# A Survey of Confidence Estimation and Calibration in Large Language Models

**Jiahui Geng**[1], **Fengyu Cai**[2], **Yuxia Wang**[1],
**Heinz Koeppl**[2], **Preslav Nakov**[1], **Iryna Gurevych**[1]
[1] Mohamed bin Zayed University of Artificial Intelligence
[2] Technical University of Darmstadt
{jiahui.geng, yuxia.wang,preslav.nakov,iryna.gurevych}@mbzuai.ac.ae,
{fengyu.cai,heinz.koeppl}@tu-darmstadt.de

## Abstract

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks in various domains. Despite their impressive performance, they can be unreliable due to factual errors in their generations. Assessing their confidence and calibrating them across different tasks can help mitigate risks and enable LLMs to produce better generations. There has been a lot of recent research aiming to address this, but there has been no comprehensive overview to organize it and to outline the main lessons learned. The present survey aims to bridge this gap. In particular, we outline the challenges and we summarize recent technical advancements for LLM confidence estimation and calibration. We further discuss their applications and suggest promising directions for future work.

## 1 Introduction

Large language models (LLMs) have demonstrated a wide range of capabilities, such as world knowledge storage, sophisticated reasoning, and in-context learning (Petroni et al., 2019; Wei et al., 2022; Brown et al., 2020). However, LLMs do not achieve good performance on all tasks (Wang et al., 2023a; Zhang et al., 2023b). Their generation still includes biases (Zhao et al., 2021; Wang et al., 2023c) and hallucinations that do not align with reality (Zhang et al., 2023b). Assessing the trustworthiness of the generations of these models remains challenging (Liu et al., 2023c).

Confidence (or uncertainty) estimation is crucial for tasks such as out-of-distribution detection and selective prediction (Kendall and Gal, 2017; Lu et al., 2022), and it has been extensively studied and applied in various contexts (Lee et al., 2018; DeVries and Taylor, 2018). A related concept is that of model calibration, which focuses on aligning predictive probabilities (estimated confidence) to actual accuracy (Guo et al., 2017).

However, applying these methods directly to LLMs presents several challenges. The output space of these models is significantly larger than that of discriminative models. The number of possible outcomes grows exponentially with the generation length, making it impossible to access all potential responses. Additionally, different expressions may convey the same meaning, suggesting that confidence estimation should consider semantics (Kuhn et al., 2023). Finally, LLMs demonstrate unique properties, such as expressing confidence in words (Lin et al., 2022; Xiong et al., 2024) and ability to perform zero-shot or few-shot learning (Brown et al., 2020). Nonetheless, their responses can be sensitive to the prompts, e.g., the examples provided and their order, which can cause a lot of instability in the results (Min et al., 2022; Wang et al., 2023b). Given this, confidence estimation and calibration for LLMs is growing as an emerging area of interest (Jiang et al., 2021; Lin et al., 2022, 2023; Shrivastava et al., 2023).

While existing surveys mainly focused on issues such as hallucination and factuality (Zhang et al., 2023b; Wang et al., 2023a, 2024b), there are no comprehensive surveys discussing recent advancements in confidence estimation and calibration for LLMs; here we aim to bridge this gap. We explore the unique challenges posed by LLMs and examine the latest studies addressing these issues. In Section 2, we first discuss key concepts such as confidence, uncertainty, and calibration in the context of neural models. We further describe different metrics for classification and generation tasks. Then, we pursue two different directions: one addressing confidence estimation and calibration techniques for generation in Section 3, and another one focusing on classification in Section 4. We conclude by exploring their practical applicability (Section 5) and looking at potential future research directions (Section 6). Figure 1 shows the work we explore in this survey, organized in a taxonomy.

*Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595
June 16-21, 2024 ©2024 Association for Computational Linguistics

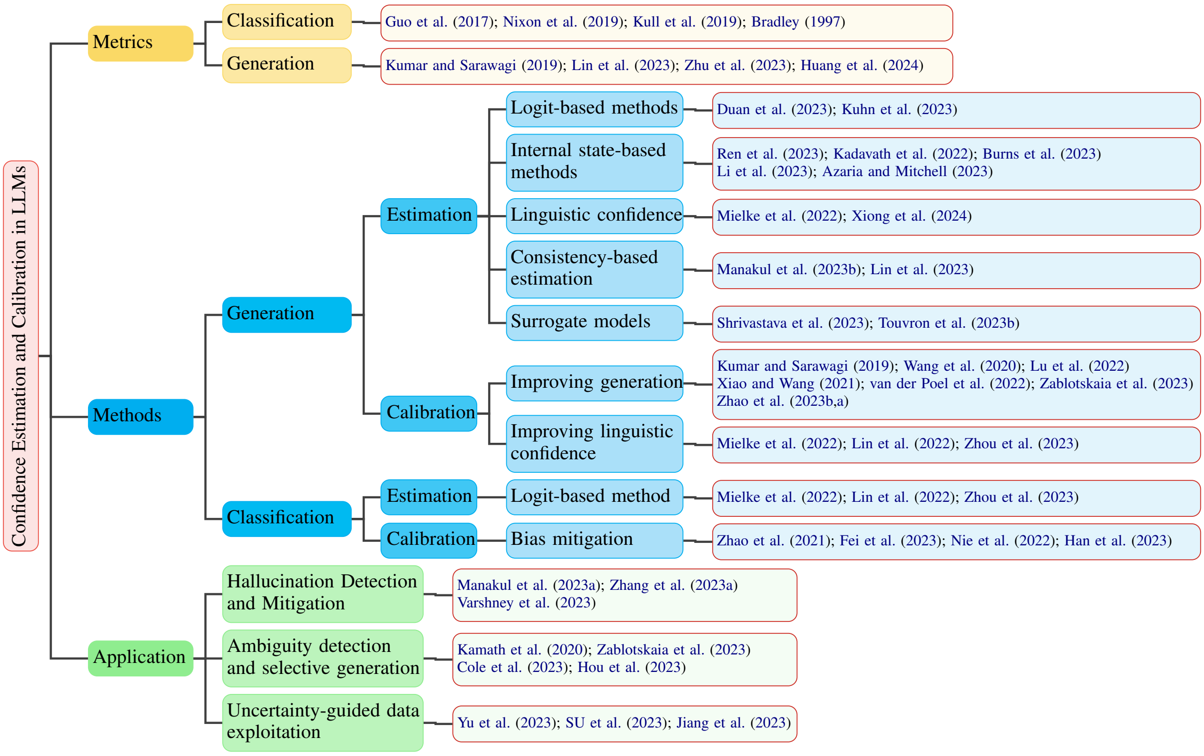

Confidence Estimation and Calibration in LLMs
Metrics
Classification
Guo et al. (2017); Nixon et al. (2019); Kull et al. (2019); Bradley (1997)
Generation
Kumar and Sarawagi (2019); Lin et al. (2023); Zhu et al. (2023); Huang et al. (2024)
Methods
Generation
Estimation
Logit-based methods
Duan et al. (2023); Kuhn et al. (2023)
Internal state-based methods
Ren et al. (2023); Kadavath et al. (2022); Burns et al. (2023)
Li et al. (2023); Azaria and Mitchell (2023)
Linguistic confidence
Mielke et al. (2022); Xiong et al. (2024)
Consistency-based estimation
Manakul et al. (2023b); Lin et al. (2023)
Surrogate models
Shrivastava et al. (2023); Touvron et al. (2023b)
Calibration
Improving generation
Kumar and Sarawagi (2019); Wang et al. (2020); Lu et al. (2022)
Xiao and Wang (2021); van der Poel et al. (2022); Zablotskaia et al. (2023)
Zhao et al. (2023b,a)
Improving linguistic confidence
Mielke et al. (2022); Lin et al. (2022); Zhou et al. (2023)
Classification
Estimation
Logit-based method
Mielke et al. (2022); Lin et al. (2022); Zhou et al. (2023)
Calibration
Bias mitigation
Zhao et al. (2021); Fei et al. (2023); Nie et al. (2022); Han et al. (2023)
Application
Hallucination Detection and Mitigation
Manakul et al. (2023a); Zhang et al. (2023a)
Varshney et al. (2023)
Ambiguity detection and selective generation
Kamath et al. (2020); Zablotskaia et al. (2023)
Cole et al. (2023); Hou et al. (2023)
Uncertainty-guided data exploitation
Yu et al. (2023); SU et al. (2023); Jiang et al. (2023)

Figure 1: The taxonomy of confidence estimation and calibration in LLMs.

## 2 Preliminaries and Background

### 2.1 Basic Concepts

In machine learning, confidence and uncertainty are two facets of a single principle: higher confidence corresponds to lower uncertainty (Xiao et al., 2022; Chen and Mueller, 2023). Research on quantifying model confidence has led to the development of two key concepts: *relative confidence score* and *absolute confidence score*, offering different ways to assess and to interpret confidence levels (Kamath et al., 2020; Vazhentsev et al., 2023a). Given an input $x$, a ground truth label $y$, and a predicted label $\hat{y}$, the model's predictive confidence is denoted as $\texttt{conf}(x, \hat{y})$. Relative confidence scores emphasize the ability to rank samples, distinguishing correct predictions from incorrect ones. Ideally, for every pair $(x_i, y_i)$ and $(x_j, y_j)$ and their corresponding predictions $\hat{y}_i$ and $\hat{y}_j$, we should have

$$\begin{aligned} &\texttt{conf}(\mathbf{x}_i, \hat{y}_i) \leq \texttt{conf}(\mathbf{x}_j, \hat{y}_j) \\ \iff &P(\hat{y}_i = y_i | \mathbf{x}_i) \leq P(\hat{y}_j = y_j | \mathbf{x}_j) \end{aligned} \tag{1}$$

An absolute confidence score indicates that a model's score reflects its true accuracy. For example, if a model predicts an event with 70% probability, that event should actually occur 70% of the time under similar circumstances. The equation for this relationship is as follows:

$$P(\hat{y} = y \mid \texttt{conf}(x, \hat{y}) = q) = q \tag{2}$$

When the model's predicted confidence scores consistently align with this principle, the model is considered to be well-calibrated.

Kendall and Gal (2017) proposed to categorize the uncertainty in machine learning into *aleatoric* and *epistemic*. Aleatoric or data uncertainty emerges from the inherent randomness or the variability of a system or a process. It is an intrinsic feature of the system and is typically irreducible. Epistemic uncertainty, in contrast, is known as model uncertainty or systematic uncertainty. It arises from the lack of knowledge or information about the system being modeled and is reducible, as it can diminish with the acquisition of more data and improved modeling (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017).

### 2.2 Evaluation Measures and Methods

**Evaluation measures** Due to the continuous nature of confidence scores, it is impossible to accurately calculate the probability as in Eq. 2. Expected calibration error (ECE; Guo et al. 2017) approximates it by clustering instances with similar confidence.

| Study | Model | Task | Calibration Methods |
|---|---|---|---|
| (Desai and Durrett, 2020) | BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) | natural language inference, paraphrase detection, commonsense reasoning | TS, LS |
| (Kim et al., 2023) | RoBERTa (Liu et al., 2019) | text classification | BL, ERL, MixUp, DeepEnsemble, MCDropout, MIMO |
| (Park and Caragea, 2022) | BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) | natural language inference, paraphrase detection, commonsense reasoning | TS, LS, MixUp, Manifold-MixUp, AUM-guided MixUp |
| (Zhang et al., 2021) | BERT-based Span Extractor (Zhang et al., 2021) | extractive question answering | FBC |
| (Si et al., 2022) | BERT-based span extractor (Si et al., 2022) | extractive question answering | LS, TS, FBC |

Table 1: **Studies on discriminative LM calibration**. **Calibration methods:** LS=label smoothing, TS=temperature scaling, BL=brier loss, ERL=entropy regularization loss, BE=Bayesian Ensemble, SNGP: spectral-normalized Gaussian process, FBC=feature-based calibrator.

The predicted probabilities are put into bins, and ECE is calculated as the weighted average of the discrepancies between the mean predicted probability and the actual accuracy across all bins $B_m$ $(m = 1, \ldots, M)$:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)| \quad (3)$$

One drawback of ECE is its sensitivity to bucket width and the variance of the samples within these buckets. Thus, more sophisticated schemes have been developed, including static calibration error (SCE), adaptive calibration error (ACE; Nixon et al. 2019), and classwise ECE (Kull et al., 2019). ECE can also be visualized as a reliability diagram: it plots predicted probabilities against observed frequencies, with points above the diagonal indicating overconfidence. Moreover, F1 score, area under receiver operating characteristic curve (AUROC; Bradley 1997) and area under accuracy-rejection curve (AUARC; Lin et al. 2023), can indicate whether the confidence score can appropriately differentiate between correct and incorrect answers.

However, it is also necessary to adapt the measures to effectively process sequences of tokens. A common approach for this is to evaluate whether the next token' probability is well-calibrated. Let $\mathbf{y}_i = y_{i1}, \cdots, y_{iT}$ denote the sequence of generated tokens (target sentence) and $\mathbf{x}_i = x_{i1}, \cdots, x_{iS}$ be the sequence of input tokens (source sentence). Then, the probability of generating the target sequence is $\prod_{t=1}^{T} P(y_{it}|\mathbf{x}_i, \mathbf{y}_{i,<t})$. For simplicity, we use $P_{it}(y_{it})$ to represent $P(y_{it}|\mathbf{y}_{i,<t}, \mathbf{x}_i)$ and $C_{it}(y) = \delta(y_{it} = y)$ to denote if $y$ matches the correct label $y_{it}$.

The ECE can be expressed as follows:

$$\frac{1}{L} \sum_{m=1}^{M} | \sum_{i,tP_{it}(\hat{y}_{it}) \in B_m} C_{it}(\hat{y}_{it}) - P_{it}(\hat{y}_{it})| \quad (4)$$

where $L = \sum_{i=1}^{N} |\mathbf{y}_i|$ is the number of generated tokens.

Kumar and Sarawagi (2019) claimed that this measure focuses solely on the highest score label, neglecting the entire probability distribution, and thereby introduced *weighted ECE* for refined calibration. Another approach analyzes the overall correctness and the confidence of the answers directly, especially in tasks like classification and question answering (Lin et al., 2022; Kadavath et al., 2022). Huang et al. (2024) treated correctness as a distribution instead of a binary value. They assessed calibration by measuring the discrepancy between the model's confidence and its correctness, using Pearson correlation and Wasserstein similarity.

**Methods in discriminative models** Common methods for confidence estimation are logit-based (Pearce et al., 2021; Pereyra et al., 2017), ensemble-based and Bayesian (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016), density-based (Lee et al., 2018), and confidence-learning methods (DeVries and Taylor, 2018). Model calibration (Guo et al., 2017) can either occur during the model's training phase, e.g., by improving loss functions (Szegedy et al., 2016) or can be applied after the model has been trained, e.g., with temperature scaling (TS; Guo et al. 2017) and feature-based calibrators (FBC; Jiang et al. 2021). Table 1 shows the significant research in discriminative LMs, with a list of models, tasks, and calibration methods.

# 3 LLMs for Generation Tasks

## 3.1 Confidence Estimation

In this section, we divide the methods into white-box and black-box. We first provide a detailed overview of these methods and then we summarize their strengths, weaknesses, and connections.

### 3.1.1 White-Box Methods

White-box methods operate on the premise that the state at every position of the LLMs is accessible during inference.

**Logit-based methods** assess the sentence uncertainty using token-level probabilities or entropy (Huang et al., 2023b). To ensure that the evaluation is consistent across sentences of different lengths, the length-normalized likelihood probability is widely used (Murray and Chiang, 2018). Moreover, alternatives such as the minimum or the average token probabilities and the average entropy are also common (Vazhentsev et al., 2023b). Logit-based methods readily adapt to scenarios involving multiple samplings (Vazhentsev et al., 2023b) and ensembles (Malinin and Gales, 2021).

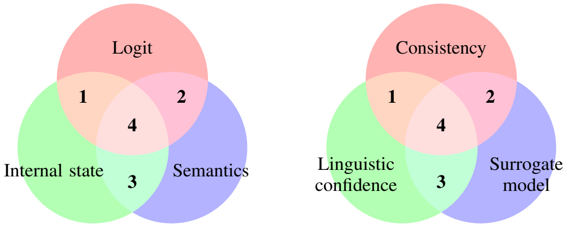
To incorporate semantics, Duan et al. (2023) introduced the concept of *token-level relevance*, which evaluates the relevance of the token by comparing the semantic change before and after moving the token with a semantic similarity metric like SBERT from Sentence Transformer (Reimers and Gurevych, 2019). Then, the sentence uncertainty can be adjusted based on the token's relevance. Duan et al. (2023) further proposed *sentence-level relevance* in multiple sampling settings, considering the similarity between the returned sentence and other sampled ones. Kuhn et al. (2023) proposed *semantic uncertainty*, which first clusters semantically equivalent samples based on the bidirectional entailment between samples and then approximates semantic entropy by summing the probabilities in each cluster. On the down side, these approaches use external models to access semantics, which adds computational costs, especially for the token level analysis (Duan et al., 2023).

Kadavath et al. (2022) discovered that LLMs can self-assess to differentiate between correct and incorrect answers. They suggested a method called *P(True)*, where the LLM first generates responses and then evaluates them as *True* or *False*. The probability the model assigns to the confidence level for the *True* label determines the confidence level.

**Internal state-based methods** Ren et al. (2023) introduced a technique for out-of-distribution detection and selective generation. The method starts by computing embeddings for both inputs and outputs in the training data, fitting them to a Gaussian distribution. It then assesses the model's confidence in its generated data by calculating the relative Mahalanobis distance of the evaluated data pair from this Gaussian distribution.

Recent studies have posited the existence of a direction in the activation space that effectively separates true and false inputs (Kadavath et al., 2022; Burns et al., 2023; Li et al., 2023; Azaria and Mitchell, 2023). Kadavath et al. (2022) proposed training a classifier (the probe), named P(IK), on the activations of the neural network to predict whether an LLM knows the answer. They sampled multiple answers for each question at a consistent temperature, labeled the correctness of each answer, and then used the question-correctness pair as training data. Similarly, Li et al. (2023) and Azaria and Mitchell (2023) used linear probes to examine whether attention heads in various layers can differentiate between correct and incorrect answers. Their empirical findings indicated that certain middle layers and a few attention heads exhibit strong performance in this task, although the layer positions vary across models. Burns et al. (2023) introduced an unsupervised approach to map hidden states to probabilities. It entails responding to questions with *Yes* and *No*, extracting and converting model activations into truth probabilities, and optimizing unsupervised loss for consistency. It ultimately gauges the model's confidence by estimating the likelihood of a *Yes* response.

**Summary** White-box methods, as illustrated in Figure 2a, primarily use logits, internal states, and semantics as sources of information. Logit-based approaches are easy to implement, but they face a limitation in that low logit probabilities may reflect various properties of language. Methods focusing on internal states (Kadavath et al., 2022; Li et al., 2023; Azaria and Mitchell, 2023) provide insights into the model's linguistic understanding, though they typically require supervised training on specially annotated data. Levinstein and Herrmann (2024) highlighted the limitations of the probing method in generalizing to unseen examples with negations. Semantics is often used to complement other methods, providing them with interpretability (Kuhn et al., 2023; Duan et al., 2023).



(a) White-box (b) Black-box

Figure 2: **Venn diagram of the taxonomy of information sources for white-box (Left) and black-box (Right) confidence estimation methods.** White-box methods rely on logit, internal state, or semantics, while black-box ones use consistency, linguistic confidence, or surrogate model, respectively. The intersections of these methods are located in Zones **1**–**4**.

To leverage the strengths of different methods, current advanced methods tend to combine different dimensions during confidence estimation. Recent work (Kuhn et al., 2023; Duan et al., 2023) achieved outstanding performance on uncertainty estimation for open-domain question answering by combining logit-based approaches with semantics, using bi-directional entailment or sentence encoders, aligning with Zone **2**. Rephrasing and round-trip translation can also be considered as using semantics to augment the remaining two methods (Jiang et al., 2021; Zhao et al., 2023c), corresponding to Zones **2** and **3**. P(True) leverages the self-evaluation capability of large language models (Kadavath et al., 2022). While it primarily uses logit probability, it is clear that this probability is influenced by internal states and semantics, related to Zone **4**. Anticipated advancements in collaborative information utilization will heighten computational demands, especially for nuanced semantic analysis (Duan et al., 2023). This underscores the need for a careful balance between performance and resource efficiency.

### 3.1.2 Black-Box Methods

Black-box methods assume access to the generations only, but no access to internal model activations or parameters.

**Linguistic confidence (verbalized method)** refers to prompting language models to express uncertainty in human language. This involves discerning different levels of uncertainty from the model's responses, such as "I don't know," "most probably," or "obviously" (Mielke et al., 2022).

This also includes prompting the model to output various verbalized words (e.g., *lowest*, *low*, *medium*, *high*, *highest*) or numbers (e.g., *85%*). Xiong et al. (2024) demonstrated that prompting strategies such as CoT (Wei et al., 2022), top-$k$ (Tian et al., 2023), and their proposed multi-step method can improve the calibration of linguistic confidence.

**Consistency-based estimation** assumes that a model's lack of confidence correlates with various responses, often leading to hallucinatory outputs. SelfCheckGPT (Manakul et al., 2023b) proposed a simple sampling-based approach that uses consistency among generations to find potential hallucinations. Five variants are utilized to measure the consistency: BERTScore (Zhang et al., 2020b), question-answering, n-gram, natural language inference (NLI) model (He et al., 2023), and LLM prompting. Lin et al. (2023) proposed to calculate the similarity matrix between generations and then estimate the uncertainty based on the analysis of the similarity matrix, such as the sum of the eigenvalues of the graph Laplacian, the degree matrix, and the eccentricity.

**Surrogate models** Shrivastava et al. (2023) introduced white-box models as surrogate models, like LLaMA-2 (Touvron et al., 2023b) and then used logit-based methods to estimate the confidence of the target model when prompted for the same task. They also demonstrated that integrating such confidence with linguistic confidence from black-box LLMs can provide better confidence estimates across various tasks.

**Summary** Figure 2b shows the information sources for confidence evaluation when the model states are not accessible: linguistic confidence, consistency, including lexical and semantic similarity, and surrogate models. Linguistic confidence can be elicited through prompts, but in practice, a mismatch between these has been observed (Lin et al., 2022; Liu et al., 2023c). Surrogate models (Shrivastava et al., 2023) facilitate white-box methods on black-box LLMs. However, they rely on the assumption of approximate parameter distribution of models, necessitating further work to validate their effectiveness. Consistency methods are computationally intensive, but have proven effective in various tasks. They can benefit the remaining two approaches (Zone **1** and **2**), such as the hybrid method proposed by Xiong et al. (2024).

| Study | Model | Proposed Methods |
|---|---|---|
| Duan et al. (2023) | OPT (Zhang et al., 2022) | SAR (Shifting Attention to Relevance): considers the semantic relevance when evaluating token and sentence-level uncertainty |
| Manakul et al. (2023b) | GPT-3 (Brown et al., 2020) | Semantic uncertainty: evaluates the consistency of the responses using various methods |
| Kuhn et al. (2023) | OPT (Zhang et al., 2022) | Clusters answers according to semantics and then computes the sum of the probabilities within each cluster to estimate confidence |
| Kadavath et al. (2022) | Anthropic LLM (Bai et al., 2022) | P(True): the probability a model assigns to its answer being True; P(IK) is the probability the model assigns to *I know* by leveraging a binary classifier |
| Xiong et al. (2024) | GPT3/3.5/4 (Brown et al., 2020), Vicuna (Chiang et al., 2023) | Hybrid methods combining linguistic confidence and consistency-based confidence |
| Lin et al. (2023) | GPT-3.5 | Estimates the confidence by evaluating the lexical and the semantic similarity between the responses |
| Shrivastava et al. (2023) | GPT-3.5/4, Claude | Hybrid methods combining confidence from the surrogate models and the linguistic confidence of the target models |

Table 2: **Recent studies of LLM confidence estimation**. These studies evaluate confidence estimation for question-answering tasks, using measures such as ECE, AUROC, etc.

Additionally, integrating all three methods (Zone **4**) has been explored by Shrivastava et al. (2023). Table 2 shows the representative research on confidence estimation for LLMs.

## 3.2 Calibration Methods

Here, we discuss related work in terms of calibration objectives: to enhance the quality of the generated text through calibration techniques and to improve the model's handling of unknowns or ambiguity by enabling it to express uncertainty more accurately. The first half of Table 3 presents recent work on calibrating LLMs over generation tasks.

### 3.2.1 Improving the Quality of Generation

Many studies (Kumar and Sarawagi, 2019; Wang et al., 2020; Lu et al., 2022) indicated that the miscalibration of token-level logit probabilities during generation is one of the reasons for the decline in generation quality. Kumar and Sarawagi (2019) introduced modified temperature scaling, where the temperature adjusts according to various factors, e.g., the entropy of the attention, token logits, token identity, and input coverage. Wang et al. (2020) noted a pronounced prevalence of over-estimated tokens compared to under-estimated ones. They introduced *graduated label smoothing*, applying heightened smoothing penalties to confident predictions. Xiao and Wang (2021) and van der Poel et al. (2022) calibrated the token probability separately by adding a weighted uncertainty estimated with model ensembles (Lakshminarayanan et al., 2017) and pointwise mutual information between the source and the target tokens. Zablotskaia et al. (2023) adapted diverse methods to improve model calibration in neural summarization.

Zhao et al. (2023b) suggested that MLE training can result in poorly calibrated sentence-level confidence, as the model has only been exposed to one gold reference. They proposed the *sequence likelihood calibration* (SLiC) technique to rectify this. It first generates $m$ multiple sequences $\{\hat{\mathbf{y}}\}_m$ from the initial model $\theta_0$, and then calibrates the model's confidence as follows:

$$\sum_{\{\mathbf{x},\bar{\mathbf{y}}\}} \mathcal{L}^{cal}(\theta, \mathbf{x}, \bar{\mathbf{y}}, \{\hat{\mathbf{y}}\}_m) + \lambda \mathcal{L}^{reg}(\theta, \theta_0, \mathbf{x}, \bar{\mathbf{y}}) \tag{5}$$

where the calibration loss $\mathcal{L}^{cal}$ aims to align models' decoded candidates' sequence likelihood according to their similarity to the reference $\bar{\mathbf{y}}$, and the regularization loss $\mathcal{L}^{reg}$ prevents models from deviating strongly. They further introduced SLiC-HF (Zhao et al., 2023a), which was designed to learn from human preferences.

### 3.2.2 Improving the Linguistic Confidence

Mielke et al. (2022) proposed a calibrator-controlled method for chatbots, which involves a trained calibrator to return the model confidence score and fine-tuned generative models to enable control over linguistic confidence. Lin et al. (2022) fine-tuned GPT-3 with a human-labeled dataset containing verbalized words and numbers to express uncertainty naturally. Zhou et al. (2023) empirically found that injecting expressions of uncertainty into prompts significantly increases the accuracy of GPT-3's answers and the calibration scores.

Different datasets (Amayuelas et al., 2023; Yin et al., 2023; Wang et al., 2024a; Liu et al., 2023a) have been presented containing questions that language models cannot answer or for which there is no clear answer.

| Study | Model | Task | Calibration Methods |
|---|---|---|---|
| Kumar and Sarawagi (2019) | LSTM (Bahdanau et al., 2015),<br>Transformer (Vaswani et al., 2017) | Machine Translation | TS with Learnable Parameters |
| Lu et al. (2022) | Transformer (Vaswani et al., 2017) | Machine Translation | Confidence-Based LS |
| Wang et al. (2020) | Transformer (Vaswani et al., 2017) | Machine Translation | LS, Dropout |
| Xiao and Wang (2021) | LSTM (Bahdanau et al., 2015),<br>Transformer (Vaswani et al., 2017) | Data2Text Generation,<br>Image Captioning | Uncertainty-Aware Decoding |
| van der Poel et al. (2022) | BART (Lewis et al., 2020) | Text Summarization | CPMI-Based Decoding |
| Zablotskaia et al. (2023) | T5 (Raffel et al., 2020) | Text Summarization | MC-Dropout, BE, SNGP,<br>DeepEnsemble |
| Zhao et al. (2023b) | PEGASUS (Zhang et al., 2020a) | Text Summarization,<br>Question Answering | SLiC |
| Zhao et al. (2023a) | T5 (Raffel et al., 2020) | Text Summarization | SLiC-HF |
| Mielke et al. (2022) | BlenderBot (Roller et al., 2021) | Dialogue Generation | Linguistic Calibration |
| Lin et al. (2022) | GPT-3 (Brown et al., 2020) | Math Question Answering | Fine-Tuning |
| Zhao et al. (2021) | GPT-3 (Brown et al., 2020) | Text Classification, Fact Retrieval<br>Information Extraction | Contextual Calibration |
| Fei et al. (2023) | PALM-2 (Anil et al., 2023),<br>CLIP (Radford et al., 2021) | Text Classification | Domain-Context Calibration |
| Han et al. (2023) | GPT-2 (Radford et al., 2019) | Text Classification | Prototypical Calibration |
| Kumar (2022) | GPT-2 (Radford et al., 2019) | Multiple Choice Question Answering | Answer-Level Calibration |
| Holtzman et al. (2021) | GPT-2(Radford et al., 2019),<br>GPT-3 (Brown et al., 2020) | Multiple Choice Question Answering | PMIDC |
| Zheng et al. (2024) | LLaMA (Touvron et al., 2023a),<br>Vicuna (Chiang et al., 2023),<br>Falcon (Penedo et al., 2023), GPT-3.5 | Multiple Choice Question Answering | PriDE |

Table 3: **Research on LLM calibration**. The first half of the table is about generation tasks, and the second half is about classification tasks. **Calibration methods:** LS: label smoothing, TS: temperature scaling, BE: Bayesian ensemble, SNGP: spectral-normalized Gaussian process, MCDropout: Monte Carlo dropout, SLiC: sequence likelihood calibration, HF: human feedback, FBC: feature-based calibrator, CPMI: conditional pointwise mutual information, PMIDC: domain conditional pointwise mutual information, PriDE: debiasing with prior estimation.

Amayuelas et al. (2023) analyzed how different large language models, including both smaller and open-source models, perform on a dataset of various unanswerable questions. They observed that LLMs showed varying accuracy levels depending on the question type, while smaller and open-source models tended to perform almost randomly for all question types. Liu et al. (2023a) evaluated both open-source models such as LLaMA-2 (Touvron et al., 2023b) and Vicuna (Chiang et al., 2023), and closed-source models such as GPT-3.5 and GPT-4, focusing on their refusal rate, accuracy, and uncertainty when handling unanswerable questions.

## 4 LLMs for Classification Tasks

Large language models are recognized for their efficiency in classification tasks, enabling rapid task implementation via prompting and few-shot learning (Brown et al., 2020; Zhao et al., 2021). Although the underlying principles of confidence estimation in a classification setup are similar to those for a generation setup, the objectives of the calibration and the approaches used differ significantly.

### 4.1 In-Context Learning

In-context learning (ICL) is a new learning paradigm with LLMs, where the model learns to perform a task based on a few examples and the context in which the task is presented. Assuming that $k$ selected input–label pairs $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_k, y_k)$ are given as demonstrations, with the predictive probability as the confidence, ICL makes predictions as follows:

$$\hat{y} = \arg\max_{y} P(y|\mathbf{x}_1, y_1, \cdots, \mathbf{x}_k, y_k, \mathbf{x}) \qquad (6)$$

When there are no demonstrations, the model performs zero-shot classification.

**Calibration methods** We refer to the input-label pairs as $\mathbf{C}$ for context, and to the original predictive probability as $P(y|\mathbf{C}, \mathbf{x})$. Zhao et al. (2021) introduced a method called *contextual calibration*. It gauges the model's bias with context-free prompts such as "[N/A]", "[MASK]" and an empty string. Then the context-free score is obtained by $\hat{\mathbf{P}}_{\texttt{cf}} = P(y|\mathbf{C}, \texttt{[N/A]})$. Subsequently, it transforms the scores with $\mathbf{W} = diag(\hat{\mathbf{p}}_{\texttt{cf}})^{-1}$ to offset the miscalibration.

Fei et al. (2023) proposed *domain-context calibration*, which first estimates the prior bias for each class using random text of an average sentence length and averaging the estimates $n$ times: $\bar{\mathbf{P}}_{rd}(y|\mathbf{C}) = \frac{1}{n}\sum_{i=1}^{n} P(y|\mathbf{C}, \texttt{[RANDOM TEXT]})$. Then, the prediction is obtained as follows:

$$\hat{y} = \arg\max_{y} \frac{P(y|\mathbf{C}, \mathbf{x})}{\bar{\mathbf{P}}_{rd}(y|\mathbf{C})} \tag{7}$$

Some methods aim to improve few-shot learning performance by combining classic statistical machine learning techniques. Nie et al. (2022) enhanced predictions by integrating a $k$-nearest-neighbor classifier with a datastore containing cached few-shot instance representations, while Han et al. (2023) introduced *prototypical calibration*, which uses Gaussian mixture models (GMM) to learn decision boundaries.

### 4.2 ICL Application: Multiple-Choice Question Answering

Multiple-choice question answering (MCQA) is an application of ICL, which is used in evaluating LLMs by prompting them to answer questions with predefined choices. The context $\mathbf{C}$ contains the question $\mathbf{q}$, and a set of options $\mathcal{I}(\mathbf{q}) = \{\mathbf{o}_1, \cdots, \mathbf{o}_K\}$, where each option is prefaced by an identifier such as *A, B*, and, if available, with a demonstration as an instruction.

Note that the implementation of the evaluation protocols can significantly impact the ranking of models. For instance, the original evaluation of the MMLU (Hendrycks et al., 2021) ranks the probabilities of the four option identifiers. The answer is considered correct when the highest probability corresponds to the correct option. The HELM implementation (Liang et al., 2023) considers probabilities over the complete vocabulary. The HARNESS implementation[1] prefers length-normalized probabilities of the entire answer sequence.

**Calibration Methods** Jiang et al. (2021) proposed various fine-tuning loss functions and temperature scaling for calibrating the performance of MCQA datasets. Additionally, they proposed techniques such as candidate output paraphrasing and input augmentation to calibrate the confidence. Holtzman et al. (2021) claimed that surface form competition occurs when different valid surface forms compete for probability.

Thus, they introduced *domain conditional pointwise mutual information*, which reweighs each option according to a term that is proportional to its prior likelihood within the context of the specific zero-shot task. To overcome the bias from the choice position, Zheng et al. (2024) proposed *PriDe*, which first decomposes the observed model prediction distribution into an intrinsic prediction over option contents and a prior distribution over option identifiers and then estimates the prior by permuting option contents on a small number of test samples. Kumar (2022) believed that under the neutral context $\mathbf{C}_\phi$, the probabilities of different options should be the same, but obviously, the LLM cannot meet this condition, so they proposed using $\log P(\mathbf{o}_k|\mathbf{C}) - sim(\mathbf{C}, \mathbf{C}_\phi) \log P(\mathbf{o}_k|\mathbf{C}_\phi)$ to make the prediction. Given that $\mathbf{C}$ is very similar to the neutral context $\mathbf{C}_\phi$, the approach will assign an equal score to each choice.

**Summary** The second half of Table 3 lists recent calibration studies over classification tasks. Current calibration methods primarily aim to mitigate biases associated with labels or choice positions in MCQA (Zhao et al., 2021; Jiang et al., 2021). A growing trend in the field is to deepen the understanding of the ICL (Holtzman et al., 2021) and to integrate semantics (Kumar, 2022). Besides, a systematic benchmark for evaluating different calibration methods is still missing.

## 5 Applications

**Hallucination Detection and Mitigation** Confidence or uncertainty can be applied as a signal for detecting and mitigating hallucinations of LLMs (Zhang et al., 2023b; Huang et al., 2023a). SelfCheckGPT (Manakul et al., 2023a) and $SAC^3$ (Zhang et al., 2023a) both explored hallucinations in the generation with self-consistency, while the latter also checked cross-model response consistency by taking generations from other models as a reference. Varshney et al. (2023) leveraged the model's logits to identify potential hallucinations, checked their correctness through a validation procedure, appended the repaired sentence to the prompt, and continued to generate. Fadeeva et al. (2024) proposed using token-level uncertainty quantification to detect hallucinations in biographies generated by LLMs. A similar idea was used to detect machine-generated text, based on perturbations in a white-box setup (Su et al., 2023).

[1] https://github.com/EleutherAI/lm-evaluation-harness/tree/v0.3.0

**Ambiguity Detection and Selective Generation** When identifying ambiguity in the data or unanswerable questions, reliable LLMs are anticipated to refrain from providing answers rather than generating responses arbitrarily (Kamath et al., 2020). Ren et al. (2023) proposed a selective generation method based on relative Mahalanobis distance. Zablotskaia et al. (2023) provided a comprehensive benchmark study that evaluates various calibration methods in neural summarization. Cole et al. (2023) and Hou et al. (2023) respectively used a disambiguate-and-answer approach and input clarification ensembling to measure data uncertainty for detecting ambiguous questions. Fadeeva et al. (2023) introduced *LM-Polygraph*, a framework with implementations of a battery of uncertainty estimation methods, focusing on improving selective generation of LLMs.

**Uncertainty-Guided Data Exploitation** Through measuring data uncertainty, the most representative instances will be selected for few-shot learning (Yu et al., 2023) or human annotation (SU et al., 2023). Regarding the knowledge enhancement to LLMs, Jiang et al. (2023) proposed an adaptive multi-retrieval method that first forecasts future content and then retrieves relevant documents stimulated by low-confidence tokens within the upcoming sentences.

## 6 Future Directions

**Comprehensive Benchmarks** The extensive utility of confidence estimation and calibration across numerous applications calls for a robust, multidimensional benchmark that covers a diverse array of tasks and domains. Moreover, fine-grained annotations of LLMs' responses, with an emphasis on long-form generation, are essential in fostering the development of more efficient approaches that improve the performance on intricate generation tasks (Tian et al., 2024; Huang et al., 2024). The extensive utility of confidence estimation and calibration across numerous applications calls for a robust, multidimensional benchmark that covers a diverse array of tasks and domains. Moreover, fine-grained annotations of LLMs' responses, with an emphasis on long-form generation, are essential in fostering the development of more efficient approaches that improve the performance on intricate generation tasks (Huang et al., 2024).

**Multi-Modal LLMs** By using additional pretraining with image–text pairings or by fine-tuning on specialized visual-instruction datasets, LLMs can be transited into the multimodal domain (Dai et al., 2023; Liu et al., 2023b; Zhu et al., 2024). However, it remains unclear whether these confidence estimation methods are effective for multimodal large language models (MLLMs) and whether these models are well-calibrated. Geng et al. (2024) found that on QA datasets focused on fact-checking, the ECE of GPT-4V's verbalized confidence is much lower than that of open-source models, which tend to be overly confident. We look forward to more efforts in detecting hallucinations in MLLMs through confidence estimation and in calibrating these models to discern events that are impossible in the real world.

**Calibration to Human Variation** Plank (2022) clarified the prevalent existence of human variation, i.e., humans have different opinions when labeling the same data. Human disagreement (Jiang and de Marneffe, 2022) can be attributed to task ambiguity (Tamkin et al., 2023), annotator's subjectivity (Sap et al., 2022), and input ambiguity (Meissner et al., 2021). Recent work (Baan et al., 2022; Lee et al., 2023) demonstrated misalignment between LLM calibration measures and human disagreement in various learning paradigms. Expressing the concern regarding different types of ambiguity (Xiong et al., 2024), abstaining from answering ambiguous questions (Yoshikawa and Okazaki, 2023), and further resolving ambiguity (Varshney and Baral, 2023) are necessary for trustworthy and reliable LLMs aligned with human variation.

## 7 Conclusion

This survey highlights the critical role of confidence estimation and calibration in addressing errors and biases in LLMs. The evolution of LLMs has paved the way for novel research opportunities and presented distinctive challenges. We first introduced the fundamental concepts of confidence and uncertainty, along with common metrics, estimation methods, and calibration techniques used in traditional discriminative models. We then identified the challenges these methods face in LLMs. Next, we delved into the latest research, introducing the principles, the advantages, and the drawbacks of various methods for generation and classification tasks. We concluded by discussing the current applications and future research directions.

## Limitations

**No experimental benchmarks** Without original experiments, we cannot offer empirical validation of the theories or the concepts that we discussed.

**Potential omissions** We made our best effort to compile the latest advancements. Due to the rapid development in the field, there is a possibility that we might have overlooked some important work.

## Ethical Considerations and Potential Risks

We anticipate no major ethical concerns for our work. Our review surveys the latest developments in this research field, and we did not conduct experiments, nor did we engage with risky datasets; we also did not employ any workers for manual annotation.

## Acknowledgement

Fengyu Cai is funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science, and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

## References

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *ArXiv preprint*, abs/2305.13712.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. PaLM 2 technical report. *ArXiv preprint*, abs/2305.10403.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *ArXiv preprint*, abs/2307.15703.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR'2015, San Diego, CA, USA.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862.

Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, NeurIPS'2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *Proceedings of the 11th International Conference on Learning Representations*, ICLR'2023, Kigali, Rwanda.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *ArXiv preprint*, abs/2308.16175.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, NeurIPS'2019, pages 2898–2909, Vancouver, BC, Canada.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*, NeurIPS'2023, New Orleans, LA, USA.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Terrance DeVries and Graham W Taylor. 2018. Learning confidence for out-of-distribution detection in neural networks. *ArXiv preprint*, abs/1802.04865.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *ArXiv preprint*, abs/2307.01379.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org.

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. 2024. Multimodal large language models to support real-world fact-checking. *arXiv preprint arXiv:2403.03627*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330, Sydney, NSW, Australia. JMLR.org.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2023. Prototypical calibration for few-shot learning of language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, ICLR'2023, Kigali, Rwanda.

Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. 2021. Training independent subnetworks for robust prediction. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR'2021.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the Eleventh International Conference on Learning Representations*, ICLR'2023, Kigali, Rwanda.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR'2021.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *ArXiv preprint*, abs/2311.08718.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv preprint*, abs/2311.05232.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *ArXiv preprint*, abs/2307.10236.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. *ArXiv preprint*, abs/2402.06544.

Abhyuday Jagannatha and Hong Yu. 2020. Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092, Online. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *ArXiv preprint*, abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5574–5584, Long Beach, CA, USA.

Jaeyoung Kim, Dongbin Na, Sungchul Choi, and Sungbin Lim. 2023. Bag of tricks for in-distribution calibration of pretrained transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 551–563, Dubrovnik, Croatia. Association for Computational Linguistics.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the Eleventh International Conference on Learning Representations*, ICLR'2023, Kigali, Rwanda.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, NeurIPS'2019, pages 12295–12305, Vancouver, BC, Canada.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *ArXiv preprint*, abs/1903.00802.

Sawan Kumar. 2022. Answer-level calibration for free-form multiple choice question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–679, Dublin, Ireland. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6402–6413, Long Beach, CA, USA.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, NeurIPS'2018, pages 7167–7177, Montréal, Canada.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Benjamin A. Levinstein and Daniel A. Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, NeurIPS'2023, New Orleans, LA, USA.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, ICCV'2017, pages 2999–3007, Venice, Italy.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *ArXiv preprint*, abs/2305.19187.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. 2023a. Prudent silence or foolish babble? Examining large language models' responses to the unknown. *ArXiv preprint*, abs/2311.09731.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, NeurIPS'2023, New Orleans, LA, USA.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023c. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. *ArXiv preprint*, abs/2308.05374.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning confidence for transformer-based neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2353–2364, Dublin, Ireland. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR2021, Virtual Event, Austria.

Potsawee Manakul, Yassir Fathullah, Adian Liusie, Vyas Raina, Vatsal Raina, and Mark Gales. 2023a. CUED at ProbSum 2023: Hierarchical ensemble of summarization models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 516–523, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023b. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. 2020. Confidence-aware learning for deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'2020*, pages 7034–7044. PMLR.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. 2020. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, NeurIPS'2020.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *ArXiv preprint*, abs/2212.02216.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–41, Long Beach, CA, USA. Computer Vision Foundation / IEEE.

Seo Yeon Park and Cornelia Caragea. 2022. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Tim Pearce, Alexandra Brintrup, and Jun Zhu. 2021. Understanding softmax confidence and uncertainty. *ArXiv preprint*, abs/2106.04972.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, New Orleans, LA, USA.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *Workshop Track Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *ICML'2021*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, ICLR'2023, Kigali, Rwanda.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your Transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what GPTs don't show: Surrogate models for confidence estimation. *ArXiv preprint*, abs/2311.08877.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongjin SU, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *Proceedings of the Eleventh International Conference on Learning Representations*, ICLR'2023, Kigali, Rwanda.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore. Association for Computational Linguistics.

Hao Sun, Boris van Breugel, Jonathan Crabbe, Nabeel Seedat, and Mihaela van der Schaar. 2022. What is flagged in uncertainty quantification? Latent density models for uncertainty categorization. *ArXiv preprint*, abs/2207.05161.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR'2016, pages 2818–2826, Las Vegas, NV, USA. IEEE Computer Society.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. 2023. Task ambiguity in humans and language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, ICLR'2023, Kigali, Rwanda.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *Proceedings of the Twelfth International Conference on Learning Representations*, ICLR'2024, Vienna, Austria.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Neeraj Varshney and Chitta Baral. 2023. Post-Abstention: Towards reliably re-attempting the abstained instances in QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 967–982, Toronto, Canada. Association for Computational Linguistics.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation. *ArXiv preprint*, abs/2307.03987.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, ICLR'2017, pages 5998–6008, Long Beach, CA, USA.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023a. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023b. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold Mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *ICML'2019*, pages 6438–6447, Long Beach, California, USA. PMLR.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *ArXiv preprint*, abs/2310.07521.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Xiaosu Wang, Yun Xiong, Beichen Kang, Yao Zhang, Philip S. Yu, and Yangyong Zhu. 2023c. Reducing negative effects of the biases of language models in zero-shot setting. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM'2023, page 904–912, New York, NY, USA. Association for Computing Machinery.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024a. Do-Not-Answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models in the year 2024. *arXiv preprint arXiv:2402.02420*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Annual Conference on Neural Information Processing*, volume 35 of *NeurIPS 2022*, pages 24824–24837, New Orleans, LA, USA.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proceeding of The Twelfth International Conference on Learning Representations*, ICLR'2024, Vienna, Austria.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028, Dubrovnik, Croatia. Association for Computational Linguistics.

Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. 2023. Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2499–2521, Toronto, Canada. Association for Computational Linguistics.

Polina Zablotskaia, Du Phan, Joshua Maynez, Shashi Narayan, Jie Ren, and Jeremiah Liu. 2023. On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2980–2992, Singapore. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *Proceeding of Sixth International Conference on Learning Representations*, ICLR'2018, Vancouver, BC, Canada.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023a. SAC$^3$: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'2020*, pages 11328–11339, Virtual Event. PMLR.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1958–1970, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *Proceedings of Eighth International Conference on Learning Representations*, ICLR'2020, Addis Ababa, Ethiopia.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *ArXiv preprint*, abs/2309.01219.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023a. SliC-HF: Sequence likelihood calibration with human feedback. *ArXiv preprint*, abs/2305.10425.

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. 2023b. Calibrating sequence likelihood improves conditional language generation. In *Proceedings of the Eleventh International Conference on Learning Representations*, ICLR'2024, Kigali, Rwanda.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023c. Knowing what LLMs do not know: A simple yet effective self-detection method. *ArXiv preprint*, abs/2310.17918.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *ICML'2021*, pages 12697–12706, Virtual Event. PMLR.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *Proceedings of the Twelfth International Conference on Learning Representations*, ICLR'2024, Vienna, Austria.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795, Singapore. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*, ICLR'2024, Vienna, Austria.

# A Appendix

## A.1 Confidence Estimation Methods

The methods for confidence estimation can generally be categorized into the following groups:

**Logit-based estimation** Given the model input $\mathbf{x}$, the logit $\mathbf{z}$, along with the prediction $\hat{y}$ (i.e., the class with the highest probability emitted by softmax activation $\sigma$), the model confidence is estimated directly using the probability value:

$$\mathtt{conf}_{sp}(\mathbf{x}, \hat{y}) = P(\hat{y}|\mathbf{x}) = \sigma(\mathbf{z})_{\hat{y}} \tag{8}$$

The confidence can also be estimated based on transformations of the logits, such as examining the gap between the two highest ones (Yoshikawa and Okazaki, 2023) or by using entropy, which indicates the uncertainty with a larger value.

**Ensemble-based & Bayesian methods** *Deep ensemble methods* (Lakshminarayanan et al., 2017) train multiple neural networks independently and estimate the uncertainty by computing the variance of the outputs from these models. *Monte Carlo dropout* (MCDropout, Gal and Ghahramani 2016) methods extend the dropout techniques to estimating uncertainty. As in the training phase, dropout is also applied during inference, and multiple forward passes are performed to obtain predictions. The final prediction is obtained through averaging the predictions, with the variability of the predictions reflecting the model uncertainty.

Methods such as deep-ensemble and MC-Dropout introduce a heavy computational overhead, especially when applied to LLMs (Malinin and Gales, 2021; Shelmanov et al., 2021; Vazhentsev et al., 2022), and there is the need to optimize the computation. For example, determinantal point process (Kulesza and Taskar, 2012) can be applied to facilitate MCDropout by sampling diverse neurons in the dropout layer (Shelmanov et al., 2021).

**Density-based estimation** approaches (Lee et al., 2018; Yoo et al., 2022) assume that the regions of the input space where the training data is dense are the regions where the model is likely to be more confident in its predictions. Conversely, regions with sparse training data are areas of higher uncertainty. Lee et al. (2018) first proposed a Mahalanobis distance-based confidence score, which calculates the distance between one test point and a Gaussian distribution fitting the test data. The confidence estimation is obtained by exponentiating the negative value of the distance.

**Confidence learning** uses a specific network branch to learn the confidence of model predictions. DeVries and Taylor (2018) leveraged a confidence estimation branch to forecast scalar confidence, and the original probability is modified by interpolating the ground truth according to the confidence to provide "hints" during the training process. Additionally, it discourages the network from always asking for hints by applying a small penalty. Corbière et al. (2019) empirically demonstrated that the confidence based on true class probability (TCP) is better for distinguishing between correct and incorrect predictions. Given the ground truth $y$, TCP can be represented as $P(y|\mathbf{x})$. However, $y$ is not available when estimating the confidence of the predictions. Hence, Corbière et al. (2019) used a confidence learning network to learn TCP confidence during training.

## A.2 Model Calibration

Calibration methods can be categorized based on their execution time as *in-training* and *post-hoc* methods.

### A.2.1 In-Training Calibration

Research indicates that model generalization methods can be used for calibration (Kim et al., 2023), and calibration methods can enhance model performance, particularly in out-of-domain generation (Desai and Durrett, 2020).

**Novel loss functions** Many studies considered the *cross-entropy* (CE) loss to be one of the causes leading to model miscalibration (Mukhoti et al., 2020; Kim et al., 2023). Mukhoti et al. (2020) demonstrated that *focal loss* (Lin et al., 2017), designed to give more importance to hard-to-classify examples and to down-weigh the easy-to-classify examples, can improve the calibration of neural networks. The *correctness ranking loss* (CRL; Moon et al. 2020) calibrated models by penalizing incorrect rankings within the same batch and by using the difference in proportions as the margin to differentiate sample confidence. Besides, *entropy regularization loss* (ERL; Pereyra et al. 2017) and *label smoothing* (LS; Szegedy et al. 2016) were introduced to discourage overly confident output distributions.

**Data augmentation** involves creating new training examples by applying various transformations or perturbations to the original data. It has been widely used for calibration of discriminative LMs by alleviating the issue of over-confidence, such as MixUp (Zhang et al., 2018), EDA (Wei and Zou, 2019), Manifold-MixUp (Verma et al., 2019), MIMO (Havasi et al., 2021), and AUM-guided MixUp (Park and Caragea, 2022).

**Ensemble and Bayesian methods** were initially introduced to quantify model uncertainty. However, both can also be valuable for model calibration, as they can enhance accuracy, mitigate overfitting, and reduce overconfidence (Kong et al., 2020; Kim et al., 2023).

### A.2.2 Post-Hoc Calibration

**Scaling methods** are exemplified by *matrix scaling*, *vector scaling* and *temperature scaling* (Guo et al., 2017). Using a validation set, they fine-tune the predicted probabilities to better align with the true outcomes, leveraging the *negative log-likelihood* (NLL) loss. Among them, temperature scaling (TS) is popular due to its low complexity and efficiency. It involves re-weighing the logits before the softmax function by a learned scalar $\tau$, known as the *temperature*.

**Feature-based calibrator** leverages both the input features and the model predictions to refine the predicted probabilities. To train the calibrator, one first applies a trained model on a validation dataset. Subsequently, both the original input features and the model's predictions from this dataset are passed to a binary classifier (Jagannatha and Yu, 2020; Jiang et al., 2021; Si et al., 2022).

## A.3 Summary

**Confidence estimation** Logit-based methods stand out as the most straightforward to implement and interpret. Reducing the computational cost and improving the sampling efficiency pose challenges to ensemble-based and Bayesian methods. Density-based estimation can be used to identify which data points are associated with different types of uncertainties. However, it makes assumptions about data distribution (Baan et al., 2023) and can also be computationally intensive when dealing with large datasets (Sun et al., 2022). Confidence learning can acquire task-relevant confidence; however, it requires modifying the neural network and performing specific training.

**Model calibration** Post-hoc methods are generally model-independent and can calibrate the probabilities without impacting the model's performance (Guo et al., 2017). Desai and Durrett (2020) empirically found that temperature scaling effectively reduces the calibration error when in-domain, whereas label smoothing is more beneficial in out-of-domain settings. Kim et al. (2023) found that augmentation can enhance both classification accuracy and calibration performance. However, ensemble methods may sometimes degrade model calibration if individual members produce similar predictions due to overfitting. Table 1 represents significant work in calibrating discriminative LMs. We have comprehensively listed the models,the tasks, and the calibration methods they used.