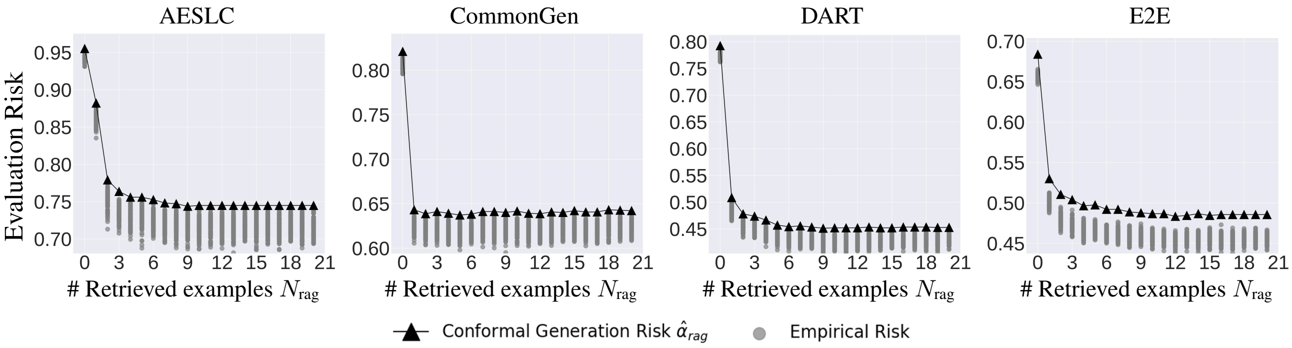


Figure 8: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and simulations of empirical risks with Biencoder-SFT for different N_{rag} and fixed $\lambda_g = 1, \lambda_s = 1.0$.

Combining Eq. (111) and Eq. (115) and using the union bound, under the condition that $\underline{N}_{\text{pos}} > N_{\text{rag}}/2 > 1/d^+$, we have:

$$\mathbb{P}\left[\hat{\alpha}_{\text{rag}} < \hat{\alpha}\right] \geq 1 - \exp\left\{-2N_{\text{cal}}\left[\Phi_M\left(\frac{d^+\left(\int_{-1}^{1}\Phi_M(v)dv - 1\right)N_{\text{rag}}}{2}\right) - \Phi_M(0)\right]^2\right\} - \frac{N_{\text{rag}}}{4(\underline{N}_{\text{pos}} - N_{\text{rag}}/2)^2}. \quad (116)$$

Combining Eqs. (108) and (116), under the condition that $N_{\text{ext}} > \frac{2\sqrt{2\ln 10}}{r^m_{\text{ext}}}$, $N_{\text{ext}}V_{\text{rag}}(\rho)^{0.25 r^m_{\text{ext}} N_{\text{ext}}} < \frac{8}{17}$, and $N_{\text{rag}} > \frac{2}{d^+}$, we can finally conclude that:

$$\begin{aligned}\mathbb{P}\left[\hat{\alpha}_{\text{rag}} < \hat{\alpha}\right] \geq & 1 - \exp\left\{-2N_{\text{cal}}\left[\Phi_M\left(\frac{d^+\left(\int_{-1}^{1}\Phi_M(v)dv - 1\right)N_{\text{rag}}}{2}\right) - \Phi_M(0)\right]^2\right\} \\ & - \frac{100}{N_{\text{rag}}}\left(8 - 17N_{\text{ext}}V_{\text{rag}}(\rho)^{0.25\left(\min_c r^{(c)}_{\text{ext}} N_{\text{ext}}\right)}\right)^{-2}.\end{aligned} \quad (117)$$

□

J. Additional evaluation results

J.1. Experiment setup

Datasets We evaluate C-RAG on four NLP datasets with retrieval augmented generation utilizing an external knowledge base. (1) AESLC: The Annotated Enron Subject Line Corpus (AESLC) dataset (Zhang & Tetreault, 2019) contains a collection of email messages from employees of the Enron Corporation. It contains two primary features: the “email_body“, which is the text of the email body, and the “subject_line“, which is the text of the email subject. The task is to generate the email subject given the email body. (2) The Generative Commonsense Reasoning (CommonGen) dataset (Lin et al., 2019) is a collection of commonsense descriptions, amounting to 79k descriptions over 35k unique concept sets, constructed

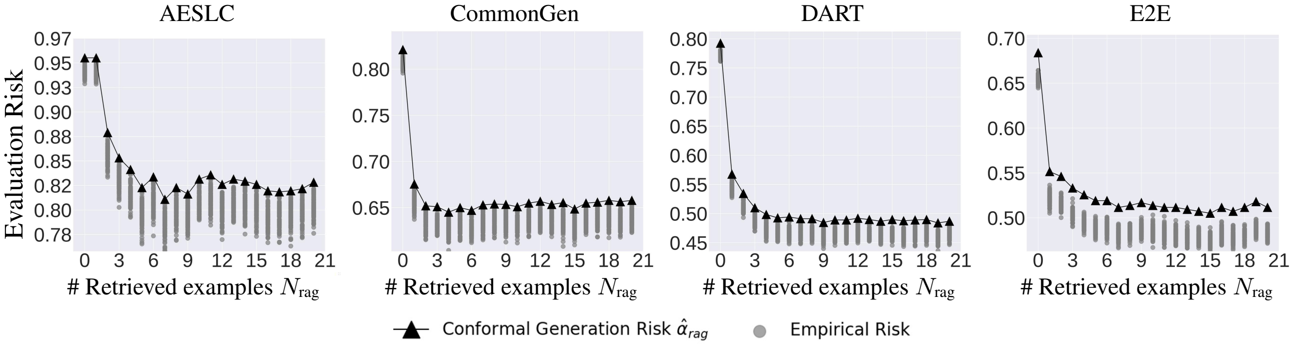


Figure 9: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and simulations of empirical risks with BM25 for different N_{rag} and fixed $\lambda_g = 1, \lambda_s = 1.0$.

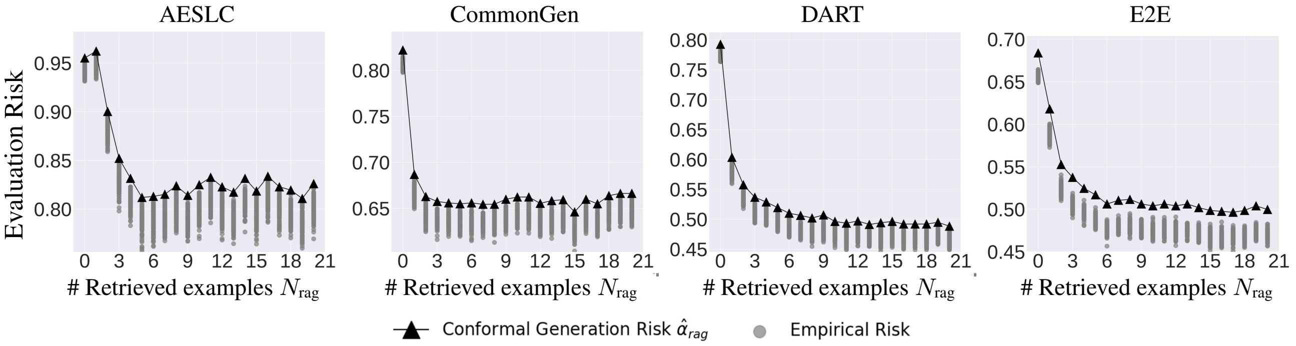


Figure 10: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and simulations of empirical risks with BAAI/bge for different N_{rag} and fixed $\lambda_g = 1, \lambda_s = 1.0$.

via crowdsourcing and existing caption corpora. The task is to generate a sentence that uses given concepts in a coherent and commonsense way. (3) The Data Record to Text (DART) dataset (Nan et al., 2020) is a large-scale dataset designed to facilitate the generation of text from structured data records. It comprises 82,191 examples spanning various domains, with each example being a set of Resource Description Framework (RDF) triples derived from data records in tables and tree ontologies of schemas. These are annotated with sentence descriptions that encapsulate all the facts presented in the RDF triplet. (4) The End-to-End Generation (E2E) dataset (Novikova et al., 2017) contains about 50k comments in the restaurant domain. The task is to generate text from meaning representations (MRs), which are structured inputs that describe various aspects of a restaurant. These MRs consist of slots and values that a model needs to convert into natural language descriptions that are coherent and fluent.

External knowledge base: Following (Wang et al., 2023c; Wei et al., 2021; Cheng et al., 2023), we construct the external knowledge base as the union of a total of 30 publicly available datasets from 9 distinct categories with over 6 million documents.

Metrics For generation tasks, we leverage ROUGE-L to quantify the quality of generations. ROUGE-L measures the longest common subsequence between a candidate generation and reference texts and is typically adopted for generation quality evaluations across the literature (Lin, 2004; Gatt & Krahmer, 2018). A low ROUGE-L implies poor quality generations and accordingly a high generation risk. Therefore, we adopt the risk function as $1 -$ ROUGE-L to bound the risk in $[0, 1]$. Note that C-RAG is agnostic to selection of risk functions, and thus, practitioners can specify any function that suits their specific use cases.

Retrieval models We consider four types of retrieval models. (1) BM25 (Robertson et al., 2009) is a sparse encoding metric used to rank the candidate documents given a query in information retrieval tasks. BM25 scores are linearly weighted combinations of token-level scores and can be analytically formulated as a function of term frequency, inverse document frequency, and length of documents. (2) BAAI general embedding (BAAI/bge) (Zhang et al., 2023a) trains the SOTA

Algorithm 4 Shifted distribution sampling protocol.

```
Input: original test set $\mathcal{D}$, test sample pool $\mathcal{D}_{\text{pool}}$, Risk function $R(\cdot,\cdot) : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$
Output: empirical risk on the shifted test set $\mathcal{Q}$ $\hat{R}_{\mathcal{Q}}$, Hellinger distance between $\mathcal{D}$ and $\mathcal{Q}$ $H_{\mathcal{P}\mathcal{Q}}$
Randomly sample $\mathcal{Q}$ from $\mathcal{D}_{\text{pool}}$ with equalized set size as $\mathcal{D}$: $|\mathcal{Q}| = |\mathcal{D}|$
Randomly sample the sample weight vector of $\mathcal{Q}$ $\boldsymbol{w} \in \Delta^{|\mathcal{D}|}$
Compute Hellinger distance as $H_{\mathcal{P}\mathcal{Q}} = \sqrt{1 - \sum_{i=1}^{|\mathcal{D}|} \sqrt{\boldsymbol{w}_i / |\mathcal{D}|}}$
Evaluate risks for all samples in $\mathcal{Q}$ with risk function $R(\cdot,\cdot)$: $\hat{\boldsymbol{R}}_{\mathcal{Q}} \in \mathbb{R}^{|\mathcal{D}|}$
Compute the empirical risk on $\mathcal{Q}$ with weight vector $\boldsymbol{w}$: $\hat{R}_{\mathcal{Q}} = \boldsymbol{w}^T \hat{\boldsymbol{R}}_{\mathcal{Q}}$
Return $\hat{R}_{\mathcal{Q}}$, $H_{\mathcal{P}\mathcal{Q}}$
```

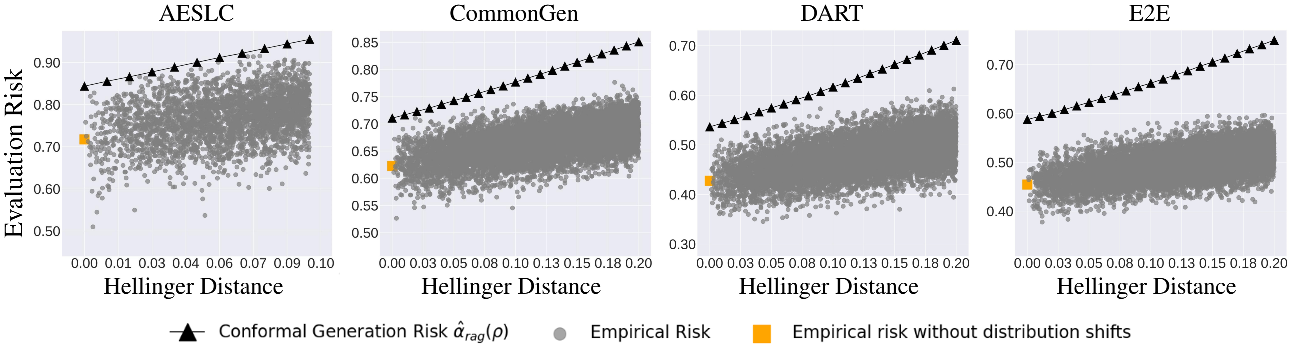


Figure 11: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and empirical risks with Biencoder-SFT under distribution shifts with $N_{\text{rag}} = 15, \lambda_g = 1, \lambda_s = 1.0$.

embedding model in the MTEB benchmark (Muennighoff et al., 2022) and supports the diverse retrieval augmentation needs of LLMs by a reward formulation based on LLMs' feedback, the stabilization of knowledge distillation, multi-task fine-tuning with explicit instructions, and homogeneous in-batch negative sampling. (3) OpenAI ada-text-embedding-02 (OpenAI/ada) is a close source text embedding models designed to convert text into high-dimensional vectors, which capture the semantic meaning of the input text and can be used for a variety of tasks, such as text similarity, classification, and clustering. (4) Biencoder-supervised fine-tuning (Biencoder-SFT) (Wang et al., 2023c) is a bi-encoder based dense retriever trained with contrastive loss and hard negative sampling strategies. It iteratively trains the retrieval model with hard negative samples identified by computing similarity scores by the current retrieval model.

Implementation details We leverage Llama-7b for inference without specification. We perform conformal calibration on validation sets for different NLP datasets and fix the confidence levels $1 - \delta = 0.9$ across all evaluations. Lastly, we concatenate the retrieved in-context examples and the query sample with line break tokens.

J.2. Conformal risk bounds evaluation

Soundness and tightness of generation risk guarantees in C-RAG. To achieve the generation risk guarantee in Eq. (3), C-RAG computes the upper confidence bound of generation risk by Prop. 1, which takes the empirical risk, calibration size and confidence level as input. We evaluate the conformal risks of RAG models α_{rag} with variations of the number of retrieved examples N_{rag} by calibration statistics on the validation set. To validate the soundness and tightness of the risk bounds, we randomly sample multiple test sets from a pool of test samples and compute the empirical risks on the sampled test sets. The sampling protocol is detailed in Alg. 3 in App. J.2. We provide the results for BM25, BAAI/bge, Biencoder-SFT in Figs. 8 to 10. The results show that **(1)** the certified conformal risks α_{rag} (black up-pointing triangles) are larger than the empirical risks of sampled test sets (grey points) and some empirical risks approach the risk bounds, demonstrating the soundness and tightness of risk bounds in C-RAG, and **(2)** the conformal risks decrease much as the number of retrieved examples N_{rag} increases, which shows the effectiveness of retrieved in-context examples in RAG model and aligns with our theoretical analysis in Thm. 1.

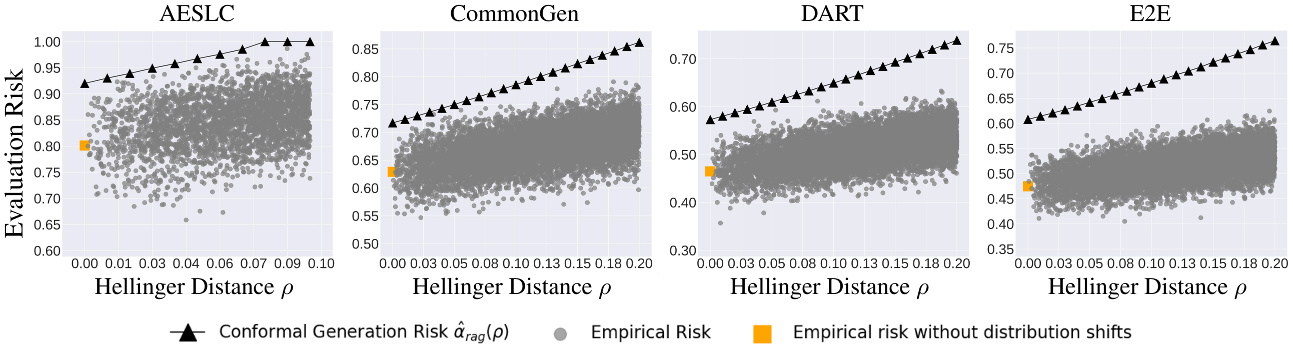


Figure 12: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and empirical risks with BM25 under distribution shifts with $N_{\text{rag}} = 15, \lambda_g = 1, \lambda_s = 1.0$.

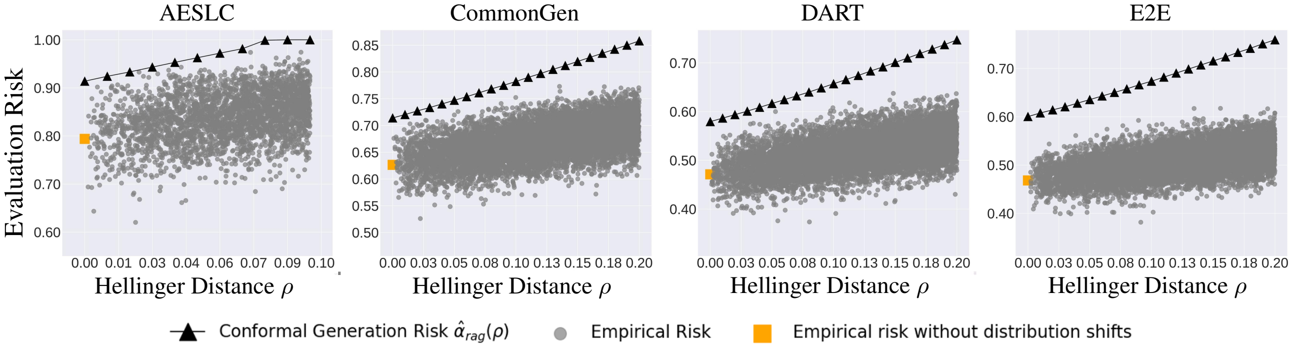


Figure 13: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and empirical risks with BAAI/bge under distribution shifts with $N_{\text{rag}} = 15, \lambda_g = 1, \lambda_s = 1.0$.

Comparisons of C-RAG risk bounds for SOTA retrieval models. We also compare the conformal risk bounds of C-RAG for different retrieval models, including sparse encoding method BM25, and SOTA dense encoding models BAAI/bge, OpenAI/ada, Biencoder-SFT. The results in Fig. 4 show that **(1)** RAG benefits in achieving a lower conformal risks of generations for different retrieval models, and **(2)** Biencoder-SFT is the most performant generally since the model is trained in the same domain as the test sets, while OpenAI/ada trained with an open corpus also demonstrates impressive effectiveness and even outperforms Biencoder-SFT on CommonGen dataset.

J.3. Conformal risk bounds evaluations under distribution shifts

Soundness and tightness of distribution-shifted conformal risk bounds in C-RAG. The test user input text can be out of the calibration distribution in practice, which needs correction of the conformal risk bounds considering the distribution drift. We provide the first distribution-shfted conformal risk bounds for general bounded risk functions in Thm. 2. We evaluate the bounds in Eq. (8) as a function of the distribution shifting distance measured by Hellinger distance ρ. To validate the bounds, we also construct various test sets with covariate shift induced by sample weight shifting. Specifically, different weights can be assigned to test samples in the shifted test sets and the Hellinger distance can also be explicitly computed by the original sample weights and shifted sample weights. We provide the detailed procedure in Alg. 4 in App. J.3. We provide the results of conformal risk bounds and simulated empirical risks with $N_{\text{rag}} = 15$ in Figs. 11 to 13 with BM25, BAAI/bge, and Biencoder-SFT. The results demonstrate that **(1)** the distribution-shifted conformal risk bounds in Thm. 2 is sound and tight for different retrieval models, and **(2)** the conformal risk bounds increases linearly with the Hellinger distance and are non-trivial for a Hellinger distance up to 0.2.

Comparisons of C-RAG distribution-shifted risk bounds for SOTA retrieval models. We compare the distribution-drift conformal risk bounds for different retrieval models in Fig. 14. The results show that **(1)** the bounds for different retrieval models increases linearly with a same slope as the Hellinger distance ρ increases, and **(2)** Biencoder-SFT and OpenAI/ada

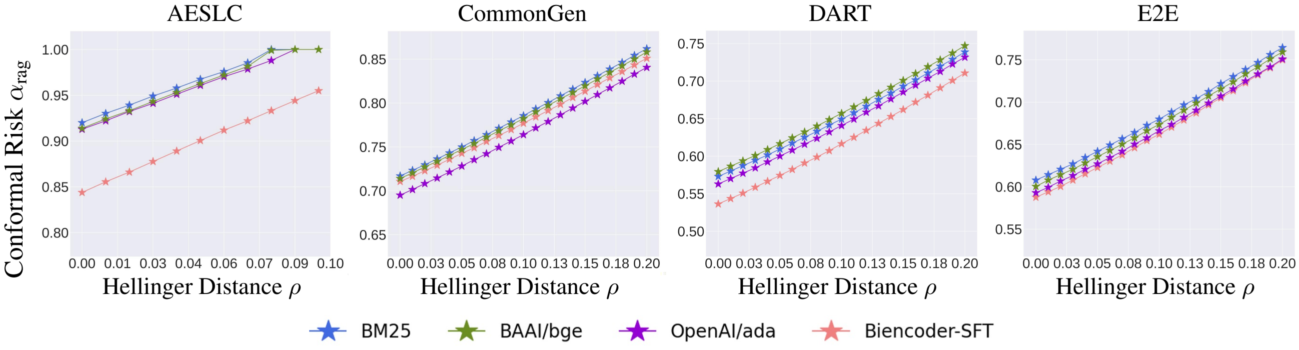


Figure 14: Conformal generation risk α_{rag} vs. Hellinger distance ρ for different retrieval models under distribution shifts with $N_{\text{rag}} = 15, \lambda_g = 1, \lambda_s = 1.0$.

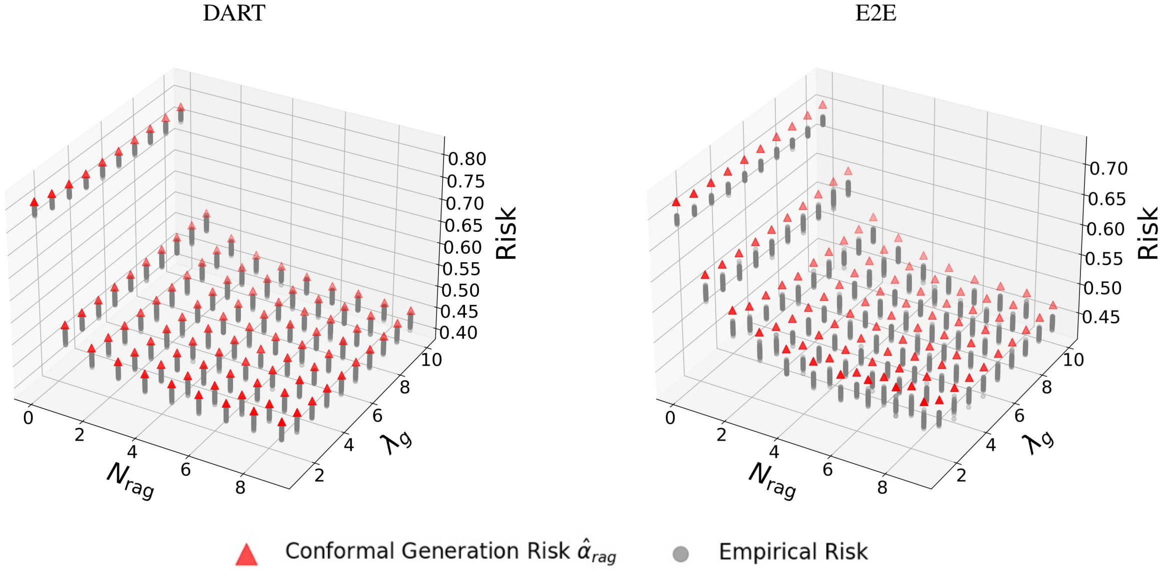


Figure 15: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and simulated empirical risks with different λ_g and N_{rag} for OpenAI/ada.

demonstrate lower conformal risk bounds for different distances due to a lower initial risk without distribution shifts.

J.4. Conformal risk bounds with multi-dimensional RAG configurations

We already theoretically and empirically demonstrate the effectiveness of retrieved in-context examples quantified by N_{rag}. To further improve the conformal risk bounds, we can addtionally consider more RAG parameters such as the number of generations λ_g in the generation set and a similarity threshold λ_s to control the diversity of generations as Alg. 1. We follow the RAG generation protocol in Alg. 1 and define the risk function as the minimal risks among all candidate generations. We similarly construct random test sets and provide the results on DART and E2E in Fig. 7. The results show that **(1)** the conformal risk bounds for multi-dimensional RAG configurations are still sound and tight, and **(2)** a larger N_{rag} can reduce the conformal risks more sensitively compared to the number of generations λ_g, demonstrating the effectiveness of more retrieved in-context examples. We also fix the number of retrieved examples $N_{\text{rag}} = 5$ and the diversity threshold $\lambda_s = 1.0$, and evaluate the certified conformal risk and empirical risk of randomly sampled test set in Fig. 17, which demonstrates that although not effective as the number of retrieved examples, a larger generation set size also benefits a low generation risk. Further more, we fix $N_{\text{rag}} = 5, \lambda_g = 20, \lambda_s = 1.0$ and evaluate the risks for different distribution drift distance in Fig. 16.

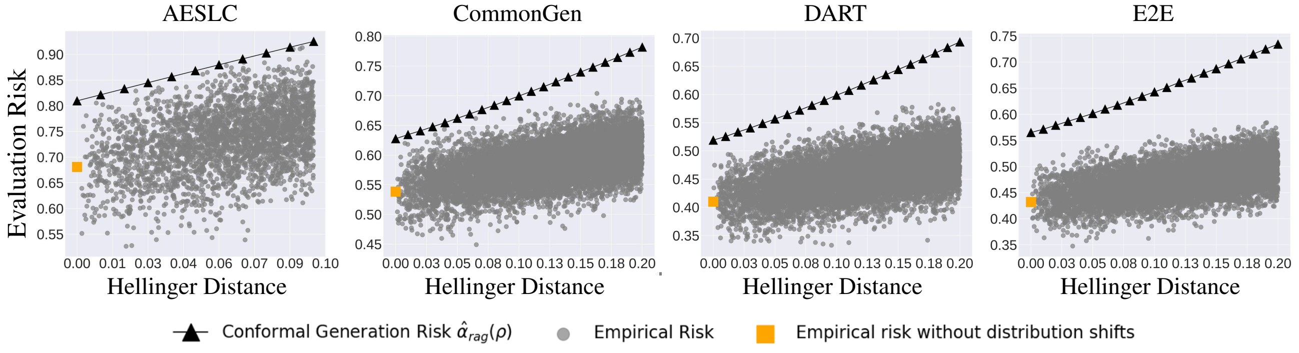


Figure 16: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and empirical risks with $N_{\text{rag}} = 5, \lambda_g = 20, \lambda_s = 1.0$ under distribution shifts for Biencoder-SFT.

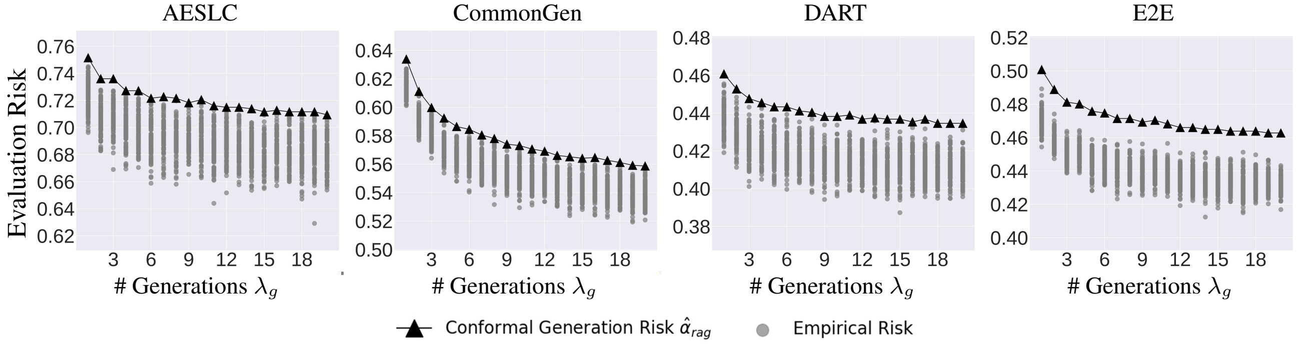


Figure 17: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and simulated empirical risks for different generation set size λ_g and fixed $N_{\text{rag}} = 5, \lambda_s = 1.0$ for Biencoder-SFT.

J.5. Valid configurations identification with specified risk levels

In conformal analysis **(2)** in Sec. 4.3, given a desired risk level α, C-RAG can certify a set of valid configurations $\hat{\Lambda}_\alpha$ such that any configuration in the set results in a conformal risk below α. We use Bonferroni correction for family-wise error rate control and evaluate empirical risks on randomly sampled test sets for the identified valid configurations. We provide the results on DART and ECE datasets in Fig. 15. The results demonstrate that **(1)** the certification is empirically sound since the empirical risks of valid configurations $\hat{\Lambda}_\alpha$ are always below the desired level α, and **(2)** a large number of retrieved examples N_{rag} and a large generation set size λ_g are effective in achieving a low generation risk since the valid configuration set includes the region with a large N_{rag} and λ_g. We also individually demonstrate the effectiveness of generation set size λ_g and diversity threshold λ_s in Figs. 17 and 18.

J.6. Conformal generation risks with different inference models

Table 1: Comparison of conformal generation risk $\hat{\alpha}_{\text{rag}}$ with different N_{rag} using Llama-2-7b, Mistral-7B-Instruct-v0.2, and Llama-2-13b. The results are evaluated on the AESLC dataset with text-embedding-ada-002 from OpenAI as the retrieval model.

Model	$N_{\text{rag}} = 0$	$N_{\text{rag}} = 1$	$N_{\text{rag}} = 2$	$N_{\text{rag}} = 3$	$N_{\text{rag}} = 4$	$N_{\text{rag}} = 5$
Llama-2-7b	0.953	0.957	0.879	0.868	0.847	0.836
Mistral-7B-Instruct-v0.2	0.897	0.829	0.813	0.795	0.792	0.793
Llama-2-13b	0.889	0.823	0.802	0.792	0.772	0.772

We conduct additional evaluations of the conformal generation risk (upper bound of the generation risk) of C-RAG with different types of inference models as well as models with different sizes. The results in Table 1 demonstrate that (1)

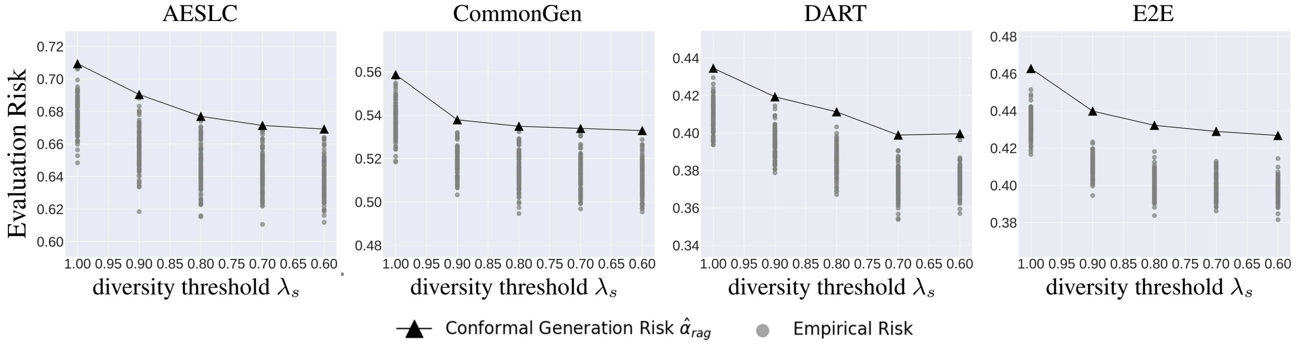


Figure 18: Conformal generation risk $\hat{\alpha}_{\text{rag}}$ and simulated empirical risks for different diversity threshold λ_s with fixed $N_{\text{rag}} = 5, \lambda_g = 20$ for Biencoder-SFT.

for different inference models, the conformal generation risk decreases effectively as the number of retrieved in-context examples N_{rag} increases, and (2) given that Llama-2-13b outperforms Llama-2-7b in the evaluations, models with larger sizes in the same model family achieve a lower conformal generation risk.

J.7. Qualitative example

Table 2: A qualitative example.

Prompt	**Generation set**	
Vanilla generation	Which team won the 2020 World Cup?	"The 2020 World Cup was not held in 2020 due to the COVID-19 pandemic."
RAG generation protocol ($N_{\text{rag}} = 1$)	The World Cup is held every four years. France won the 2018 World Cup in Russia. Argentina won the 2022 World Cup in Katar. Which team won the 2020 World Cup?	"The 2020 World Cup has not yet been held."
RAG generation protocol ($N_{\text{rag}} = 1, \lambda_g = 3$)	The World Cup is held every four years. France won the 2018 World Cup in Russia. Argentina won the 2022 World Cup in Katar. Which team won the 2020 World Cup?	"The 2020 World Cup has not yet been held.", "2020 World Cup is not held yet.", "2020 World Cup has not held yet."
RAG generation protocol ($N_{\text{rag}} = 1, \lambda_g = 3, \lambda_s = 0.5$)	The World Cup is held every four years. France won the 2018 World Cup in Russia. Argentina won the 2022 World Cup in Katar. Which team won the 2020 World Cup?	"The 2020 World Cup has not yet been held.", "The 2020 World Cup was not held due to the COVID-19 pandemic", "2020 was not the year for any World Cup as the tournament"

In Table 2, we use the following qualitative example to illustrate the effectiveness of our RAG protocol. Consider the input prompt, "Which team won the 2020 World Cup?" This prompt is inherently misleading, as there was no World Cup event scheduled for 2020. In the vanilla generation process, the model (LLAMA-2-7B-32K-INSTRUCT) produces a misconception stating, "The 2020 World Cup was not held in 2020 due to the COVID-19 pandemic." By integrating one retrieved result ($N_{rag} = 1$) into the prompt: "The World Cup is held every four years." the model no longer falsely attributes the absence of the event to COVID-19. However, it still fails to recognize that 2020 was not a designated year for the World Cup. Simply increasing the generation set size to $\lambda_g = 3$ yields similar results, failing to address the core issue. Nonetheless, by imposing diversity constraints on the generations $\lambda_s = 0.5$, the model correctly identifies the crux of the matter: "2020 was not the year for any World Cup as the tournament".