Random-set Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large Language Models (LLMs) are known to produce very high-quality texts and responses to our queries. But how much can we trust this generated text? In this paper, we study the problem of uncertainty quantification in LLMs. We propose a novel Random-Set Large Language Model (RS-LLM) approach which predicts finite random sets (belief functions) over the token space, rather than probability vectors as in classical LLMs. In order to allow so efficiently, we also present a methodology based on hierarchical clustering to extract and use a budget of "focal" subsets of tokens upon which the belief prediction is defined, rather than using all possible collections of tokens, making the method scalable yet effective. RS-LLMs encode the epistemic uncertainty induced in their generation process by the size and diversity of its training set via the size of the credal sets associated with the predicted belief functions. The proposed approach is evaluated on multiple datasets using Llama2-7b, Llama3.1-8b, Mistral-7b and Phi-2 models and compared against existing uncertainty methods, and is shown to outperform the standard model and existing uncertainty methods in both datasets in terms of correctness of answer, while also showing potential in estimating the second-level uncertainty of its predictions and empowering the ability to detect when it is hallucinating.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Anil et al., 2023; Touvron et al., 2023) have made significant waves in the recent years and have shown to perform very well on a wide range of NLP tasks like question-answering Wu et al. (2023), common sense reasoning Wei et al. (2022), mathematical problem-solving Lewkowycz et al. (2022) and code generation Roziere et al. (2023). Typically, these models are pre-trained on a large corpus of text to predict the next token in an unsupervised fashion. To make these models usable, they are further fine-tuned on a particular application Cobbe et al. (2021) and aligned to make sure the model behaves in an ethical manner and according to the human preferences Bai et al. (2022).

However, LLMs still have limitations in their capacity to understand information and often produce false statements or "hallucinations" Maynez et al. (2020). This makes them less trustworthy and causes hindrance to deploying LLMs in high-stake decision-making applications where the consequences of incorrect decisions are severe. Therefore, there needs to be a mechanism to make LLMs more truthful (calibration) and to associate uncertainty estimation with the model's generation Manchingal et al. (2025b). Furthermore, they need to distinguish the source of this uncertainty Cuzzolin (2024); Manchingal et al. (2025a), as the latter can be of either aleatoric (relating to chance) or epistemic (relating to knowledge) nature (Kendall & Gal, 2017; Hüllermeier & Waegeman, 2021). However, LLMs, as they are now, are not equipped to express or reason about uncertainty in a structured or interpretable way.

Next-token prediction in language models can be thought of simply as classification over the all the tokens in the vocabulary. Therefore, it appears natural to apply to language models uncertainty quantification approaches developed for classification tasks, However, LLM-scale modeling demands solutions that are both scalable and interpretable.

In this paper, we take a principled and novel step in this direction by proposing Random-Set Large Language Models (RS-LLMs)—a new class of LLMs that model uncertainty using belief functions derived from the mathematical theory of random sets (Molchanov, 2005; 2017; Cuzzolin, 2023). As the vocabulary in LLMs lives in a finite space, we model the random sets using *belief functions* Shafer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

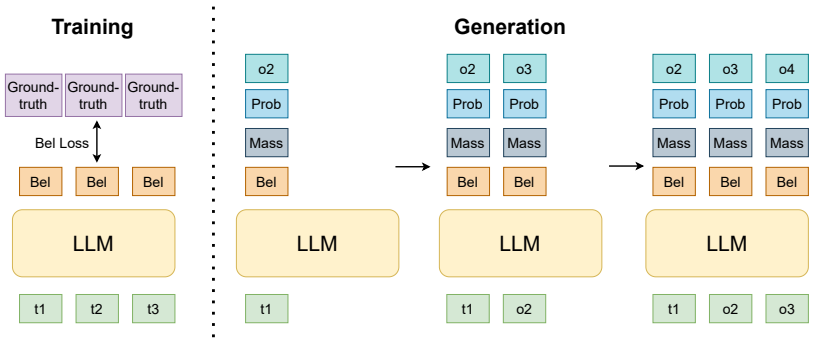


Figure 1: Training and generation flow of RS-LLM. Training is performed in a parallel fashion using the teacher forcing method. Generation is done sequentially. For each token, the model predicts a belief function. Then the mass function, probability distribution and next token is subsequently computed/sampled from that belief function.

(1976a), the finite incarnation of random sets Cuzzolin (2020); Manchingal & Cuzzolin (2022). They assign probabilities to sets of tokens directly, in a non-additive way, and can therefore naturally model the fact that observations come in the form of sets, e.g. when data components are missing either in the feature space or in time. Our approach builds on the recent success of Random-Set Neural Networks (RS-NNs) in classification (Manchingal et al., 2025d), where random sets outperformed Bayesian and ensemble methods in both accuracy and uncertainty quantification.

Instead of predicting a probability distribution over the vocabulary for the next token, an RS-LLM predicts a belief function Cuzzolin (2010c) over the sets of tokens in the vocabulary. However, it is computationally impossible to work with all possible sets of tokens. To address this, we propose an ingenious mechanism, together with an analysis, to select a fixed budget of most appropriate focal sets based on hierarchical clustering (Müllner, 2011; Ackermann et al., 2014). The predicted belief function assigns a belief value $bel(A) \in [0, 1]$ to all $A \in \mathcal{O}$, where $\mathcal{O}$ is a finite budget of sets of tokens in the vocabulary. Such a belief function Cuzzolin (2014) is mathematically equivalent to a convex set of probability vectors (*credal set* (Levi, 1980; Cuzzolin, 2008)) on the vocabulary of tokens and can be interchangeably be represented by a mass function with mass values $m(A) \in [0, 1] \; \forall \; A \in \mathcal{O}$. The center of mass of this mass function gives a probability distribution, termed as *pignistic probability* (Smets & Kennes, 1994). The uncertainty associated with a particular output can then be expressed by either the entropy of the pignistic prediction or by the width of the credal prediction. The uncertainty associated with the whole generated sentence can then be computed as the maximum of the uncertainties among all predicted tokens.

The benefit of having a belief function as output, as a way of expressing epistemic uncertainty about the prediction, can be illustrated using an example. Suppose that a standard LLM was tasked to predict the next token in the sentence *Joe likes to play* ___ given the vocabulary $\mathcal{V} = \{baseball, basketball, \cdots\}$ and it predicted $\mathcal{P} : \{P(baseball) = 0.5, P(basketball) = 0.5\}$. So, there is a 50% chance of the next token being baseball and 50% of being basketball. However, this does not tell us if this predicted probability distribution is equally distributed because the LLM did not quite know the answer, or if both answers are valid. Allowing the model to predict a belief function, in opposition, gives the LLM more expressive power to clarify the source of the uncertainty.

Consider the two belief functions

$Bel_1 : \{bel(\{baseball\} = 0,\ bel(\{basketball\} = 0,\ bel(\{baseball, basketball\} = 1\}$ and

$Bel_2 : \{bel(\{baseball\} = 0.5,\ bel(\{basketball\} = 0.5,\ bel(\{baseball, basketball\} = 1\}.$

As it can easily be shown, both have the above predicted probability distribution $\mathcal{P}$ as their center of mass (pignistic distribution). So, the corresponding pignistic prediction would be the same. However, we can now associate an explanation with that answer. If the model predicts Bel_1, this means that the LLM does not know which one of baseball or basketball Joe likes, only that the answer is in the set $baseball, basketball$. Whereas, predicting Bel_2 means that Joe likes both baseball and basketball equally, so either can be the answer.

Fig. 1 encapsulates the training and generation process of RS-LLM. At training time, the RS-LLM is tasked to predict a belief function and is updated using the loss function highlighted in Sec 4.3. At generation time, the RS-LLM once again predicts a belief function, which is then transformed into a mass function. Its pignistic probability is then computed to obtain the final token. The training is done in parallel fashion using the teacher forcing method (Williams & Zipser, 1989) while generation needs to be done sequentially token by token. Our method is the first and only one that, as shown by our experiments, uniquely provides the LLMs with the capability to deal with second-level uncertainty in a principled manner using random-sets, so boosting its accuracy, while also providing tools to reason about hallucination by using Pignistic entropy and Credal Set Width (Sec. 4.4).

Contributions. This paper makes the following contributions to the wider research in LLMs:

1. We propose a unique Random-set Large Language Model (RS-LLM) approach based on predicting belief functions over the vocabulary instead of just a probability distribution. This leads to better uncertainty estimation and explainability of an LLM's generation.
2. We advance an intuitive approach for selecting the budget of the focal sets of hierarchical clustering over the embedding of the tokens, mitigating the computational complexity associated with working with random sets and making the approach scalable.
3. We set out a method to measure the uncertainty associated with an LLM's generation process using both the entropy of pignistic prediction and the width of the credal set Wang et al. (2024a); Caprio et al. (2024); Wang et al. (2025; 2024a) associated with the predicted belief function.
4. We show significant experimental results of our approach while also exploring it in the context of hallucination. RS-LLMs outperforms Standard LLMs and other methods in terms of closeness to the groundtruth, while also displaying remarkable capabilities in hallucination detection using credal set width.

Paper Outline. This paper is structured as follows. Sec. 1 introduces the need for and concept of Random-Set Large Language Model. Sec. 2 outlines the existing approaches and methods for uncertainty quantification in LLMs. In Sec. 3, we recall the notions of random sets, belief functions, credal and pignistic predictions. Sec. 4 details our RS-LLM approach, loss and uncertainty representation. Finally, Sec. 5 provides empirical evidence supporting out approach and discusses the results. Conclusions are given in Sec. 6.

2 Related Work

Uncertainty quantification in LLMs is a hot topic and indeed, significant work has recently been done in this regard. Plaut et al. (2024) proposes to simply use softmax entropy quantification, whereas (Kuhn et al., 2023; Farquhar et al., 2024) introduce a semantic entropy, measured over multiple runs of the model, as the uncertainty measure. Some methods use calibration for uncertainty evaluation (Glushkova et al., 2021; Wang et al., 2022) and use calibration errors to evaluate if the models can be trusted. One stream of methods Chen et al. (2024) uses the hidden state information within the models to quantify uncertainty. Other methods address uncertainty by prompting the model to evaluate its own generation Mielke et al. (2020) or by fine-tuning the model so that it predicts its own uncertainty (Lin et al., 2022; Kadavath et al., 2022). In addition, general-purpose methods like Monte-Carlo Dropout (MCD) Gal & Ghahramani (2016), Bayes by Backprop (BBB) Blundell et al. (2015) and Deep Ensembles Balabanov & Linander (2024) have also been applied to LLMs to quantify uncertainty. Laplace-LORA Yang et al. (2023) and BLoB Wang et al. (2024b) apply Bayesian methods to LoRA Hu et al. (2021). LoRA provides an efficient way of fine-tuning LLMs by injecting low-rank weight matrices into transfomer layers and training them instead of the whole network. Finally, ENN-LLM Osband et al. (2022) measures uncertainty by using their ensemble-inspired Epinet with LLMs.

Our proposed method does not directly compare with these uncertainty methods as they (i) mostly measure uncertainty by modifying the prompt or over multiple runs (Mielke et al., 2020; Osband et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024; Wang et al., 2024b; Farquhar et al., 2024) and (ii) because our method is a complete training regimen rather than just being a wrapper method (Mielke et al., 2020; Glushkova et al., 2021; Wang et al., 2022; Plaut et al., 2024). (Lin et al., 2022; Kadavath et al., 2022) and Chen et al. (2024) directly train their model to predict uncertainty or

use hidden state information to compute uncertainty, which is, once again, not directly comparable to our method. However, for the sake of completeness, we do provide a comparison against two representative methods.

3 Random sets and belief functions

Random sets. Consider a sensor measuring a 2D position vector, $\mathbf{x} = (x, y)$, where each component represents a spatial coordinate, modelled by a data probability distribution. Ideally, the sensor outputs precise values for each coordinate. However, suppose due to occlusion or sensor failure, the second coordinate y is unreadable.

Instead of assigning a single value, we can model the missing component y as the set Y containing all possible values for that component. The underlying random process that generates the data remains, but the resulting random variable assumes values which can be sets:

$$\mathbf{X} = (x, Y), \tag{1}$$

where Y is a set accounting for the missing information. The data generation process can thus be modelled as a set-valued random variable, i.e., a *random set*. Random sets generalise classical additive probability measures by representing uncertainty through sets, rather than precise values, making random sets a powerful tool for modeling imprecise observations (Matheron, 1975; Nguyen, 1978; Molchanov, 2005).

Belief functions. Random sets have been proposed by (Dempster, 2008) and (Shafer, 1976a) as a mathematical model for subjective belief, alternative to Bayesian reasoning. Thus, on finite domains (e.g., for classification) they assume the name of *belief functions*. While classical discrete mass functions assign normalised, non-negative values to *elements* $\theta \in \Theta$ of their sample space, a belief function independently assigns normalised, non-negative mass values to its *subsets* $A \subset \Theta$:

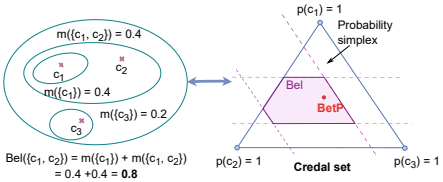


Figure 2: A belief function measures the total belief (sum of masses of its subsets) for a set (Eq. 3).

$$m(A) \geq 0, \ \forall A \subset \Theta, \quad \sum_{A} m(A) = 1. \tag{2}$$

The belief function associated with a mass function m measures the total mass of the subsets of each *focal set* A. Mass functions can be recovered from belief functions via Moebius inversion (Shafer, 1976a; Cuzzolin, 2010a), which, in combinatorics, plays a role similar to that of the derivative in calculus:

$$Bel(A) = \sum_{B \subseteq A} m(B), \quad m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B). \tag{3}$$

For example, consider a classification problem where an object belongs to one of three possible categories, $\Theta = \{c_1, c_2, c_3\}$. A belief function might express uncertainty by assigning mass as follows:

$$m(\{c_1\}) = 0.4, m(\{c_3\}) = 0.2, m(\{c_1, c_2\}) = 0.4. \tag{4}$$

Here, 40% of the belief supports c_1, 20% supports c_3, and 40% supports the composite hypothesis that the object belongs to either c_1 or c_2 but not c_3, without being able to specify which. The belief value $Bel(A)$ of a set of classes A accumulates mass from all the subsets of A:

$$Bel(\{c_1, c_2\}) = m(\{c_1\}) + m(\{c_1, c_2\}) = 0.4 + 0.4 = 0.8. \tag{5}$$

This means we have an 80% belief that the object belongs to either c_1 or c_2, reflecting epistemic uncertainty rather than simply assigning a probability to each class. Belief functions generalise Bayesian probability by allowing explicit representation of uncertainty, making them useful in applications where knowledge is incomplete or ambiguous Shafer (1976b); Bouckaert (1995).

Belief functions have been employed in the past for machine learning purposes, e.g. classification Cuzzolin (2018a) regression Gong & Cuzzolin (2017); Cuzzolin & Frezza (2000).

Credal prediction. As anticipated, RS-LLM is designed to predict a belief function (a finite random set) on the set of classes. A belief function, in turn, is associated with a convex set of probability distributions (a *credal set* (Levi, 1980; Zaffalon & Fagiuoli, 2003; Cuzzolin, 2010b; Antonucci & Cuzzolin, 2010)) on the same domain. This is the set:

$$Cre = \{P : \Theta \to [0, 1] | Bel(A) \leq P(A)\}, \tag{6}$$

of probability distributions P on Θ which dominate the belief function on each and every focal set A.

The size of the resulting credal prediction measures the extent of the related epistemic uncertainty (see Sec. 4.4). Indeed, the use of credal set size as a measure of epistemic uncertainty is well-supported in literature (Hüllermeier & Waegeman, 2021; Bronevich & Klir, 2008), as it aligns with established concepts of uncertainty such as conflict and non-specificity (Yager, 2008; Kolmogorov, 1965). A wider credal set indicates higher uncertainty, as the model refrains from committing to a specific probability distribution due to limited or conflicting evidence. In contrast, a narrower credal set implies lower uncertainty (Manchingal et al., 2025c; Caprio et al., 2025), signifying a more confident prediction based on substantial, consistent evidence.

Pignistic probability. The *pignistic probability* is the probability distribution obtained by re-distributing the masses of all the focal sets A of a belief function Bel to their constituent elements $\theta \in A$:

$$BetP(\theta) = \sum_{A \ni \theta} \frac{m(A)}{|A|}. \tag{7}$$

Smets (Smets, 2005b) originally proposed to use the pignistic probability for decision making using belief functions, by applying expected utility to it. Notably, the pignistic probability is geometrically the centre of mass of the credal set (6) associated with a belief function (Cuzzolin & Frezza, June 2001; Cuzzolin, 2008; 2018b; 2020).

4 Random-Set Large Language Models

4.1 Representation

RS-LLM predicts for each input sequence of tokens a belief function for the next token, rather than a vector of softmax probabilities as in a traditional LLM. A 'vanilla' RS-LLM, having a vocabulary of T tokens, would have 2^N outputs (as 2^N is the cardinality of $\mathbb{P}(T)$), each being the belief value of the focal set of tokens $A \in \mathbb{P}(T)$ corresponding to that output neuron. This would requires the output layer of the LLM to be amended to have 2^N outputs, as the only change needed to convert any LLM architecture to an RS-LLM architecture. In practice, for efficiency only a budget of focal sets is identified from the training data (Sec. 4.2).

Since the model is tasked to predict a belief function, the ground-truth for training needs to be modified. Given an input sequence, the ground-truth in a classical LLM is the one-hot encoding of the next token. The belief-encoded ground-truth for RS-LLM, instead, is given by the vector $\mathbf{bel} = \{Bel(A), A \in \mathbb{P}(T)\}$ of belief values for each focal set of tokens $A \in \mathbb{P}(T)$. $Bel(A)$ is set to 1 iff the true next token is contained in the subset A, 0 otherwise. This corresponds to full certainty that the element belongs to that set and there is complete confidence in this proposition. Fig. 1 shows the training and generation flow of RS-LLM (detailed in Sec. 1)

4.2 Budgeting

Random-sets are defined on the power set of all subsets of their domain. However, this is computationally impossible when the domain cardinality is extremely large for large language models where the typical size of vocabulary is ~32K tokens. The number of subsets in this case would be an astronomical 2^{32K}. Therefore, to tackle this problem, we need a method to select a budget of most appropriate focal sets and use only those as our focal sets.

We propose a novel strategy based on clustering the most similar tokens by utilize hierarchical clustering. This is appropriate because, by intuition, the model will most likely be confused among tokens which are similar to each other and potentially have similar semantics Farquhar et al. (2024). This is, in principle, different from the budgeting mechanism proposed in Manchingal et al. (2025d)

as this i) is model dependent instead of being data dependent, and ii) does not require multiple embedding instances of the same class/token.

As a first step, token embeddings are computed using a pre-trained LLM (in fact, any method for obtaining embedding of tokens can be used here). Then, hierarchical clustering is applied to cluster similar tokens. The number of clusters K is a hyper-parameter which determines the granularity of the belief function.

The final budget $\mathcal{O}$ of focal sets of the belief functions to be predicted is the union of the clusters of tokens so obtained and the collection of singleton sets containing the T original tokens, amounting to a total of $(K + T)$ sets. Fig. 3 shows a detailed overview of the proposed budgeting method. An analysis of the focal sets obtained by this budgeting strategy is presented in Sec. 5 and § A.

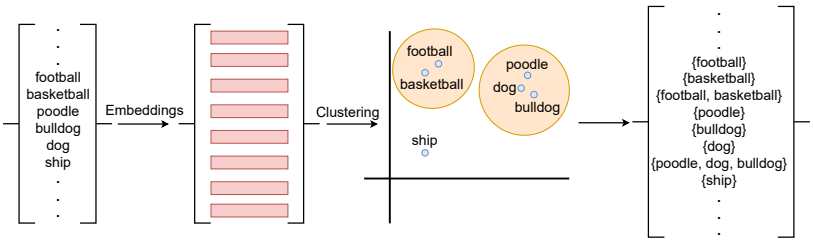


Figure 3: Proposed budgeting method for RS-LLM. First, embeddings are computed for all the tokens in vocabulary. Then, focal sets are computed using hierarchical clustering.

4.3 Loss function

As highlighted in Manchingal et al. (2025d), a random-set prediction problem is mathematically similar to the multi-label classification problem, for in both cases the corresponds to a vector of 0s and 1s and the prediction is vector in $[0, 1]$. They do have different semantics, as in the case of random sets, they indicate if the set contains true class or not; whereas in multi-label classification, it is the probability of input belonging to a particular class.

Despite the different semantics, we can thus adopt as loss Binary Cross-Entropy (BCE) (8) with sigmoid activation, to drive the prediction of a belief value for each focal set in the identified budget:

$$\mathcal{L}_{BCE} = -\frac{1}{b_{size}} \sum_{i=1}^{b_{size}} \frac{1}{seq_len} \sum_{j=1}^{seq_len} \frac{1}{|\mathcal{O}|} \sum_{A \in \mathcal{O}} \Big[Bel_{ij}(A) \log(\hat{Bel}_{ij}(A)) + (1 - Bel_{ij}(A)) \log(1 - \hat{Bel}_{ij}(A)) \Big]. \tag{8}$$

Here, i is the index of the training text and j is the index of the token in that text of length seq_len within a batch of cardinality b_{size}. A is a focal set of classes in the budget $\mathcal{O}$. $Bel_{ij}(A)$ is the A-th component of the vector $\mathbf{bel}_{ij}$ encoding the ground truth belief values for the j-th token of i-th training text, and $\hat{Bel}_{ij}(A)$ is the corresponding belief value in the predicted vector $\hat{\mathbf{bel}}_{ij}$ for the same training point. Both $\mathbf{bel}_{ij}$ and $\hat{\mathbf{bel}}_{ij}$ are vectors of cardinality $|\mathcal{O}|$ for all i and j.

To be valid, a belief function needs to be such that the corresponding mass values are non-negative and should sum up to 1 (Shafer, 1976a) (see Eq. (2)). To ensure that the predicted scores amount to a valid belief function, a mass regularisation term M_r and a mass-sum term M_s (9) are incorporated into the loss function. These terms encourage non-negativity and normalization (sum equal to 1) of the (predicted) mass values $\hat{m}(A)$, $A \in \mathcal{O}$, respectively:

$$M_s = \max\left(0, \frac{1}{b_{size}} \sum_{i=1}^{b_{size}} \sum_{A \in \mathcal{O}} \hat{m}_i(A) - 1\right), \quad M_r = \frac{1}{b_{size}} \sum_{i=1}^{b_{size}} \sum_{A \in \mathcal{O}} \max(0, -\hat{m}_i(A)). \tag{9}$$

These mass values are obtained from the predicted belief function $\hat{\mathbf{bel}}$ by rearranging the belief definition, i.e.,

$$m(A) = Bel(A) - \sum_{B \subset A} m(B). \tag{10}$$

The reason is that the Moebius inverse cannot be used here, as it requires the belief values for all subsets of a given set to compute the corresponding mass function. This distinguishes it from the loss

function used in (Manchingal et al., 2025d), as this makes it more generalizable and usable with all kinds of budget focal sets.

All loss components $\mathcal{L}_{BCE}$, M_r and M_s are computed during batch training. Two hyperparameters, α and β, control the relative importance of the two mass terms, yielding as the total loss for RS-LLM:

$$\mathcal{L}_{RS} = \mathcal{L}_{BCE} + \alpha M_r + \beta M_s. \tag{11}$$

To ensure adherence to valid belief functions, regularisation terms are introduced to discourage deviations, akin to how training-time regularisation in neurosymbolic learning fosters predictions that align with commonsense reasoning (Giunchiglia et al., 2023). Nevertheless, when α and β are too small, regularisation alone may not suffice to maintain the validity of belief-function predictions. In such instances, a corrective post-processing step is implemented: negative mass values are reset to zero, and an additional 'universal' set — comprising all classes — is incorporated into the final budget. This ensures that any residual mass is assigned to this set, preserving the requirement that the total mass across all focal sets in $\mathcal{O}$ sums to 1. This adjustment follows well-established approximation methods Cuzzolin (2009a;b; 2011), as detailed in (Cuzzolin, 2020), Part III or Cuzzolin (2013).

4.4 Uncertainty estimation

Pignistic prediction. The pignistic probability (Eq. 7) Smets (2005a) provides a representative estimate derived from a belief function, which can be understood as defining a credal set. Although it resembles a conventional probability distribution in behavior, the RS-LLM framework is explicitly designed to incorporate class set relationships during training, fostering a more structured form of learning. As a result, the pignistic probabilities produced by this model encode information from the underlying mass functions and subset structures, leading to predictions that, as shown in our experiments, offer greater reliability than those obtained through traditional softmax outputs.

Entropy of the pignistic prediction. The Shannon entropy of the pignistic prediction $BetP_j$ for a particular token at index j can then be used as a measure of the uncertainty associated with that token. This is in some way analogous to the entropy of a softmax probability prediction.

The cumulative uncertainty associated with the whole generated text is the mean uncertainty of that text.

$$H_{RS} = mean\left\{-\sum_{v\in V} BetP_j(v)\log BetP_j(v)\right\}. \tag{12}$$

A higher entropy value (12) thus indicates greater uncertainty in the model's predictions.

Size of the credal prediction. A meaningful way to quantify the epistemic uncertainty in a random-set prediction $\hat{\mathbf{bel}}$ is by assessing the size of its corresponding credal set. Since a credal set forms a convex polytope, various metrics can be used to evaluate its size (Sale et al., 2023). For a given token t, the probability range within the predicted credal set $\hat{Cre}$ is determined by its upper Cuzzolin (July 2003) and lower probability bounds:

$$\overline{P}(t) = \max_{P\in\hat{Cre}} P(t), \quad \underline{P}(t) = \min_{P\in\hat{Cre}} P(t), \tag{13}$$

The difference between these bounds, $\overline{P}(t) - \underline{P}(t)$, serves as a measure of the size of the credal set, particularly when considering the most probable next token as determined by the pignistic prediction. Crucially, for every token t, not just the most likely next token $\hat{t}$, the computed pignistic probability estimate $BetP(t)$ is guaranteed to fall within the range $[\underline{P}(t), \overline{P}(t)]$. The width of this interval reflects the degree of epistemic uncertainty in the prediction, with larger intervals indicating greater uncertainty.

5 Experiments

5.1 Implementation

Datasets. We evaluate the efficacy of our method on four datasets: one free-text dataset, CoQA (Reddy et al., 2019); two multiple-choice datasets, OBQA (Mihaylov et al., 2018) and ARC-E (Clark et al., 2018); and one Boolean dataset, BoolQ (Clark et al., 2019). Furhter details of these datasets are mentioned in § B.

Training Details. We make use of Llama2-7B-hf (Touvron et al., 2023), Llama-3.1-8B (Grattafiori et al., 2025), Mistral-7b (Jiang et al., 2023) and Phi-2 (Javaheripi et al., 2023) as our base models. As highlighted in Sec. 4.1, any LLM can be converted to a random-set LLM by simply redefining the last layer to represent sets of tokens instead of individual tokens. This last layer is then trained using the loss function detailed in Sec. 4.3.

By employing the budgeting technique highlighted in Sec. 4.2, we extract $K = 8,000$ focal sets from 32,000 tokens in Llama2 and Mistral, 51,200 tokens in Phi-2 and 128,256 tokens in Llama3.1, so that the output size of the last layer for RS-Llama2 and RS-Mistral is $32,000 + 8,000 = 40,000$, $51,200 + 8,000 = 59,200$ for Phi-2, and $128,256 + 8,000 = 136,256$ for Llama3.1. The models are trained using the Supervised Fine-tuning (SFT) method for LLMs. Further implementation details are provided in § B.

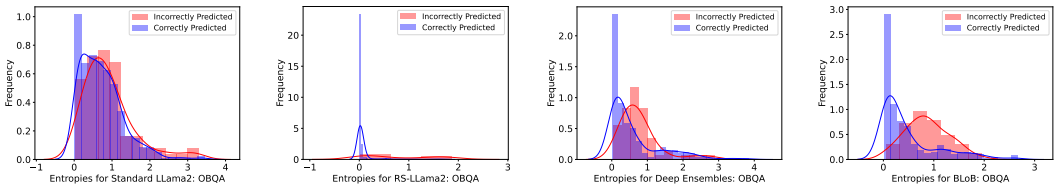


Figure 4: Behavior of entropy measures for all models with respect to the correctness on OBQA datasets

5.2 Results and Analysis

Performance comparison. We evaluate all model architecures on CoQA and OBQA datasets to higlight the performance improvments over standard LLMs. For CoQA, we produce free text. To evaluate the closeness of generated and ground truth text, we employ the cosine similarity metric between the two Banerjee et al. (2023). Cosine Similarity is a metric used to determine the cosine of the angle between two non-zero vectors in a multi-dimensional space. It is used in CoQA to measure semantic closeness between question and context (or candidate answers) text embeddings. Whereas, for OBQA, the option label is expected and generated which allows us to conveniently measure the accuracy w.r.t. the ground truth. Table 1 reports the cosine similarity and accuracy on the CoQA and OBQA datasets, respectively, for both models. RS-LLMs clearly outperform the standard LLMs model on both datasets across all models, even though all models are trained using the exact same regimen. This clearly shows the representative prowess of random sets in the output space. Some additional experiments are shown in § C. Sample generations are shown in § D.

Table 1: Performance of Standard LLMs and RS-LLMs on CoQA and OBQA datasets over 5 runs. For CoQA, cosine similarity is reported while accuracy is reported for OBQA.

Model	*CoQA* *(Cosine Similarity)*	OBQA *(Accuracy)*
Llama2	0.68 ± 0.005	83.20 ± 0.4
RS-Llama2	**0.71 ± 0.006**	**89.6 ± 0.3**
Mistral	0.67 ± 0.003	91.60 ± 0.4
RS-Mistral	**0.72 ± 0.004**	**93.00 ± 0.2**
Phi2	0.72 ± 0.005	87.60 ± 0.2
RS-Phi2	**0.73 ± 0.003**	**91.80 ± 0.2**
Llama3.1	0.73 ± 0.002	91.80 ± 0.1
RS-Llama3.1	**0.74 ± 0.003**	**93.20 ± 0.2**

Comparison with other methods. We also compare our approach with other uncertainty methods; namely BLoB (Wang et al., 2024b), a Bayesian-based method, and LoRA Ensembles (Balabanov & Linander, 2024), an ensemble based method on the ARC-E Clark et al. (2018), BoolQ Clark et al. (2019) and OBQA Mihaylov et al. (2018) datasets. All methods are evaluated using Llama-2 models. Table 2 shows the quantitative results. RS-Llama2 significantly outperforms all the other methods on all datasets in terms of accuracy, clearly showcasing its superiority.

Table 2: Performance of Standard LLM, RS-LLM, Bayesian and ensemble LLM on ARC-E, BoolQ and OBQA datasets.

Model	ARC-E	BoolQ	OBQA
Llama2	83.89	87.51	83.20
BLoB	84.78	87.23	85.12
LoRA Ensembles	84.31	87.09	81.38
RS-Llama2	**86.12**	**88.56**	**89.60**

Uncertainty quantification. In Fig. 4, instead, we show how the uncertainty measures produced by the four models behave with respect to the correctness of the answer. It illustrates the entropy

distributions of correct vs. incorrect predictions for all models on the OBQA dataset. Ideally, these distributions should be clearly separable. Both BLoB and RS-Llama2 perform well in this regard, with RS-Llama2 further distinguishing itself by exhibiting a sharp spike at zero in the correct prediction distribution and showing better separation.

Budgeting Analysis. Furthermore, we analyse the obtained focal sets using our budgeting method. In general, they show appropriate closeness. Some examples include '[cattle, sheep]', '[shame, pity]', '[maintain, retain, maintained, retained]' and '[delight, pleasure, pleased, proud, pride]'. Moreover, we also analyse the semantics of the sets obtained. Primarily, we try to see whether the sets only contain tokens with same root-word or different words with similar meanings. Upon quantitative analysis, we find that out of 8,000 sets, 4,508 have 2 or less length of common substring, whereas only 721 have 6 or more length of common substring. In terms of synonyms, there are 2,703 sets which have one or more synonyms in them. Therefore, we can conclude that the budget contains sets with both possibilities. This is good because this makes our budgeted sets more general. This highlights the efficacy of our proposed budgeting method. Further analysis is provided in § A.

5.3 Hallucination

Lastly, we experiment to evaluate the models on their hallucination. For this, we deliberately provide the LLM with incorrect context and then measure how confidently the model answers the given question. More specifically, for BoolQ we randomly replace the context of given test sample with another context from the test set and for OBQA we replace the additional information along with the answer choices with some other random test point. Tab. 3 presents the uncertainty evaluation and AUROC of all uncertainty methods under correct and incorrect context. As evident, RS-Llama2 exhibits significantly superior performance on BoolQ while also showing comparable performance on the OBQA dataset. Further analysis is provided in § C.

Table 3: Uncertainty evaluation using pignisitic entropy on correct and incorrect context.

Model	BoolQ			OBQA		
	CC (↓)	IC (↑)	AUROC	CC (↓)	IC (↑)	AUROC
LLama2	0.50 ± 0.52	0.97 ± 0.64	78.43	0.75 ± 0.59	1.10 ± 0.93	66.45
BLoB	0.33 ± 0.26	0.45 ± 0.30	56.34	0.48 ± 0.53	1.46 ± 0.47	**92.45**
LoRA Ensembles	0.42 ± 0.36	0.60 ± 0.44	61.62	0.62 ± 0.72	1.12 ± 0.91	63.21
RS-LLama2	0.13 ± 0.22	0.62 ± 0.36	**86.34**	0.14 ± 0.38	0.61 ± 0.68	90.23

6 Conclusions

This paper proposes a novel training and fine-tuning regimen for Large Language Models (LLMs) that not only provides a higher-quality representation, leading to significantly improved performance, but also allows the model to encode second-order uncertainty on its own predictions, in the random set formalism. Through our experiments, we conclusively demonstrate the efficacy of our proposed approach. The results demonstrate the potential of our approach to change the fundamental nature of LLMs, and move towards language models more aware of their own uncertainty and thus, potentially, less prone to hallucination.

A current limitation of our work is having to manually set the number of focal elements K; a dynamic strategy adjusting K based on overlap would enhance flexibility and effectiveness and will be subject to future work.

Future work includes the exploration of other clustering algorithms (e.g., fuzzy clustering) for more effective budgeting, and the design of a large-scale mathematical framework for training an LLM from scratch in the Random-Set approach, as our next objective. Another interesting line of investigation could be exploring the semantic relation of tokens in a focal set, to potentially propose random-sets as alternative to individual tokens altogether.

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work. Large Language Models (LLMs) have the potential to revolutionize fields like healthcare, education, and business by improving accessibility, enhancing creativity, and supporting decision-making at scale. However, their widespread use raises concerns about bias, job displacement, and privacy risks.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Marcel R Ackermann, Johannes Blömer, Daniel Kuntze, and Christian Sohler. Analysis of agglomerative clustering. *Algorithmica*, 69:184–215, 2014.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Alessandro Antonucci and Fabio Cuzzolin. Credal sets approximation by lower probabilities: application to credal networks. In *Computational Intelligence for Knowledge-Based Systems Design: 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010, Dortmund, Germany, June 28-July 2, 2010. Proceedings 13*, pp. 716–725. Springer, 2010.

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

Oleksandr Balabanov and Hampus Linander. Uncertainty quantification in fine-tuned llms using lora ensembles. *arXiv preprint arXiv:2402.12264*, 2024.

Debarag Banerjee, Pooja Singh, Arjun Avadhanam, and Saksham Srivastava. Benchmarking llm powered chatbots: methods and metrics. *arXiv preprint arXiv:2308.04624*, 2023.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Remco Ronaldus Bouckaert. *Bayesian belief networks: from construction to inference*. PhD thesis, 1995.

Andrey Bronevich and George J Klir. Axioms for uncertainty measures on belief functions and credal sets. In *NAFIPS 2008-2008 Annual Meeting of the North American Fuzzy Information Processing Society*, pp. 1–6. IEEE, 2008.

Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal learning theory. *arXiv preprint arXiv:2402.00957*, 2024.

Michele Caprio, Shireen K Manchingal, and Fabio Cuzzolin. Credal and interval deep evidential classifications. *arXiv preprint arXiv:2512.05526*, 2025.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. Wikisum: Coherent summarization dataset for efficient human-evaluation. 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)*, pp. 212–219, 2021.

Fabio Cuzzolin. On the credal structure of consistent probabilities. In *European Workshop on Logics in Artificial Intelligence*, pp. 126–139. Springer, 2008.

Fabio Cuzzolin. Complexes of outer consonant approximations. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 10th European Conference, ECSQARU 2009, Verona, Italy, July 1-3, 2009. Proceedings 10*, pp. 275–286. Springer, 2009a.

Fabio Cuzzolin. Credal semantics of bayesian transformations in terms of probability intervals. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(2):421–432, 2009b.

Fabio Cuzzolin. Three alternative combinatorial formulations of the theory of evidence. *Intelligent Data Analysis*, 14(4):439–464, 2010a.

Fabio Cuzzolin. Credal semantics of Bayesian transformations in terms of probability intervals. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(2):421–432, 2010b.

Fabio Cuzzolin. Three alternative combinatorial formulations of the theory of evidence. *Intelligent Data Analysis*, 14(4):439–464, 2010c.

Fabio Cuzzolin. On consistent approximations of belief functions in the mass space. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pp. 287–298. Springer, 2011.

Fabio Cuzzolin. Lp consonant approximations of belief functions. *IEEE Transactions on Fuzzy Systems*, 22(2):420–436, 2013.

Fabio Cuzzolin. *Belief functions: theory and applications*. Springer, 2014.

Fabio Cuzzolin. Generalised max entropy classifiers. In *Belief Functions: Theory and Applications: 5th International Conference, BELIEF 2018, Compiègne, France, September 17-21, 2018, Proceedings 5*, pp. 39–47. Springer, 2018a.

Fabio Cuzzolin. Visions of a generalized probability theory. *arXiv preprint arXiv:1810.10341*, 2018b.

Fabio Cuzzolin. *The Geometry of Uncertainty: The Geometry of Imprecise Probabilities*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, 2020. ISBN 9783030631536. URL https://books.google.co.uk/books?id=jNQPEAAAQBAJ.

Fabio Cuzzolin. Reasoning with random sets: An agenda for the future. *arXiv preprint arXiv:2401.09435*, 2023.

Fabio Cuzzolin. Uncertainty measures: A critical survey. *Information Fusion*, pp. 102609, 2024.

Fabio Cuzzolin. Geometry of upper probabilities. In *Proceedings of the 3rd Internation Symposium on Imprecise Probabilities and Their Applications (ISIPTA'03)*, July 2003.

Fabio Cuzzolin and Ruggero Frezza. Integrating feature spaces for object tracking. *Proc. of MTNS2000 (21Y25 June 2000)*, 2000.

Fabio Cuzzolin and Ruggero Frezza. Geometric analysis of belief space and conditional subspaces. In *Proceedings of the 2nd International Symposium on Imprecise Probabilities and their Applications (ISIPTA'01)*, June 2001.

Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pp. 57–72. Springer, 2008.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Eleonora Giunchiglia, Mihaela Cătălina Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. Road-r: The autonomous driving dataset with logical requirements. *Machine Learning*, 112(9):3261–3291, 2023.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. Uncertainty-aware machine translation evaluation. *arXiv preprint arXiv:2109.06352*, 2021.

Wenjuan Gong and Fabio Cuzzolin. A belief-theoretical approach to example-based pose estimation. *IEEE Transactions on Fuzzy Systems*, 26(2):598–611, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. 2025.

Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. trlx: A framework for large scale reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8578–8595, 2023.

https://sbert.net/. SentenceTransformers Documentation &x2014; Sentence Transformers documentation — sbert.net. `https://sbert.net/`. [Accessed 23-07-2025].

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

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

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Andrei N Kolmogorov. Three approaches to the quantitative definition of information. *Problems of information transmission*, 1(1):1–7, 1965.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. The MIT Press, Cambridge, Massachusetts, 1980.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

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

Shireen Kudukkil Manchingal and Fabio Cuzzolin. Epistemic deep learning. *arXiv preprint arXiv:2206.07609*, 2022.

Shireen Kudukkil Manchingal, Armand Amaritei, Mihir Gohad, Maryam Sultana, Julian FP Kooij, Fabio Cuzzolin, and Andrew Bradley. Uncertainty-aware autonomous vehicles: Predicting the road ahead. *arXiv preprint arXiv:2510.22680*, 2025a.

Shireen Kudukkil Manchingal, Andrew Bradley, Julian FP Kooij, Keivan Shariatmadar, Neil Yorke-Smith, and Fabio Cuzzolin. Epistemic artificial intelligence is essential for machine learning models to trulyknow when they do not know'. *arXiv preprint arXiv:2505.04950*, 2025b.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, and Fabio Cuzzolin. A unified evaluation framework for epistemic predictions. In *International Conference on Artificial Intelligence and Statistics*, 2025c. URL https://api.semanticscholar.org/CorpusID:275932472.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025d. URL https://openreview.net/forum?id=pdjkikvCch.

Georges Matheron. *Random sets and integral geometry*. Wiley Series in Probability and Mathematical Statistics, New York, 1975.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Sabrina J Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *arXiv preprint arXiv:2012.14983*, 11, 2020.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Ilya Molchanov. Random sets and random functions. *Theory of Random Sets*, pp. 451–552, 2017.

Ilya S Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Hung T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.

Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-tuning language models via epistemic neural networks. *arXiv preprint arXiv:2211.01568*, 2022.

Benjamin Plaut, Khanh Nguyen, and Tu Trinh. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *arXiv preprint arXiv:2402.13213*, 2024.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. doi: 10.1162/tacl_a_00266. URL https://aclanthology.org/Q19-1016.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. Is the volume of a credal set a good measure for epistemic uncertainty? In *Uncertainty in Artificial Intelligence*, pp. 1795–1804. PMLR, 2023.

Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976a.

Glenn Shafer. A theory of statistical evidence. In W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume 2, pp. 365–436. Reidel, Dordrecht, 1976b.

Philippe Smets. Decision making in the tbm: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005a. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2004.05.003. URL https://www.sciencedirect.com/science/article/pii/S0888613X04000593.

Philippe Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005b.

Philippe Smets and Robert Kennes. The transferable belief model. *Artificial intelligence*, 66(2): 191–234, 1994.

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

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, Hans Hallez, et al. Credal deep ensembles for uncertainty quantification. *Advances in Neural Information Processing Systems*, 37: 79540–79572, 2024a.

Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural Networks*, pp. 107198, 2025.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *arXiv preprint arXiv:2406.11675*, 2024b.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.

Ronald R Yager. Entropy and specificity in a mathematical theory of evidence. *Classic works of the Dempster-Shafer theory of belief functions*, pp. 291–310, 2008.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*, 2023.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Marco Zaffalon and Enrico Fagiuoli. Tree-based credal networks for classification. *Reliable computing*, 9(6):487–509, 2003.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.

A Analysis of Budgeted Focal Sets Using the Proposed Budgeting Technique

In this section, we present a comprehensive analysis of the budgeted focal sets generated using our novel budgeting approach. To conduct this evaluation, we employ the Llama2-7b-hf model, which supports a substantial token context size of 32,000. For the budgeting parameter, we set $K = 8000$, meaning that from the enormous theoretical space of 2^{32000} possible focal sets, we constrain and select a manageable subset of 8,000 focal sets to analyze.

A.1 Qualitative Assessment of Focal Sets

The resulting focal sets exhibit strong semantic coherence, demonstrating that the budgeting technique effectively clusters related tokens. For instance, several focal sets group together semantically or contextually related words, such as:

- {`cattle`, `sheep`} — grouping related livestock terms.
- {`shame`, `pity`} — expressing related emotional states.
- {`quiet`, `calm`, `quietly`} — capturing synonyms and related adverbs describing tranquility.
- {`maintain`, `retain`, `maintained`, `retained`} — grouping verb forms and their conjugations.
- {`delight`, `pleasure`, `pleased`, `proud`, `pride`} — combining expressions of positive emotions and states.

These examples highlight the model's ability to capture both lexical similarity and subtle semantic relationships within the focal sets.

A.2 Quantitative Semantic Analysis: Centroid Distance

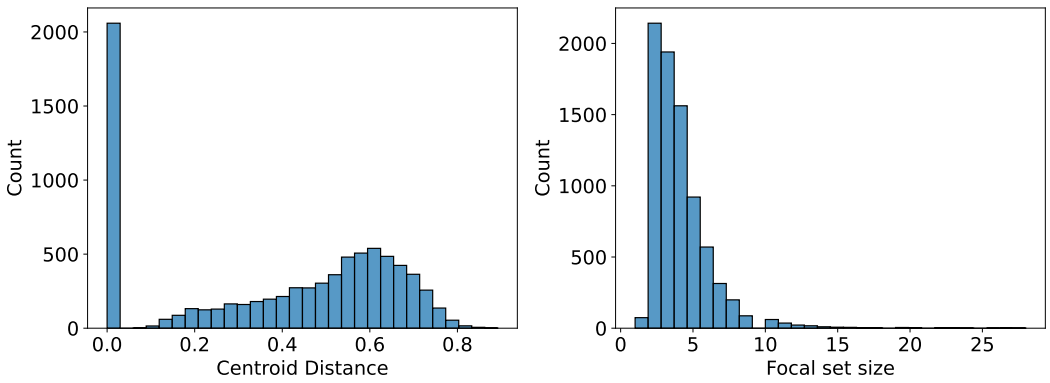
To quantify the semantic coherence of these focal sets, we perform a detailed semantic analysis by calculating the centroid distance for each set. Specifically, we use sentence-transformers https://sbert.net/ to obtain vector embeddings for each token within a focal set. The centroid distance is then computed as the mean Euclidean distance between each token's embedding and the centroid (mean vector) of the entire set.

Fig. 5a illustrates the distribution of these centroid distances across all 8,000 budgeted focal sets. The majority of the focal sets exhibit very small centroid distances, clustering tightly around zero, indicating high semantic similarity among the tokens in each set. Only a small fraction of sets show larger distances, which may correspond to more diverse or loosely connected token groups. This empirical evidence strongly supports the effectiveness of our budgeting method in generating semantically meaningful and coherent focal sets.

A.3 Distribution of Focal Set Sizes

Further insights are gained by examining the distribution of focal set sizes, as depicted in Fig. 5b. The data shows that most focal sets are relatively small, with sizes of 2, 3, and 4 being the most prevalent. Specifically, we observe 2,142 sets of size 2, 1,940 sets of size 3, and 1,562 sets of size 4, making these the dominant group sizes.

The largest focal set identified contains 162 tokens. Closer inspection reveals that this set predominantly consists of punctuation marks and symbols such as periods (.), exclamation marks (!), hashes (#), dollar signs ($), and percent signs (%). The formation of such a large set comprising non-lexical tokens is intuitive, as these symbols often co-occur or share syntactic roles, making them a natural cluster in the budgeting process.



(a) (b)

Figure 5: (a) Frequency distribution of centroid distances of the obtained budgeted focal sets. (b) Frequency distribution of sizes of obtained budgeted focal sets. Note: There are 8 focal sets with sizes > 30. They are excluded here for better visualisation.

A.4 Morphology and synonymity analysis

Furthermore, we conduct a detailed semantic analysis of the obtained focal sets to understand the nature of the token groupings. Specifically, we investigate whether the sets primarily consist of tokens sharing the same root word—indicating morphological similarity—or if they instead contain different words that are semantically related, such as synonyms.

To quantitatively evaluate morphological similarity, we measure the length of the longest common substring among tokens within each set. Our analysis reveals that out of the 8,000 budgeted focal sets, 4,508 sets have a longest common substring length of two characters or fewer, suggesting that these sets mostly consist of tokens with little or no shared root structure. In contrast, only 721 sets have a longest common substring length of six characters or more, indicating a smaller subset of sets with strong morphological ties.

In addition to morphological analysis, we assess synonymy within the focal sets. Using a lexical resource to identify synonyms, we find that 2,703 sets contain at least one pair of synonymous tokens. This indicates that a substantial portion of the focal sets cluster semantically similar but lexically distinct words.

Taken together, these findings demonstrate that our budgeting technique produces a diverse collection of focal sets: some clusters are formed by morphologically related variants of the same root word, while others group distinct but semantically similar words. This diversity enhances the generality and flexibility of our budgeted sets, allowing them to capture a broader range of linguistic relationships rather than being restricted to one type of semantic similarity.

A.5 Impact of Distance Metrics on Performance

Finally, we investigate the performance of various distance metrics in hierarchical clustering. Specifically, we evaluate the effectiveness of different distance metrics for RS-Llama2 embeddings on the CoQA dataset. Tab. 4 presents the cosine similarities obtained using several distance measures for hierarchical clustering. Our analysis reveals that all the metrics exhibit comparable performance, with only minor variations observed across the different methods. Notably, Euclidean distance and Cosine similarity perform slightly better than Manhattan distance, likely due to their higher expressiveness in capturing the geometric relationships within the embedding space.

In addition to evaluating individual metrics, we also explore the performance when combining the distance metrics into a richer, multi-metric budget. The results show that it remains competitive, performing similarly to when as compared to stand-alone metric budgets. However, a key advantage of using a combined budget approach is that it yields better uncertainty estimates. This enhanced uncertainty estimation can allow for more accurate detection of hallucinations (see Tab. 2).

Table 4: Cosine similarity values for different distance metrics in hierarchical clustering.

Metric	Euclidean Distance	Manhattan Distance	Cosine Similarity	Combined Budget (16,354)
Cosine Similarity	**0.71**	0.70	**0.71**	0.70

Table 5: Uncertainty evaluation using pignistic entropy of budgeting with a stand-alone metric and with combined metrics.

	CoQA		
	CC (↓)	IC (↑)	AUROC
Stand-alone Metric (Euclidean Distance)	2.39 ± 1.42	4.70 ± 3.25	72.14
Combined Metrics	2.62 ± 1.74	6.34 ± 3.87	**80.09**

B Implementation Details & Ablation Studies

B.1 Datasets

We evaluate the efficacy of our method on four diverse datasets, each with distinct characteristics and challenges: one free-text dataset, two multiple-choice datasets, and one Boolean dataset. Specifically, we evaluate on:

- **CoQA** (Reddy et al., 2019), a free-text, conversational question-answering (QA) dataset where the task is to generate answers based on a provided story. The goal is to generate free text and provide relevant, context-aware answers. CoQA consists of 7,200 training samples and 500 validation samples. For our evaluation, we use the validation set as the test set.
- **OBQA** (Mihaylov et al., 2018), a multiple-choice QA dataset focused on STEM (Science, Technology, Engineering, Mathematics) subjects. The task involves selecting the correct answer from a set of options, with each question accompanied by a fact (additional helpful information) that aids in answering. In this paper, we use the OBQA-Additional variant (referred to as OBQA) of the dataset. OBQA contains 4,957 training samples and 500 test samples.
- **ARC-E** (Clark et al., 2018), a multiple-choice QA dataset designed for evaluating models on their ability to answer science questions. It includes questions from elementary-level science topics. ARC-E consists of 2,251 training questions and 2,376 test questions, designed to challenge models with reasoning skills beyond simple fact retrieval.
- **BoolQ** (Clark et al., 2019), a Boolean question-answering dataset where the task is to predict a true or false answer based on the given context. It includes 9,427 training and 3,270 testing samples across a wide range of topics, with a binary answer format (True/False), making it an ideal dataset for evaluating models in terms of their ability to reason over short passages and provide accurate answers.

B.2 Implementation Details

All models are trained on NVIDIA A100 80GB GPUs using Huggingface's trl framework (Havrilla et al., 2023) with default training parameters for 5 epochs, with a batch size of 8. To further boost the training, we add LoRA adapters (Yu et al., 2023) of rank 64 to all blocks of the model. Furthermore, we load and train the model in 4-bit mode for enhanced efficiency. Similary setting is used for BLoB and LoRA Ensembles with number of samples and number of ensemble set to 2 for them respectively. For RS-Llama2, we set $\alpha = \beta = 1e-2$ as hyperparameter values in the loss function.

Fig. 6 highlights the training prompt template. The blue text represents the instructions to the LLM while black represents the actual question. The model is trained to predict the text in green. At generation time, the model is given input in the same template but without the answer. The model then continues the statement and produces the answer.

CoQA

Story: The Vatican Apostolic Library (), more commonly called the Vatican Library or simply the Vat, is the library of the Holy See, located in Vatican City. Formally (.....) though some are very significant.

Answer the following question based on the above story.

When was the Vat formally opened?
Answer: It was formally established in 1475

OBQA

Fact:
the sun is the source of energy for physical cycles on Earth

Answer the following question based on the above fact by selecting the correct option.

The sun is responsible for
A) puppies learning new tricks
B) children growing up and getting old
C) flowers wilting in a vase
D) plants sprouting, blooming and wilting

Answer: D

Figure 6: Training examples from CoQA and OBQA datasets. The text in black highlights the actual question, while the blue text represents prompt instructions. The model is trained to predict the text in green.

B.3 Ablation on Hyperparameters α & β

The terms M_s and M_r, as defined in Equations 9 and 10 respectively, are included in the loss function (Equation 12) to enforce the validity of the belief functions generated by the model. Specifically, these terms encourage two key properties: the sum of the masses must be equal to 1, and all masses must remain non-negative. These constraints ensure that the model outputs consistent and valid belief functions. However, it is crucial to strike an appropriate balance in determining the weight of these terms within the overall loss function. Excessively penalizing deviations from these constraints can lead to undesirable consequences, as it may impair the model's ability to make accurate predictions.

This issue is similar to the problem faced in Variational Autoencoders (VAEs), where an overemphasis on the KL divergence term in the loss function can cause the model to focus too heavily on matching the latent distribution, potentially at the expense of reconstructing the input data faithfully. As a result, while the latent space may be well-regularized, the quality of the data reconstruction may suffer, leading to subpar performance.

To investigate the optimal balance between constraint enforcement and predictive performance, we conducted experiments with different values for the hyperparameters α and β. These hyperparameters control the relative importance of the mass-related terms in the loss function. Tab. 6 presents the cosine similarities for RS-Llama2 with varying values of α and β. Our findings indicate that values of $\alpha = \beta = 0.01$ and $\alpha = \beta = 0.0001$ yield similar performance in terms of cosine similarity. Whereas, setting them to zero leads to poor performance, concluding that belief function training objective does provide performance gains over simple clustered hierarchical output. However, to further encourage the generation of valid belief functions, we opt for $\alpha = \beta = 0.01$ in our experiments, as this choice strikes a better balance between model accuracy and constraint satisfaction.

Table 6: Cosine Similarity values for different $\alpha = \beta$ settings

$\alpha = \beta$	0.1	0.01	0.001	0.0001	0
Cosine Similarity	0.66	**0.71**	0.69	**0.71**	0.68

B.4 Ablation on Number of Focal Sets

The number of non-singleton focal sets, denoted as K, to be budgeted is a critical hyperparameter that can influence model performance. It is important to explore the impact of different values of K, as it can affect the balance between model expressiveness and computational complexity. Specifically, a lower value of K may lead to results that closely resemble those of traditional large language models (LLMs), as it restricts the number of focal sets the model generates, thereby simplifying its task. On the other hand, a higher value of K introduces greater complexity, as the model is tasked with managing a larger number of focal sets, potentially increasing the diversity and richness of its predictions, but at the cost of computational efficiency and potential overfitting.

To better understand the effect of K on model performance, we conducted an ablation study on the CoQA dataset, as shown in Tab. 7. In this study, we evaluated the model's performance across various values of K, ranging from smaller budgets (2,000 focal sets) to larger ones (16,000 focal sets).

We also included a "Combined" budget, which represents the union of the budgets for 2,000, 4,000, 8,000, and 16,000 focal sets. This approach allows us to observe how combining different budgets impacts the model's ability to generalize and maintain performance.

Our findings indicate that using a medium value of K—neither too small nor too large—yields the most optimal results. This balance ensures that the model has enough focal sets to generate meaningful predictions while avoiding excessive complexity that could hinder its ability to generalize effectively. These experiments were conducted using Llama2 with hyperparameters $\alpha = \beta = 0.01$, and the results suggest that tuning K is essential for achieving the best performance.

Table 7: Cosine Similarity values across different budget sizes

Budget Size	2000	4000	8000	16000	Combined (24,538)
Cosine Similarity	0.67	0.68	**0.71**	0.66	0.69

C Additional Experiments

C.1 Efficiency Analysis

The RS-LLM architecture introduces only a minimal set of additional matrix multiplication operations during both the training and inference phases. Consequently, the resulting computational overhead remains negligible when compared to the baseline. To substantiate this claim, we present a quantitative analysis of the temporal performance for both the Standard LLM and RS-LLM variants in Table 8.

For robustness, all reported timing metrics are averaged over 5 independent runs. The experimental setup utilized a batch size of 1 and a budget size of 8000, conducted on the Mistral-7b backbone. Note that the reported training duration corresponds to a single training step. As the data demonstrates, RS-LLM maintains near-parity with the standard model, confirming that the proposed modifications do not impose a significant runtime penalty.

Table 8: Quantitative Analysis of Training and Inference Times (Mistral-7b)

Model	Training (sec)	Inference (sec)
Standard LLM	0.761 ± 0.032	0.186 ± 0.002
RS-LLM	0.771 ± 0.076	0.196 ± 0.002

C.2 Hallucination Detection

We further evaluate the hallucination detection on CoQA dataset using entropy measures. To assess this, we deliberately introduce errors into the context or choices provided to the model, and then measure how confidently the model answers the given question. We simulate hallucination by randomly replacing the question in a given test sample with another question from the test set. This forces the model to respond to a question based on an unrelated context, testing its robustness in recognizing such discrepancies.

Tab. 9 presents the uncertainty evaluation and Area Under the Receiver Operating Characteristic Curve (AUROC) for Llama2 and RS-Llama2 under both correct and incorrect context conditions. Note that the results for BLoB and LoRA Ensembles are not available for CoQA, as their implementations do not currently support free-text generation tasks, limiting their applicability to this particular dataset.

As shown in the results, RS-Llama2 demonstrates superior performance in detecting hallucinations on the CoQA dataset, outperforming Llama2 and further validating its effectiveness in handling hallucinations across different types of tasks. This reinforces the model's robustness in uncertainty estimation, making it a reliable choice for real-world applications where context errors may occur.

C.3 Hallucination Detection with Credal set width

We also evaluate the hallucination detection for RS-Llama2 using credal set wdith. The experiment setting is same as for previously. Only the metric for uncertainty measurement is changed.

Table 9: Uncertainty evaluation using entropy for Llama2 and RS-Llama2 for CoQA on correct and incorrect context.

Model	CoQA		
	CC (↓)	IC (↑)	AUROC
LLama2	0.39 ± 0.45	0.90 ± 0.80	70.56
RS-LLama2	2.39 ± 1.42	4.70 ± 3.25	**72.14**

Tab. 10 presents the uncertainty evaluation and Area Under the Receiver Operating Characteristic Curve (AUROC) for RS-Llama uncertainty methods under both correct and incorrect context conditions. As shown in the results, RS-Llama2 effectively detects hallucinations under credal set width too.

Table 10: Uncertainty evaluation using credal set width for RS-Llama2 on correct and incorrect context.

Dataset	CC (↓)	IC (↑)	AUROC
CoQA	0.13 ± 0.13	0.28 ± 0.21	69.24
OBQA	0.01 ± 0.05	0.10 ± 0.09	81.90
BoolQ	0.00 ± 0.01	0.01 ± 0.02	67.23

C.4 Dataset Shift

To rigorously assess the generalization capabilities of our proposed architecture, we extended our evaluation to out-of-distribution datasets. While both the Standard Llama and RS-Llama2 models were trained on the OpenBookQA (OBQA) dataset, we tested their inference performance on the ARC-Easy (ARC-E), Chemistry (Chem), and Physics (Phy) benchmarks.

This experimental setup allows us to quantify the impact of dataset shift and determine how well the models adapt to unseen domains. As shown in Table 11, RS-Llama2 demonstrates superior generalization on the ARC-E and Physics tasks, outperforming the standard baseline, while maintaining competitive performance on the Chemistry dataset.

Table 11: Evaluation of Dataset Shift (Trained on OBQA)

Model	ARC-E	Chem	Phy
Standard Llama	73.71	**41.29**	26.63
RS-Llama	**75.52**	40.95	**28.51**

C.5 Align Score

In addition to using cosine similarity for the evaluation on the CoQA dataset, we also introduce another evaluation metric known as Align Score Zha et al. (2023). The Align Score is a measure of factual consistency between the generated and predicted text. While traditionally used for measuring alignment in machine translation tasks, we adapt it for the Question Answering (QA) setting. Specifically, in our case, we compute the Align Score between the ground truth and the generated answer, providing a more meaningful and relevant assessment for QA models. This adaptation allows us to assess how consistent the model's generated answers are with the true answers, focusing on factual accuracy rather than just semantic similarity.

Tab. 12 presents the Align Score between the ground truth and the generated answers for different models on the CoQA dataset. From the results, it is evident that RS-LLMs (Random-Set LLMs) consistently outperform standard LLMs across all tested models. This demonstrates the superiority of RS-LLMs in generating more factually consistent and reliable answers, highlighting their effectiveness in tasks where factual accuracy is critical.

Table 12: AlignScore on CoQA Dataset for Llama2 and RS-Llama2

Model	Llama2	RS-Llama2
Llama2	0.44	**0.57**
Mistral	0.46	**0.52**
Phi2	0.51	**0.56**

C.6 Additional Dataset

To further evaluate the efficacy of our proposed approach, we extend our analysis to a summarization task. Specifically, we use the XSum dataset (Narayan et al., 2018), which requires the model to generate a concise, one-sentence summary for a given document and WikiSum (Cohen et al., 2021), a long summary dataset. This task serves as an excellent benchmark for testing how well a model can condense information into a succinct and accurate form while preserving the essential content of the original document.

Tab. 13 presents the performance of Mistral and RS-Mistral on the XSum and WikiSUm dataset. The results clearly demonstrate that RS-Mistral outperforms Standard Mistral in terms of cosine similarity between the generated summaries and the ground truth. This improvement highlights the effectiveness of our approach in capturing and preserving the semantic content of the document while generating a more precise and accurate summary. By leveraging the RS-Llama framework, RS-Mistral is able to produce summaries that are not only more aligned with the ground truth but also more representative of the key concepts in the source text, offering a substantial improvement over the baseline model.

These findings provide further validation for the proposed method, demonstrating its applicability and effectiveness across different types of natural language understanding and generation tasks, including summarization.

Table 13: Performance (cosine similarity) of Standard Mistral and RS-Mistral on XSum and WikiSum datasets.

	Mistral	RS-Mistral
XSum	0.56	**0.60**
WikiSum	0.42	**0.44**

D Generated Results

In this section, we present the generations by all models. Tab. 14 displays the their generations on the CoQA test set while Tab. 15 displays the generated results on the OBQA test set. For better readability, the 'story' in Tab. 14 has been shortened.

Table 14: Generated results by RS-Llama2 and Llama2 on CoQA test set.

Story	Question	RS-Llama2	Llama2	Groundtruth
The Six-Day War (Hebrew: , ”Milhemet Sheshet Ha Yamim”; Arabic: , ”an-Naksah”, ” . . . ing to the conflict.	When was the Six-Day War fought?	1967.	1967.	June 5 and 10, 1967
(CNN) – President Barack Obama has nominated his top counterterrorism adviser, . . . the past year or so.	Who did Obama nominate?	John Brennan, to be the next director of the CIA.	John Brennan.	John Brennan.
(CNN) – There are plenty of country songs about how to be a man. One way is to . . . oses at him over it.	Who came out?	2 country crooners.	Ty Herndon and Billy Gilman came out as being gay.	Ty Herndon and Billy Gilman
There once was a little old lady who had no friends. She rode her bike to and fr . . . nd her first friend.	Who was antosocial at first?	the little old lady.	the little old lady.	the old lady
Weimar Republic was an unofficial, historical designation for the German state b . . . the Eastern border.	What was an unofficial designation	1919 and 1933.	Weimar Republic It was first used in a book by Wolfgang Mommsen in 1983.	the Weimar Republic
CHAPTER XX A BLIZZARD A bitter wind searched the poplar bluff where George and h . . . the withered brush.	Who were they waiting for?	Edgar, who should have joined them with the team and sledge.	West and the mail-carrier.	Edgar
There was a princess named Janet who was very pretty. She wanted to marry a prin . . . after in the castle.	Who was the princess?	Janet.	Janet.	Janet
Jenny’s family lived in a small apartment in Seattle. One day Jenny came home fr . . . going to like Utah!	Did Jenny help her mom pack?	yes.	Yes.	yes

Table 15: Generated results by all models on OBQA test set.

Fact	Question	RS-Llama2	Llama2	Ground-truth	BLoB	LoRA Ensembles
Sugars are transported from the leaves to the roots of a plan	Glucose travels A) from roots to leaves of a daffodil B) from a rose's leaves to the atmosphere C) from a daisy's leaves into it's underground support system D) from the sun to a sunflower's buds	C	C	C	C	C
The use of alternative fuels increases , the use of gasoline will decrease	As gasoline costs rise, alternative fuels are being used, which means that A) wind power will be expensive B) gas costs will rise C) oil costs will be maintained D) gasoline will be needed less	D	D	D	D	C
The Grand Canyon was formed by the Colorado River flowing over long periods of time	The Grand Canyon was formed by A) a volcano erupting in 1782 B) a river named after the 20th state to join the union flowing over time C) a river named after the 38th state to join the union flowing over time D) the Great Lakes drying up	B	B	C	B	C
Storms cause bodies of water to increase amount of water they contain	After a torrential downpour over a week, a man notices that the pond in his backyard is A) melted B) dehydrated C) bloated D) salted	C	C	C	A	C
Thick feathers can be used for keeping warm	What can feathers on Spheniscidae be used for? A) keeping warm B) flying C) sleeping D) eating	A	A	A	A	B
decomposition increases the amount of nutrients in the soil	They looked where the log decayed to garden as it would leave the earth A) richer B) dryer C) sandy D) harder	A	A	A	A	A
plant requires soil for to grow	A plant will grow strong if it has A) love B) heat C) earth D) sand	C	C	C	C	C
eye color is an inherited characteristic	What is the best way to guess a babies eye color? A) The surroundings they are born in. B) Their parents usual diet. C) Just take a random guess. D) The genealogy records of their family.	D	B	D	D	D