A Survey of Confidence Estimation and Calibration in Large Language Models

Anonymous ACL submission

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 language-based reasoning, and in-context learning (Petroni et al., 2019; Wei et al., 2022; Brown et al., 2020a). However, LLMs do not consistently achieve good performance (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). Evaluating the trustworthiness of responses from these models remains challenging (Liu et al., 2023c).

Confidence (or uncertainty) estimation is crucial for tasks like 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; Ren et al., 2022; Vazhentsev et al., 2023b). A related concept is that of model calibration, which focuses on aligning predictive probabilities (estimated confidence) to actual accuracy (Guo et al., 2017). LLMs show unique properties in this regard, such as expressing confidence in words (Lin et al., 2022; Xiong et al., 2023) and the ability to perform zero-shot or few-shot learning (Brown et al., 2020a). However, their responses can be sensitive to the prompts, e.g., the examples provided and their order, which can cause a lot of instability of the results. In view of 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 in LLMs (Zhang et al., 2023b; Wang et al., 2023b), there are no comprehensive surveys systematically discussing the technical advancements in LLMs, and here we aim to bridge this gap.

We explore the unique challenges posed by LLMs and examining the latest studies addressing these issues. We first discuss key concepts such as confidence, uncertainty, and calibration in the context of neural models, as detailed in Section 2. This is followed by a focused discussion on the specific challenges associated with LLMs (Section 3). We pursue two different directions: one addressing confidence estimation and calibration techniques for generation tasks in Section 4, and the other for classification tasks in Section 5. We conclude by exploring their practical applications (Section 6) and looking at potential future research directions (Section 7).

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*

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

Table 1: **Recent studies of LLM confidence estimation**. These studies evaluate confidence estimation in question-answering tasks, utilizing metrics such as ECE, AUROC, etc.

and *absolute confidence score*, offering different methods to assess and to interpret confidence levels (Kamath et al., 2020; Vazhentsev et al., 2023a). Given input x, ground truth y, and prediction $\hat{y}$, 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 of (x_i, y_i) and (x_j, y_j) and their corresponding predictions $\hat{y}_i$ and $\hat{y}_j$, we 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 in real-world scenarios. For example, if a model predicts an event with a 70% probability, that event should actually happen about 70% of the time under similar circumstances. The equation for this relationship is s 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 categorizing uncertainty in machine learning into *aleatoric* and *epistemic* uncertainty. Aleatoric or data uncertainty emerges from the inherent randomness or 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 techniques (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017).

2.2 Metrics and Methods

Metrics 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. The predicted probabilities are first segmented into various bins. ECE is then calculated by taking the weighted average of the discrepancies between the mean predicted probability and the actual accuracy across all bins. One drawback of the ECE metric is its sensitivity to various factors such as bucket width and the variance of samples within these buckets. To overcome these issues, 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, which plots predicted probabilities against observed frequencies, with points or lines above the diagonal indicating overconfidence. Additionally, metrics such as 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.

Methods in discriminative models Common methods for confidence estimation include logit-based methods (Pearce et al., 2021; Pereyra et al., 2017), ensemble-based and Bayesian methods (Lakshminarayanan et al., 2017; Gal and

Study	Model	Task	Calibration Methods
Kumar and Sarawagi (2019)	LSTM (Bahdanau et al., 2015), 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), Transformer (Vaswani et al., 2017)	Data2Text Generation, 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, DeepEnsemble
Zhao et al. (2022)	PEGASUS (Zhang et al., 2020a)	Text Summarization, 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., 2020b)	Math Question Answering	Fine-Tuning
Zhao et al. (2021)	GPT-3 (Brown et al., 2020b)	Text Classification, Fact Retrieval Information Extraction	Contextual Calibration
Fei et al. (2023)	PALM-2 (Anil et al., 2023), CLIP (Radford et al., 2021)	Text Classification	Domain-Context Calibration
Han et al. (2022)	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), GPT-3 (Brown et al., 2020b)	Multiple Choice Question Answering	PMIDC
Zheng et al. (2023)	LLaMA (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), Falcon (Penedo et al., 2023), GPT-3.5	Multiple Choice Question Answering	PriDE

Table 2: **Studies of LLM calibration**. The first half 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.

Ghahramani, 2016), density-based methods (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, for example, by improving loss functions (Szegedy et al., 2016), or be applied after the model has been trained, such as temperature scaling (TS; Guo et al. 2017) and feature-based calibrators (FBC; Jiang et al. 2021). Table 3 represents significant research in the discriminative LMs, with a list of models, tasks, and calibration methods. Due to space limitation, please refer to the Appendix A for detailed principles and comparisons.

3 Challenges of Confidence Estimation and Calibration in LLMs

This section elaborates on the challenges related to confidence estimation and calibration of LLMs.

Exponential output space growth Discriminative models readily provide probability scores for distinct categories. In contrast, LLMs encounter a significant challenge due to the exponential increase in their output space as the sentence length grows. This increase renders it impossible to assess all possible predictions. This complexity hinders the effective calculation of confidence or uncertainty metrics (Wang et al., 2023a).

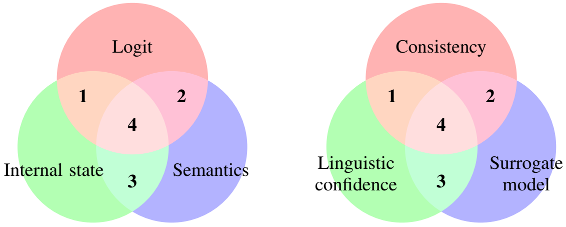
Semantics Unlike fixed-category labels, the outputs of LLMs capture the diversity of natural language, where the same words or sentences can have varied meanings across different contexts, while at the same time, superficially distinct phrases may convey the same meaning (Kuhn et al., 2023). Additionally, within the same sentence, different tokens have varying levels of semantic importance. Some lengthy sentences can be almost entirely expressed using just a few keywords (Duan et al., 2023).

Emergent capabilities and diverse tasks The emerging capabilities of LLMs, allow to evaluate the truthfulness of their answers or to express uncertainty when addressing unknown or ambiguous questions, which presents new research directions (Kadavath et al., 2022; Lin et al., 2022; Amayuelas et al., 2023). Moreover, methods and motivations are distinctly different when calibrating LLMs in generation and classification tasks (Duan et al., 2023; Zhao et al., 2021; Kuhn et al., 2023). Most work on classification tasks focuses on mitigating prior biases across different categories (Zhao et al., 2021; Jiang et al., 2021). However, it often involves enabling LLMs to produce better outputs or to express more precise confidence in generation tasks (Zhao et al., 2022; Mielke et al., 2022). Therefore, we further discuss recent advances in generation tasks and classification tasks in Section 4 and Section 5, respectively.

4 LLMs for Generation Tasks

4.1 Confidence Estimation

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



(a) White-box

(b) Black-box

Figure 1: Venn diagram: the taxonomy of information sources for white-box (**Left**) and black-box (**Right**) confidence estimation methods. These two families of methods can be categorized into the methods relying on logit, internal state, or semantics, and those relying on consistency, linguistic confidence, or surrogate model, respectively. The intersections of these methods are located in Zone **1** - **4**.

4.1.1 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 The logit-based method evaluates sentence uncertainty using token-level probabilities or entropy. Assuming that $\mathbf{y} = y_1, \cdots, y_T$ denotes the sequence of generated tokens (target sentence), and that $\mathbf{x} = x_1, \cdots, x_S$ denotes the sequence of input tokens (source sentence), the sentence confidence can be represented by the factorized probability: $\prod_{i=1}^{T} P(y_i|\mathbf{x}, \mathbf{y}_{<i})$. To ensure an evaluation consistent across sentences of different lengths, the length-normalized likelihood probability is widely utilized (Murray and Chiang, 2018). Moreover, alternatives such as the minimum or average token probabilities and the average entropy are also widely used (Vazhentsev et al., 2023b). Logit-based techniques readily adapt to scenarios involving multiple sampling (Vazhentsev et al., 2023b) or ensemble models (Malinin and Gales, 2021a).

To incorporate semantics, Duan et al. (2023) introduced the concept of *token-level relevance*, which evaluates the relevance of the token by comparing semantic change before and after moving the token with a semantic similarity metric like Sentence Transformer (Reimers and Gurevych, 2019). Then, 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 probabilities in each cluster.

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 the confidence level to "True" determines the confidence level.

Internal state-based methods Ren et al. (2022) 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 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 a 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 the training data. Similarly, Li et al. (2023) and Azaria and Mitchell

(2023) employed 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" or "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 1a, primarily utilize logits, internal states, and semantics as sources of information. Logit-based approaches, easy to implement during inference, 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. Semantics are often used to complement other methods, providing them with interpretability (Kuhn et al., 2023; Duan et al., 2023).

To leverage their respective strengths, the current advanced methods tend to combine different dimensions during confidence estimation. Recent works (Kuhn et al., 2023; Duan et al., 2023) achieve outstanding performance on uncertainty estimation for open-domain question answering by combining logit-based approaches with semantics, using tools like 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., 2023b), 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**. We anticipate better collaborative use of diverse information types in the future.

4.1.2 Black-box Methods

Black-box methods assume that all parameters during inference are unknown, allowing access only to the generations.

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) or prompting the model to output various verbalized words (e.g., "lowest", "low", "medium", "high", "highest") or numbers (e.g., "85%"). Xiong et al. (2023) demonstrated that prompting strategies like 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 employed logit-based methods to estimate the confidence of the target model when prompted with the same task. They also showed that integrating such confidence with linguistic confidence from black-box LLMs can provide better confidence estimates across various tasks.

Summary Figure 1b illustrates the information sources for confidence evaluation when 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. (2023). Additionally, integrating all three methods (Zone **4**) has been explored by Shrivastava et al. (2023) to offer further benefits. Table 1 presents the latest representative works in confidence estimation for large language models, briefly describing their proposed methods.

4.2 Calibration Methods

This section categories related work in terms of calibration objectives: to enhance the quality of generated text through calibration techniques and to improve the model's handling of unknown or ambiguous issues by enabling it to express uncertainty more accurately. The first half of Table 2 presents recent work on calibrating LLMs over generation tasks.

4.2.1 Improve 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 a modified temperature scaling approach where the temperature value adjusts according to various factors, including the entropy of attention, token logit, 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 tasks.

Zhao et al. (2022) suggested that MLE training can result in poorly calibrated sentence-level confidence, as the model is only 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 θ_0, then calibrates the model's confidence with:

$$\sum_{\{\mathbf{x},\bar{\mathbf{y}}\}} \mathcal{L}^{cal}(\theta, \mathbf{x}, \bar{\mathbf{y}}, \{\hat{\mathbf{y}}\}_m) + \lambda \mathcal{L}^{reg}(\theta, \theta_0, \mathbf{x}, \bar{\mathbf{y}}) \tag{3}$$

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.

4.2.2 Improve 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 the 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., 2023d; Liu et al., 2023a) have been presented on questions that language models cannot answer or for which there is no clear answer. Amayuelas et al. (2023) analyzed how different language models, including both smaller and open-source models, classify a dataset of various unanswerable questions. They observed that LLMs show varying accuracy levels depending on the question type, while smaller and open-source models tend to perform almost randomly. Liu et al. (2023a) evaluated both open-source models like LLaMA-2 (Touvron et al., 2023b), 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 in handling unanswerable questions.

5 LLMs for Classification Tasks

LLMs are recognized for their efficiency in classification tasks, enabling rapid task implementation via prompts (Brown et al., 2020a; Zhao et al., 2021). Although the underlying principles of confidence estimation are similar to those in generation tasks, the objectives of calibration and the approaches differ significantly.

5.1 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:

$$\hat{y} = \arg\max_{y} P(y|\mathbf{x}_1, y_1, \cdots, \mathbf{x}_k, y_k, \mathbf{x}) \qquad (4)$$

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 the original predictive probability is denoted 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 estimates the prior bias for each class with n times model average with random text of an average sentence length: $\bar{\mathbf{P}}_{rd}(y|\mathbf{C}) = \frac{1}{n}\sum_{i=1}^{n} P(y|\mathbf{C}, \texttt{[RANDOM TEXT]})$. The prediction is obtained with:

$$\hat{y} = \arg\max_{y} \frac{P(y|\mathbf{C}, \mathbf{x})}{\bar{\mathbf{P}}_{rd}(y|\mathbf{C})} \qquad (5)$$

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. (2022) introduced *prototypical calibration*, which employs Gaussian mixture models (GMM) to learn decision boundaries.

5.2 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 the set of options $\mathcal{I}(\mathbf{q}) = \{\mathbf{o}_1, \cdots, \mathbf{o}_K\}$, where each is prefaced with an identifier such as "A", and, if available, with a demonstration as an instruction.

It is worth noting that implementing 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., 2022) 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 MQCA 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* (PMIDC), 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. (2023) 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 2 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

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

to integrate semantics (Kumar, 2022). Besides, a systematic benchmark for evaluating different calibration methods is still missing.

6 Applications

Confidence estimation and calibration can be effectively employed in the following applications as an indispensable component in ensuring reliable AI.

Hallucination detection and mitigation Confidence or uncertainty can be applied as a signal for the detection and mitigation generated by LLMs. 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 the reference. Varshney et al. (2023) proposed a method that leverages the model's logits to identify potential hallucinations, checks their correctness through a validation procedure, appends the repaired sentence to the prompt, and continues to generate.

Ambiguity detection and selective generation When identifying ambiguity in 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. (2022) 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 employed a disambiguate-and-answer approach and input clarification ensembling to measure data uncertainty for detecting ambiguous questions.

Uncertainty-guided data exploitation Through measuring data uncertainty, the most representative instances will be selected for few-shot learning (Yu et al., 2022) or human annotation (Su et al., 2022). Regarding the knowledge enhancement to LLMs, Jiang et al. (2023) proposed an adaptive multi-retrieval method that first forecasts future content and retrieves relevant documents stimulated by low-confidence tokens within upcoming sentences.

7 Future Directions

Multi-modal LLMs By employing additional pre-training 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., 2023). However, it remains unclear whether these confidence estimation methods are effective for multimodal large language models (MLLMs) and whether these models are well-calibrated. 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., 2022), 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 the misalignment between LLM calibration measures and human disagreement in various learning paradigms. Expressing the concern regarding different types of ambiguity (Xiong et al., 2023), 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.

8 Conclusion

This survey highlights the critical role of confidence estimation and calibration in addressing errors and biases in large language models (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, advantages, and drawbacks of various methods in generation and classification tasks. We concluded by discussing the current applications and future research directions.

Limitations

This survey mainly has the following limitations:

No experimental benchmarks Without original experiments, this paper cannot offer empirical validation of the theories or concepts. This limits the paper's ability to contribute new, verified knowledge to the field.

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

Ethical Considerations and Potential Risks

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

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 its lying. *ArXiv preprint*, abs/2304.13734.

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 *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *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 R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *ArXiv preprint*, abs/2305.14613.

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 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2898–2909.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint*, abs/2305.06500.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 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.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. *ArXiv preprint*, abs/2305.19148.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

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, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2022. Prototypical calibration for few-shot learning of language models. *ArXiv preprint*, abs/2205.10183.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

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 F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *ArXiv preprint*, abs/2305.06983.

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, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584.

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. *ArXiv preprint*, abs/2302.09664.

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, December 8-14, 2019, Vancouver, BC, Canada*, pages 12295–12305.

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, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

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, December 3-8, 2018, Montréal, Canada*, pages 7167–7177.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models infer and disagree like humans? *ArXiv preprint*, abs/2305.13788.

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. *ArXiv preprint*, abs/2306.03341.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *ArXiv preprint*, abs/2205.14334.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

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. *ArXiv preprint*, abs/2304.08485.

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 J. F. Gales. 2021a. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Andrey Malinin and Mark J. F. Gales. 2021b. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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 JF Gales. 2023b. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv preprint*, abs/2303.08896.

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.

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, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 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, December 6-12, 2020, virtual*.

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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 38–41. 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, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *ArXiv preprint*, abs/2306.01116.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *ArXiv preprint*, abs/1701.06548.

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, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. 2022. Out-of-distribution detection and selective generation for conditional language models. *ArXiv preprint*, abs/2209.15558.

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, et al. 2022. Selective annotation makes language models better few-shot learners. *ArXiv preprint*, abs/2209.01975.

Hao Sun, Boris van Breugel, Jonathan Crabbe, Nabeel Seedat, and Mihaela van der Schaar. 2022. Daux: a density-based approach for uncertainty explanations. *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 *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. 2022. Task ambiguity in humans and language models. *ArXiv preprint*, abs/2212.10711.

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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. 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.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023b. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *ArXiv preprint*, abs/2310.07521.

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 '23, page 904–912, New York, NY, USA. Association for Computing Machinery.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023d. Do-not-answer: A dataset for evaluating safeguards in LLMs. *ArXiv preprint*, abs/2308.13387.

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. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *ArXiv preprint*, abs/2306.13063.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *ArXiv preprint*, abs/2305.18153.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation. *ArXiv preprint*, abs/2203.01677.

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. 2022. Cold-start data selection for few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. *ArXiv preprint*, abs/2209.06995.

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. *ArXiv preprint*, abs/2304.08653.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Kumar Sricharan. 2023a. Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *ArXiv preprint*, abs/2311.01740.

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, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. 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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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. 2022. Calibrating sequence likelihood improves conditional language generation. *ArXiv preprint*, abs/2210.00045.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023b. 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, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *ArXiv preprint*, abs/2309.03882.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *ArXiv preprint*, abs/2302.13439.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, abs/2304.10592.

Study	Model	Task	Calibration Methods
(Desai and Durrett, 2020)	BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)	Nature 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)	Nature 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 3: **Studies of 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

A Appendix

A.1 Confidence Estimation Methods

The methods for confidence estimation have been extensively studied and 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 σ), the model confidence is estimated directly using the probability value:

$$\texttt{conf}_{sp}(\mathbf{x}, \hat{y}) = P(\hat{y}|\mathbf{x}) = \sigma(\mathbf{z})_{\hat{y}} \qquad (6)$$

There are methods for estimating confidence based on transformations of the logit probabilities, such as examining the gap between the two highest probabilities (Yoshikawa and Okazaki, 2023) or utilizing 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 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, 2021b; 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 Density-based approaches (Lee et al., 2018; Yoo et al., 2022) are based on the assumption that regions of the input space where training data is dense are 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 test data. The confidence estimation is obtained by exponentiating the negative value of the distance.

Confidence learning employs 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 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-weight 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-weighting the logits before the softmax function by a learned scalar τ, known as the *temperature*.

Feature-based calibrator leverages both input features and 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 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 requires assumptions of 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 probabilities without impacting the model's performance (Guo et al., 2017). Desai and Durrett (2020) empirically found that temperature scaling effectively reduces calibration error 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 3 represents significant work in calibrating discriminative LMs. We have comprehensively listed the models, tasks, and calibration methods they employed.