

Refer, Infer, or Abstain: Investigating Latent Skill Selection of Large Language Models under Contextual Question Answering Uncertainty

Emil Kalbaliyev

Institute of Computer Science
University of Tartu
Tartu, Estonia
emil.kalbaliyev@ut.ee

Kairit Sirts

Institute of Computer Science
University of Tartu
Tartu, Estonia
kairit.sirts@ut.ee

Abstract

Contextual question answering (CQA) involves uncertainty regarding the use of available knowledge in forming an answer. In CQA, large language models (LLMs) should decide whether to *Refer* to explicitly stated information, *Infer* from contextual evidence using commonsense knowledge, or *Abstain* when an answer cannot be supported. We investigate whether LLMs implicitly select among these skills under CQA uncertainty. Combining prompting analysis with layer-wise probing, we examine latent skill selection across instruction-tuned LLMs. We find that the models distinguish *Refer* from *Infer* more easily than *Infer* from *Abstain*. In particular, erroneous abstentions disproportionately occur on inference-based questions, and the *Infer*–*Abstain* distinction is the least separable in model representations. We further find that models represent contextual knowledge gaps more strongly than contextual knowledge sufficiency. These results characterize the emergence and limitations of latent skill selection in LLMs under CQA uncertainty.

1 Introduction

Large Language Models (LLMs) are provided with context, either to supplement their knowledge to answer factual questions or to provide background to answer context-specific (contextual) questions. In Contextual Question Answering (CQA), LLMs are expected to provide grounded answers to questions with user-provided context, such as a narrative (Kočíský et al., 2018; Xu et al., 2022), a situation (Gu et al., 2022), or a scenario (Li et al., 2021; Sun et al., 2023). In contrast to RAG-based approaches to factual or open-domain questions, where context is used as an additional source of knowledge (Neeman et al., 2023; Farahani and Johansson, 2024; Cheng et al., 2024; Zhao et al., 2024; Du et al., 2024; Basmov et al., 2025; Huang et al., 2025; Schumacher et al., 2025; Zhao et al.,

2026), context is the main source of knowledge for tailored answers in CQA. Especially, LLMs should refer to the read context to answer explicit questions, infer from context to answer implicit questions (Kalbaliyev and Sirts, 2024), and abstain from unanswerable questions where no correct or plausible answer is derivable. Rogers et al. (2020) describe this as “the full spectrum of uncertainty in QA,” since the task involves determining not only whether the information is explicitly stated in the text, but also whether the missing information can be inferred with reasonable confidence.

While LLMs often produce answers, how they navigate distinct forms of uncertainty remains poorly understood. In this paper, we study QA uncertainty from a dimensional skill perspective (Kalbaliyev and Sirts, 2024, 2025), conceptualizing CQA as requiring three core contextual knowledge skills: *Refer*, *Infer*, or *Abstain*. Under this view, solving a CQA instance involves selecting and applying the appropriate skill. However, these skills are exercised implicitly through latent reasoning processes (Zhu et al., 2025; Li et al., 2025b), and it remains unclear whether LLMs can select the skill required for a given question. We therefore investigate whether LLMs are implicitly aware of the skill needed to answer contextual questions.

To investigate latent skill selection in LLMs, we adopt two different approaches. First, in a prompting-based approach, we examine wrongful abstentions to identify potential confusions between *Abstain* and other skills. This analysis provides behavioral evidence of whether models select the appropriate skill when answering contextual questions. Second, in a probing-based approach, we use layer-wise probing to examine the model’s internal representations and track how skill-related information emerges across layers. This allows us to identify which layers encode information about the required skill and where the model has the capacity to linearly recover it.

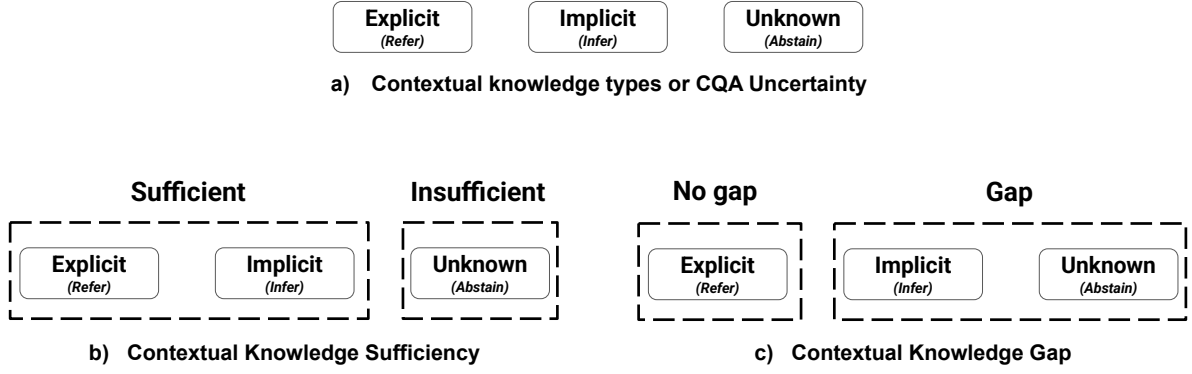


Figure 1: Knowledge types in Contextual Question Answering (CQA) and their groupings in Contextual Knowledge Sufficiency and Gap subdimensions.

2 Background: Contextual Question Answering

There are two types of context-based QA: (1) questions that focus on general knowledge, where the provided contexts serve to supplement that knowledge, and (2) questions that are directly about the contexts themselves (Shaier et al., 2024). In this work, we focus on the second type of questions and refer to it as Contextual Question Answering (CQA). CQA concerns generating answers to questions grounded in a provided context, such as a narrative (Kočíský et al., 2018; Xu et al., 2022), a situation (Gu et al., 2022), or a scenario (Li et al., 2021; Sun et al., 2023). A model is expected to produce a valid response only when the answer is supported by the given context, potentially augmented with appropriate contextual commonsense knowledge. When the available information is insufficient to derive a reliable answer, the model should abstain rather than speculate or hallucinate.

The knowledge required to bridge a context and a question can be broadly categorized into three types:

- *Explicit knowledge*: refers to information directly stated in the context.
- *Implicit knowledge*: refers to information not explicitly stated but inferable by combining the context with world or commonsense knowledge.
- *Unknown knowledge*: refers to information that neither is provided in context nor inferable with world or commonsense knowledge.

We can also organize three knowledge types along two subdimensions that characterize the rea-

soning challenges in CQA: (1) Contextual Knowledge Sufficiency (Joren et al., 2025) and (2) Contextual Knowledge Gap (Kalbaliyev and Sirts, 2024) as shown in Figure 1.

The Contextual Knowledge Sufficiency subdimension captures whether a valid answer exists in principle:

- *Sufficient Knowledge*: Questions for which sufficient knowledge exists to derive an answer, including both explicit and implicit cases.
- *Insufficient Knowledge*: Questions for which available knowledge is insufficient, requiring the model to abstain.

Contextual Knowledge Sufficiency also shares the same categorization as the known-unknown (Amayuelas et al., 2024) and answerable-unanswerable (Madhusudhan et al., 2025) classifications.

The Contextual Knowledge Gap subdimension captures whether the provided context alone is sufficient to answer the question:

- *No Gap*: The answer can be derived directly from information explicitly stated in the context.
- *Gap*: The information required to answer is not directly present in the text. The information can be recovered in implicit cases but not in unknown cases.

CQA Uncertainty and ambiguous position of implicit knowledge. A central uncertainty in CQA lies in the dual nature of implicit knowledge. From the perspective of sufficiency, implicit questions

align with explicit ones, as both are, in principle, answerable. However, from the perspective of the contextual gap, implicit questions align with unknown ones, as both require information beyond what is directly stated.

This dual alignment makes the application of implicit knowledge inherently ambiguous: it is proximal to explicit knowledge in terms of the existence of answers, yet proximal to unknown knowledge in terms of the contextual gap. As a result, implicit reasoning introduces a spectrum-like notion of uncertainty that extends beyond a simple binary distinction between answerable and unanswerable or known and unknown cases. This perspective aligns with prior views that conceptualize question answering as spanning a full spectrum of uncertainty (Rogers et al., 2020).

3 Methodology

3.1 CQA Uncertainty as Latent Skill Selection

We study Contextual Question Answering (CQA) uncertainty from a dimensional skill perspective (Kalbaliyev and Sirts, 2024, 2025), framing CQA in terms of three core contextual knowledge skills as shown in Figure 1:

- **Refer:** refer to information directly from the context (explicit)
- **Infer:** integrate contextual evidence with background knowledge (implicit)
- **Abstain:** recognize and reject unsupported or unanswerable cases (unknown).

This framework treats uncertainty as a latent skill-selection problem, allowing us to examine whether LLMs distinguish among these skills and how this distinction is reflected in their behavior and internal representations.

3.2 Contextual Question Answering

We formulate Contextual Question Answering as a multiple-choice QA task. Let the dataset consist of tuples

$$\mathcal{D} = \{(c_i, q_i, O_i, y_i, s_i)\}_{i=1}^N, \quad (1)$$

where c_i denotes the context, q_i the question, $O_i = (o_1, \dots, o_K)$ the ordered list of candidate answer options, and $y_i \in \{0, \dots, K\}$ the index of the correct option. Each example is assigned a skill label

$$s_i \in \{\text{Refer}, \text{Infer}, \text{Abstain}\}. \quad (2)$$

3.3 Prompting

In prompting, each option o_k starts with a letter label $\ell_k \in \{A, B, C, \dots\}$, in a fixed order. Given the input prompt $x_i = (c_i, q_i, O_i)$, the language model f_θ predicts the answer by comparing the log-probabilities it assigns to the letter tokens at the next-token position:

$$\hat{y}_i = \arg \max_{k \in \{0, \dots, K\}} \log p_\theta(\ell_k | x_i). \quad (3)$$

A prediction is correct if $\hat{y}_i = y_i$.

3.4 Probing

To characterize the emergence of latent skill representations throughout the network, we adopt a layer-wise probing methodology. Given an input instance x_i , we extract the hidden representation $h_i^{(l)} \in \mathbb{R}^d$ from each transformer layer l . For every layer, we train an independent linear probe using logistic regression with L2 regularization to predict the target skill label. The probe receives only the frozen hidden representation as input, while the underlying language model parameters remain unchanged. Since linear probes have limited expressive capacity, high probe performance indicates that information about the required reasoning skill is explicitly encoded and linearly separable within the representation space. Analyzing probe performance across layers allows us to trace where and how latent skill selection emerges during the model’s forward computation.

We design a suite of probing tasks to analyze latent skill representations at multiple levels of abstraction. The three-way probing task evaluates fine-grained latent skill selection by predicting whether a given instance requires Refer, Infer, or Abstain. Pairwise probing tasks further examine the separability of skill representations by focusing on binary distinctions between specific skill pairs, thereby characterizing local decision boundaries in representation space. Finally, meta-skill probing tasks assess higher-order abstractions related to contextual knowledge sufficiency and gaps, capturing whether the model encodes a global awareness. Together, these probing tasks provide a structured, layer-wise characterization of how language models organize and represent reasoning skills under contextual QA uncertainty.

3.4.1 Three-Way Latent Skill Selection

Our primary probing task evaluates whether a model internally represents the specific reasoning

skill required for a contextual QA instance.

For every layer l , a multiclass logistic regression probe is trained to estimate

$$p(s_i | h_i^{(l)}). \quad (4)$$

Predicted skill labels are obtained as

$$\hat{s}_i^{(l)} = \arg \max_{s \in \mathcal{S}} p(s | h_i^{(l)}). \quad (5)$$

This task measures whether the hidden representations encode awareness of the reasoning strategy required to answer the question.

3.4.2 Pairwise Skill Discrimination

While three-way classification evaluates overall skill awareness, it does not reveal how distinctly different skills are represented. To analyze pairwise separability, we construct binary classification tasks for all skill pairs: (Refer, Infer), (Refer, Abstain), (Infer, Abstain).

For each skill pair (s_a, s_b) , we define a restricted dataset

$$\mathcal{D}_{ab} = \{(x_i, s_i) : s_i \in \{s_a, s_b\}\}. \quad (6)$$

A binary logistic regression probe is trained according to

$$p(s | h^{(l)}) = \sigma(w^\top h^{(l)} + b), \quad (7)$$

where $\sigma(\cdot)$ denotes the sigmoid function.

These experiments quantify the separability of skill representations and reveal which skill distinctions emerge most clearly across model layers.

3.4.3 Meta-Skill Awareness

Beyond individual skill prediction, we investigate whether language models encode higher-level abstractions regarding contextual knowledge sufficiency and gap.

Contextual Knowledge Sufficiency Awareness. The first meta-skill task tests whether the model represents if sufficient knowledge exists to answer a question. We define the binary meta-label

$$m_i^{\text{suff}} = \begin{cases} 0, & s_i \in \{\text{Refer}, \text{Infer}\}, \\ 1, & s_i = \text{Abstain}. \end{cases} \quad (8)$$

A binary probe is trained to estimate

$$p(m_i^{\text{suff}} | h_i^{(l)}). \quad (9)$$

This task measures whether the model distinguishes answerable questions from questions with insufficient contextual evidence.

Contextual Knowledge Gap Awareness. The second meta-skill task tests whether the model recognizes a gap in the contextual information needed to answer a question. We define

$$m_i^{\text{gap}} = \begin{cases} 0, & s_i = \text{Refer}, \\ 1, & s_i \in \{\text{Infer}, \text{Abstain}\}. \end{cases} \quad (10)$$

A binary probe is trained to estimate

$$p(m_i^{\text{gap}} | h_i^{(l)}). \quad (11)$$

Refer examples are those where the answer is directly stated in the context, while Infer and Abstain examples involve a contextual gap. The task thus probes whether models internally distinguish direct grounding from cases that require inference or lack the necessary information.

4 Experimental Setup

Dataset. We conduct our experiments using the QuAIL dataset (Rogers et al., 2020), a reading comprehension dataset that provides human annotations for three types of questions: text-based questions, questions requiring world knowledge, and unanswerable questions. We map these categories to the labels *Refer*, *Infer*, and *Abstain*, respectively, and use them in our prompting analysis and as probing targets in our probing experiments.

Models. We evaluate nine instruction-tuned language models from three model families:

- **Llama 3** (Grattafiori et al., 2024): Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct
- **Qwen 2.5** (Qwen et al., 2025): Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct
- **Gemma 3** (Team et al., 2025): gemma-3-1B-IT, gemma-3-4B-IT, gemma-3-12B-IT

These models span multiple parameter scales, enabling us to study how skill representations vary across model families and sizes.

Prompting Setup. We prompt LLMs in a way that log-based evaluation on choice letters would be possible. User Prompt format for question answering is given below:

User Prompt: Question Answering

Read the context and choose the correct option to answer the question. Only respond with the correct option letter.

CONTEXT:
<context>

QUESTION: <question>

OPTIONS:
A. <answer0>
B. <answer1>
C. <answer2>
D. <answer3>

ANSWER:

Probing Setup. For each model, we extract layer-wise hidden representations from last token during prompting and train linear probes to predict skill labels. We use the LogisticRegression implementation in scikit-learn with default parameter values and train each probe for up to 10,000 iterations. All probes are trained on the training split and evaluated on the validation and challenge splits. We report the results on the challenge set in the main text and the results on the validation set in the Appendix A.

Evaluation Metrics. For probing performance, we report macro-averaged F1 (F1-macro) to account for potential class imbalance across skill categories. To evaluate the CQA task performance, we report overall accuracy, as well as skill-wise and skill-balanced accuracy, following [Kalbaliyev and Sirts \(2025\)](#). To observe skill awareness during prompting, we focus on questions for which the model abstains by selecting ‘not enough information’ option. We report the false abstention rate (FAR), i.e., the percentage of Refer and Infer questions on which a model wrongly abstains, and the skill distribution of abstained questions (skill confusion), alongside total number of abstentions for each model.

5 Results

5.1 Prompting

Table 1 summarizes the prompting results of the evaluated LLMs on the contextual QA, challenge set of QuAIL dataset.

5.1.1 Skill Performance

Based on overall skill performance, the gemma-3-12B-it model achieves the highest accuracy, closely followed by Llama-3.1-8B-Instruct, indicating that the larger models are generally better at resolving context-dependent questions. Among the Llama models, performance improves with model size. The Qwen models show lower overall accuracy in larger model sizes compared to other model families. The balanced skill accuracy, which accounts for performance across the three skills, is again highest for gemma-3-12B-it, followed by Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. Notably, all Qwen models score higher on balanced than on overall accuracy, whereas the opposite holds for all Llama and Gemma models. Examining individual skills reveals substantial variation across models. *Refer* accuracy is highest for gemma-3-12B-it, followed by Llama-3.1-8B-Instruct, whereas *Infer* accuracy is highest for Llama-3.1-8B-Instruct. In contrast, *Abstain* accuracy is highest for Qwen2.5-3B-Instruct, substantially exceeding the corresponding scores of the other models. However, the high *Abstain* accuracy of Qwen2.5-3B-Instruct comes at the cost of *Refer* and *Infer* accuracy, indicating over-abstention rather than stronger context-dependent reasoning.

5.1.2 Latent Skill Selection During Abstention

We investigate latent skill selection by analyzing the skills of instances on which the model chooses to abstain. We focus on abstention because it provides an explicit observable signal of the model’s latent decision over whether the available context is sufficient to answer, making it possible to assess which underlying skills are being selected or confused at this decision point. Specifically, we treat the skill labels of abstained instances as a confusion: an abstention on a *Refer* or *Infer* instance indicates that the model has incorrectly selected the *Abstain* skill.

We report false abstention rate (FAR) for skills, skill confusion, and number of abstentions for models on the Abstention section of Table 1. Across model families, *Infer* is consistently the dominant source of erroneous abstentions. This confusion is particularly strong for small language models like Llama-3.2-1B-Instruct (64.49%) and gemma-3-1B-it (63.04%). The FAR shows that this is not merely a size ef-

Model	Skill Performance					Abstention					
						FAR for skills		Skill Confusion			#
	<i>All</i>	<i>Balanced</i>	<i>Refer</i>	<i>Infer</i>	<i>Abstain</i>	<i>Refer</i>	<i>Infer</i>	<i>Refer</i>	<i>Infer</i>	<i>Abstain</i>	
Llama-3.2-1B-Instruct	45.86	43.48	52.69	44.41	33.33	8.60	22.70	14.95	64.49	20.56	107
Llama-3.2-3B-Instruct	73.56	69.79	82.80	72.04	54.55	2.15	9.87	5.71	42.86	51.43	70
Llama-3.1-8B-Instruct	82.01	76.43	88.17	83.55	57.58	2.15	4.93	7.02	26.32	66.67	57
Qwen2.5-1.5B-Instruct	61.15	65.33	73.12	51.64	71.21	10.75	34.87	11.56	61.27	27.17	173
Qwen2.5-3B-Instruct	61.51	69.25	65.05	53.29	89.39	26.88	39.47	21.83	52.40	25.76	229
Qwen2.5-7B-Instruct	72.30	74.37	84.41	64.47	74.24	8.06	30.26	9.62	58.97	31.41	156
gemma-3-1b-it	49.46	41.62	58.06	51.64	15.15	3.76	9.54	15.22	63.04	21.74	46
gemma-3-4b-it	74.28	65.23	83.87	76.97	34.85	1.08	3.62	5.56	30.56	63.89	36
gemma-3-12b-it	82.37	78.26	91.40	81.25	62.12	3.23	13.49	6.82	46.59	46.59	88

Table 1: LLM prompting results on the *challenge* set of the QuAIL dataset that contains 186 *Refer*, 304 *Infer*, and 66 *Abstain* questions. We report overall, skill-balanced, and skill-wise accuracy, following [Kalbaliyev and Sirts \(2025\)](#). We additionally report the false abstention rate (FAR), the percentage of falsely abstained questions belonging to the $\{Refer, Infer\}$ skills, and skill confusion during abstention, i.e., the skill distribution of questions on which the LLMs abstain and the total number (#) of abstentions per model.

fect: every model has a higher FAR for *Infer* than for *Refer*, for example, 22.70% versus 8.60% for Llama-3.2-1B-Instruct and 13.49% versus 3.23% for gemma-3-12B-it. With increasing model scale, Llama exhibits a clearer separation of *Abstain* from the other skills, reflected in the higher proportion of genuinely abstainable instances among abstentions (66.67% for Llama-3.1-8B-Instruct) and a corresponding reduction in infer confusion (26.32%). This is accompanied by a steady drop in *Infer* FAR (22.70% $\rightarrow$ 9.87% $\rightarrow$ 4.93%) and in the number of abstentions (107 $\rightarrow$ 70 $\rightarrow$ 57), which approaches the 66 *Abstain* questions in the set.

Qwen models do not show such a scaling effect. *Infer* remains the dominant source of abstention at every size (52.40–61.27%), and Qwen models abstain far more often than the other families (156–229 abstentions, against 66 *Abstain* questions), with the highest FAR for both *Infer* (30.26–39.47%) and *Refer* (up to 26.88% for Qwen2.5-3B-Instruct). As a result, only 25.76–31.41% of Qwen abstentions fall on genuinely abstainable instances. This suggests that their strong *Abstain* accuracy stems from abstaining indiscriminately rather than from separating *Abstain* from the other skills. Gemma shows a non-monotonic pattern. gemma-3-4b-it has the cleanest separation (63.89% *Abstain* and 30.56% *Infer* among abstentions), the lowest FAR (1.08% for *Refer*, 3.62% for *Infer*), and only 36 abstentions. However, gemma-3-12B-it regresses: *Infer* confusion rises to 46.59%, matching *Abstain*, its *Infer* FAR

increases to 13.49%, and abstentions grow to 88. Model scale therefore does not by itself guarantee cleaner skill selection.

In contrast, *Refer*-to-*Abstain* confusion is consistently smaller, remaining below 22% across models, and the *Refer* FAR stays below 11% for all models except Qwen2.5-3B-Instruct (26.88%). Overall, the pattern suggests that latent skill selection is most vulnerable to confusing the *Infer* skill with *Abstain*. Even the best-separated models, Llama-3.1-8B-Instruct and gemma-3-4b-it, still show far higher *Infer* than *Refer* confusion (26.32% vs. 7.02% and 30.56% vs. 5.56%, respectively).

5.2 Probing

Figure 2 shows the probing results on the challenge set of the QuAIL dataset for the Llama and Qwen models. For better readability, the results for the Gemma models are reported separately in Figure 3 of the Appendix.

5.2.1 Emergence of Latent Skill Selection

Figure 2 shows how skill information emerges across model layers. The Skill score ($\rightarrow$), which captures the three-way distinction between *Refer*, *Infer*, and *Abstain*, starts low in the early layers of all models. It then increases through the early and middle layers as the models integrate more contextual and semantic information. The score reaches a plateau around the middle of the network and remains relatively stable in the later layers. More broadly, the models appear to first encode simpler

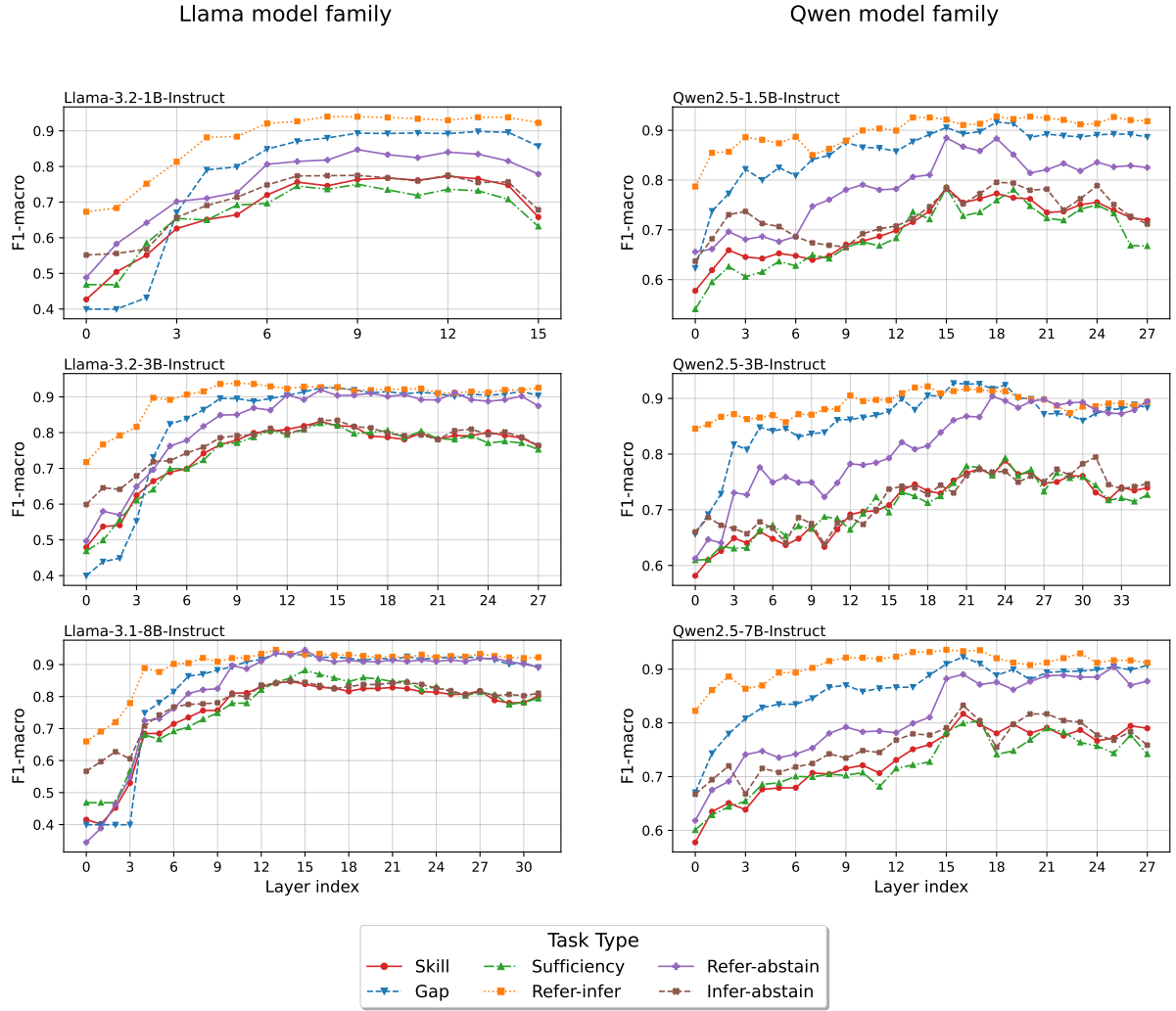


Figure 2: Layer-wise skill probing results on the *challenge* set of the QuAIL dataset for the Llama and Qwen model families. Results for the Gemma model family, as well as results for all model families on the validation set, are provided in the Appendix.

task distinctions and only later develop a more complete representation of the required skill.

5.2.2 Difficulty of Skill Distinctions

The pairwise results show a consistent difference in difficulty. *Refer-Infer* (—■—) is the easiest distinction across all models. It increases rapidly in the early layers and reaches a high, stable level. This indicates that the models can quickly distinguish whether an answer is directly supported by the context or requires inference. *Refer-Abstain* (—◆—) is somewhat harder but still reaches a relatively high level. In contrast, *Infer-Abstain* (—✕—) is consistently more difficult. Its curve rises more slowly and reaches a lower plateau. Thus, distinguishing between cases that require inference and abstention is the hardest pairwise distinction. These results suggest that the difficulty lies mainly in distinguish-

ing *Abstain* from *Infer* while *Refer* is comparatively easy to separate from the other two skills. A similar result was also observed in skill confusion analysis during prompting.

5.2.3 Sufficiency vs. Gap Awareness

The two meta-skills show a similar difference. *Gap* (—▼—), which captures whether the model recognizes a gap in the contextual information needed to answer a question, is highly linearly separable. It rises quickly in the early layers and remains at a high level through most of the network. In contrast, *Sufficiency* (—▲—), which captures whether the model represents if sufficient knowledge exists to answer a question, develops more gradually and reaches a lower plateau. This pattern is consistent with, and can largely be explained by, the pairwise results reported in Section 5.2.2. *Gap* separates

Refer from $\{\text{Infer}, \text{Abstain}\}$, drawing on two of the easier pairwise distinctions, while Sufficiency separates $\{\text{Refer}, \text{Infer}\}$ from *Abstain*, drawing in part on the harder *Infer–Abstain* distinction. Although the meta-skill and pairwise probes are trained independently, their results are conceptually linked. The ease of Gap and the difficulty of Sufficiency both reflect the same underlying asymmetry: *Refer* is easily separated from the other two skills, while *Infer* and *Abstain* are not from each other.

5.2.4 Cross-Model Trends

Several trends are consistent across the different model settings. Model scale affects the stability of skill representations. The smaller models (1B–1.5B) show more variation across layers and retain a larger gap between easy and difficult tasks. The larger models (7B–12B) generally show faster increases and reach stable levels earlier relative to network depth. Their harder tasks also become closer to the easier ones in the later layers. Finally, the three model families: Llama, Qwen, and Gemma, show the same overall pattern: an early increase in separability, a divergence between task types in the middle layers, and a later plateau. This consistency suggests that the observed pattern is not specific to one model family but reflects a broader property of instruction-tuned transformer language models.

6 Related Works

Our work relates to research on contextual question answering, contextual knowledge use, uncertainty and abstention in language models, and model knowledge awareness. We focus on how LLMs select among different knowledge-use skills when answering questions under contextual uncertainty, with particular attention to the distinction between inference and abstention.

Contextual Question Answering and Knowledge Skills. Theoretical work on contextual question answering has explored the prerequisite skills underlying reading comprehension (Sugawara et al., 2017a,b), multidimensional skill taxonomies (Schlegel et al., 2020; Rogers et al., 2023; Kalbaliyev and Sirts, 2024), and desiderata for effective context use (Shaier et al., 2024). These works emphasize that answering a question from context can require different forms of knowledge use, including retrieving explicitly stated information and integrating information through inference.

In our work, we adopt the knowledge-skill dimension defined by Kalbaliyev and Sirts (2024), with the individual skills adapted from the taxonomy of Rogers et al. (2020). We extend the knowledge-skill dimension by considering *Abstain* as an additional skill when neither contextual evidence nor appropriate background knowledge supports an answer.

A substantial body of work has developed datasets for evaluating different aspects of contextual question answering (Kočíský et al., 2018; Choi et al., 2018; Huang et al., 2019; Rogers et al., 2020; Li et al., 2021; Gu et al., 2022; Xu et al., 2022; Sun et al., 2023). We use the QuAIL dataset (Rogers et al., 2020), which is particularly suitable for our analysis because its human annotations can be mapped to the knowledge skills examined in our experiments. Rather than only using QA instances to evaluate answer accuracy, we use mapped skill annotations to investigate whether LLMs internally distinguish among knowledge-use strategies.

Contextual and Parametric Knowledge. Another line of work investigates how LLMs use and combine contextual and parametric knowledge (Zhao et al., 2024; Jin et al., 2025a; Yuan et al., 2025; Hagström et al., 2026). This issue has received particular attention in retrieval-augmented generation, where contextual information is typically introduced to complement or correct knowledge stored in model parameters (Neeman et al., 2023; Farahani and Johansson, 2024; Cheng et al., 2024; Zhao et al., 2024; Du et al., 2024; Basmov et al., 2025; Huang et al., 2025; Schumacher et al., 2025; Zhao et al., 2026). While this literature primarily examines whether and how models incorporate external information, we consider a different aspect of context use: cases in which contextual evidence is necessary to determine the appropriate knowledge source or response behavior. In particular, we investigate whether models can distinguish questions that can be answered by referring to the context, those requiring inference from contextual evidence and background knowledge, and those for which abstention is warranted.

Uncertainty and abstention in LLMs. Uncertainty in language models has been studied from several perspectives, including general frameworks for characterizing uncertainty (Cui et al., 2025), semantic uncertainty (Kuhn et al., 2023), episodic uncertainty (Ahdritz et al., 2024; Abbasi-Yadkori et al., 2024), known-unknown uncertainty

(Amayuelas et al., 2024; Li et al., 2025a), and denotational uncertainty (Cole et al., 2023). Other work has investigated uncertainty quantification (Shorinwa et al., 2025; Shelmanov et al., 2025), including in contextual question answering (Bakman et al., 2026), as well as how uncertainty is expressed by language models (Zhou et al., 2024). Closely related work studies abstention, including surveys of the field (Wen et al., 2025), methods for eliciting hallucination or improving abstention (Feng et al., 2024; Kirichenko et al., 2025; Kim et al., 2025; Liu et al., 2025), and empirical investigations of LLM abstention abilities (Madhusudhan et al., 2025). These studies generally evaluate whether models can recognize uncertainty or refrain from answering when appropriate. Our work complements this perspective by examining abstention as part of a broader skill-selection problem. Specifically, we ask whether models can distinguish an unsupported question from an inference-based question, and whether failures to abstain arise from difficulty in identifying a knowledge gap or from confusion with inference.

Knowledge Awareness and Internal Representations. Our work is also closely related to studies of knowledge awareness in language models. Prior research has examined whether models know what they know (Jiang et al., 2020, 2021; Kadavath et al., 2022; Yin et al., 2023; Prato et al., 2024; Amayuelas et al., 2024; Kapoor et al., 2024), including whether internal representations contain signals about the correctness or availability of model knowledge. This line of work provides evidence that LLMs can encode information about their own knowledge states. Probing (Alain and Bengio, 2018; Belinkov, 2022) has been widely used to analyze the information encoded in LLM representations. Prior work has used probing to investigate whether models encode separable representations of true and false statements (Azaria and Mitchell, 2023; Marks and Tegmark, 2024), as well as to examine how knowledge and concepts are acquired and represented across different layers of LLMs (Jin et al., 2025b). Through layer-wise probing, we examine whether representations encode distinctions among referring to contextual information, inferring an answer, and abstaining.

7 Conclusion

We investigated uncertainty in contextual question answering (CQA) through the lens of three

latent skills: *Refer*, *Infer*, and *Abstain*. Across nine instruction-tuned large language models (LLMs), we combined prompting analysis with layer-wise probing to examine whether models recognize the contextual knowledge skill required by a question.

The prompting results show that overall and balanced accuracy do not always align and high *Abstain* accuracy can also be misleading. Some models achieve it by abstaining indiscriminately, rather than by accurately recognizing when context is insufficient. Analysis of skill selection during abstention shows that *Infer* is consistently the dominant source of erroneous abstentions across models. While larger models within a family sometimes better separate *Abstain* from the other skills, this improvement is not guaranteed by scale alone and varies across model families.

Our probing results reinforce this pattern at the representational level: models more easily distinguish *Refer* from *Infer* and *Abstain* skills, while the *Infer*–*Abstain* boundary remains consistently the hardest to represent, mirroring the errors observed during prompting. We also find that models encode contextual knowledge gaps more readily than contextual knowledge sufficiency due to the same underlying asymmetry in pairwise separability. Finally, model scale improves the stability and speed with which skill representations emerge across layers, but does not close the gap between easy and hard distinctions, and the overall layer-wise trajectory is consistent across model families.

Taken together, our findings show that uncertainty in CQA is not simply a binary answer-or-abstain decision but involves distinct, unequally represented contextual knowledge skills, with the boundary between inference-based answering and abstention posing the most persistent challenge for current instruction-tuned LLMs.

Limitations

Our findings should be interpreted in light of several methodological limitations. First, our analysis relies primarily on a single dataset. To our knowledge, QuAIL (Rogers et al., 2020) is the only dataset with human-annotated labels that map directly onto the three skills we study. This limits the extent to which we can assess the generality of the observed skill-selection patterns across different contextual QA settings. Second, probing demonstrates that skill information is linearly recoverable from model representations, but does not

establish that models causally use this information when making their final decisions. Similarly, interpreting abstentions as evidence of latent skill selection assumes that they reflect abstention behavior, whereas they may also be influenced by instruction-following behavior or other generation preferences. Third, we use the dataset as is and do not test for biases in multiple-choice question answering (Balepur et al., 2025) or other consistency issues in knowledge probing (Zhao et al., 2025). Finally, as the primary goal of this work is to introduce and characterize latent skill selection under contextual QA uncertainty, we focus on relatively lightweight instruction-tuned models rather than computationally demanding larger language models or reasoning models. We left the evaluation of our methods on additional datasets and models, and the use of additional methods to address the stated limitations for future work.

Acknowledgments

The model training and inference were carried out in the High Performance Computing Center of the University of Tartu (University of Tartu, 2018).

References

- Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. [To Believe or Not to Believe Your LLM: Iterative Prompting for Estimating Epistemic Uncertainty](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. 2024. [Distinguishing the Knowable from the unknowable with language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 503–549. PMLR.
- Guillaume Alain and Yoshua Bengio. 2018. [Understanding intermediate layers using linear classifier probes](#). *Preprint*, arXiv:1610.01644.
- Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhui Chen, and William Yang Wang. 2024. [Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, Bangkok, Thailand. Association for Computational Linguistics.
- Amos Azaria and Tom Mitchell. 2023. [The Internal State of an LLM Knows When It’s Lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.
- Yavuz Faruk Bakman, Sungmin Kang, Zhiqi Huang, Duygu Nur Yaldiz, Catarina G Belém, Chenyang Zhu, Anoop Kumar, Alf Samuel, Daben Liu, Salman Avestimehr, and Sai Praneeth Karimireddy. 2026. [Uncertainty as Feature Gaps: Epistemic Uncertainty Quantification of LLMs in Contextual Question-Answering](#). In *The Fourteenth International Conference on Learning Representations*.
- Nishant Balepur, Rachel Rudinger, and Jordan Boyd-Graber. 2025. [Which of These Best Describes Multiple Choice Evaluation with LLMs? A\) Forced B\) Flawed C\) Fixable D\) All of the Above](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.
- Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2025. [LLMs’ Reading Comprehension Is Affected by Parametric Knowledge and Struggles with Hypothetical Statements](#). *Preprint*, arXiv:2404.06283.
- Yonatan Belinkov. 2022. [Probing Classifiers: Promises, Shortcomings, and Advances](#). *Computational Linguistics*, 48(1):207–219.
- Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. [Understanding the Interplay between Parametric and Contextual Knowledge for Large Language Models](#). *Preprint*, arXiv:2410.08414.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentaoh Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. [QuAC: Question Answering in Context](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.
- Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. [Selectively Answering Ambiguous Questions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.
- Shaobo Cui, Luca Mouchel, and Boi Faltings. 2025. [Uncertainty in Causality: A New Frontier](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8022–8044, Vienna, Austria. Association for Computational Linguistics.
- Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. [Context versus Prior Knowledge in Language Models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand. Association for Computational Linguistics.

- Mehrdad Farahani and Richard Johansson. 2024. [Deciphering the Interplay of Parametric and Non-parametric Memory in Retrieval-augmented Language Models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16966–16977, Miami, Florida, USA. Association for Computational Linguistics.
- Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. [Don’t Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022. [DREAM: Improving Situational QA by First Elaborating the Situation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.
- Lovisa Hagström, Youna Kim, Haeun Yu, Sang goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. 2026. [CUB: Benchmarking Context Utilisation Techniques for Language Models](#). *Preprint*, arXiv:2505.16518.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning](#). 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.
- Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025. [To Trust or Not to Trust? Enhancing Large Language Models’ Situated Faithfulness to External Contexts](#). In *The Thirteenth International Conference on Learning Representations*.
- 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, Jun Araki, and Graham Neubig. 2020. [How Can We Know What Language Models Know?](#) *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025a. [Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding](#). In *Forty-second International Conference on Machine Learning*.
- Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025b. [Exploring Concept Depth: How Large Language Models Acquire Knowledge and Concept at Different Layers?](#) In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 558–573, Abu Dhabi, UAE. Association for Computational Linguistics.
- Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. [Sufficient Context: A New Lens on Retrieval Augmented Generation Systems](#). In *The Thirteenth International Conference on Learning Representations*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language Models \(Mostly\) Know What They Know](#). *Preprint*, arXiv:2207.05221.
- Emil Kalbaliyev and Kairit Sirts. 2024. [On Narrative Question Answering Skills](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 814–820, Mexico City, Mexico. Association for Computational Linguistics.
- Emil Kalbaliyev and Kairit Sirts. 2025. [Towards Evaluation of Language Models with Skill Dimensions: A Case Study on Narrative Question Answering](#). In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 430–440, Suzhou, China. Association for Computational Linguistics.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. [Large Language Models Must Be Taught to Know What They Don’t Know](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. 2025. [When to Speak, When to Abstain: Contrastive Decoding with Abstention](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

- Long Papers*), pages 9710–9730, Vienna, Austria. Association for Computational Linguistics.
- Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. [AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions](#). *Preprint*, arXiv:2506.09038.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. [The NarrativeQA reading comprehension challenge](#). *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation](#). In *The Eleventh International Conference on Learning Representations*.
- Jiaqi Li, Yixuan Tang, and Yi Yang. 2025a. [Know the unknown: An uncertainty-sensitive method for LLM instruction tuning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2972–2989, Vienna, Austria. Association for Computational Linguistics.
- Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025b. [Implicit Reasoning in Large Language Models: A Comprehensive Survey](#). *Preprint*, arXiv:2509.02350.
- Xiao Li, Yawei Sun, and Gong Cheng. 2021. [TSQA: Tabular Scenario Based Question Answering](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13297–13305.
- Jingyu Liu, JingquanPeng, Xiaopeng Wu, Xubin Li, Tiezheng Ge, Bo Zheng, and Yong Liu. 2025. [Do not Abstain! Identify and Solve the Uncertainty](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17177–17197, Vienna, Austria. Association for Computational Linguistics.
- Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. [Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345, Abu Dhabi, UAE. Association for Computational Linguistics.
- Samuel Marks and Max Tegmark. 2024. [The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets](#). *Preprint*, arXiv:2310.06824.
- Ella Neeman, Roei Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. [DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.
- Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, and Sarath Chandar. 2024. [Do Large Language Models Know How Much They Know?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6070, Miami, Florida, USA. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 Technical Report](#). *Preprint*, arXiv:2412.15115.
- Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. [Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension](#). *ACM Comput. Surv.*, 55(10).
- Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. [Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.
- Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. [A Framework for Evaluation of Machine Reading Comprehension Gold Standards](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.
- Dan Schumacher, Fatemeh Haji, Tara Grey, Niharika Bandlamudi, Nupoor Karnik, Gagana Uday Kumar, Cho-Yu Jason Chiang, Peyman Najafirad, Nishant Vishwamitra, and Anthony Rios. 2025. [RASTeR: Robust, Agentic, and Structured Temporal Reasoning](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3098–3123, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Sagi Shaiyer, Lawrence Hunter, and Katharina von der Wense. 2024. [Desiderata For The Context Use Of Question Answering Systems](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 777–792, St. Julian’s, Malta. Association for Computational Linguistics.
- Artem Shelmanov, Maxim Panov, Roman Vashurin, Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin. 2025. [Uncertainty Quantification for Large Language Models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

- Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4, Vienna, Austria. Association for Computational Linguistics.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. [A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions](#). *ACM Comput. Surv.*, 58(3).
- Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017a. [Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817, Vancouver, Canada. Association for Computational Linguistics.
- Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017b. [Prerequisite Skills for Reading Comprehension: Multi-Perspective Analysis of MCTest Datasets and Systems](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2023. [Scenario-based Question Answering with Interacting Contextual Properties](#). In *The Eleventh International Conference on Learning Representations*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 Technical Report](#). *Preprint*, arXiv:2503.19786.
- University of Tartu. 2018. [UT Rocket](#).
- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. [Know Your Limits: A Survey of Abstention in Large Language Models](#). *Transactions of the Association for Computational Linguistics*, 13:529–556.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. [Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. [Do Large Language Models Know What They Don’t Know?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.
- Xiaowei Yuan, Zhao Yang, Ziyang Huang, Yequan Wang, Siqu Fan, Yiming Ju, Jun Zhao, and Kang Liu. 2025. [Exploiting contextual knowledge in LLMs through \$\mathcal{V}\$ -usable information based layer enhancement](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31726–31741, Vienna, Austria. Association for Computational Linguistics.
- Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2026. [Understanding Parametric and Contextual Knowledge Reconciliation within Large Language Models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Raoyuan Zhao, Abdullatif Köksal, Ali Modarressi, Michael A. Hedderich, and Hinrich Schuetze. 2025. [Do We Know What LLMs Don’t Know? A Study of Consistency in Knowledge Probing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23254–23280, Suzhou, China. Association for Computational Linguistics.
- Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. [Enhancing Contextual Understanding in Large Language Models through Contrastive Decoding](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237, Mexico City, Mexico. Association for Computational Linguistics.
- Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. [Relying on the Unreliable: The Impact of Language Models’ Reluctance to Express Uncertainty](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.
- Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, and 14 others. 2025. [A Survey on Latent Reasoning](#). *Preprint*, arXiv:2507.06203.

A Additional Results

Figure 3 shows results for the challenge set of the QuAIL dataset, which complements Figure 2.

Table 2 and Figures 4, 5, 6 show results for the validation set of the QuAIL dataset and display similar trends to the challenge set discussed in the main text.

gemma model family

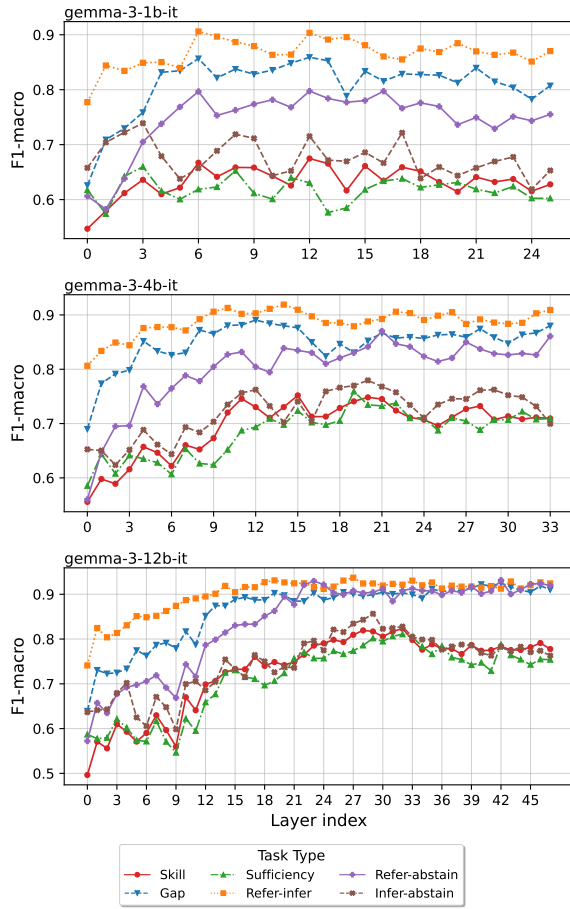


Figure 3: Layer-wise skill probing results on the *challenge* set of the QuAIL dataset for the Gemma model family.

B Data and Code Availability

The QuAIL dataset is publicly available on Hugging Face.¹ The mapping from the question types to text-based questions (*Refer*), questions requiring world knowledge (*Infer*), and unanswerable questions (*Abstain*) is based on the QuAIL dataset paper (Rogers et al., 2020). The code for experiments is available on GitHub.²

¹<https://huggingface.co/datasets/textmachinelab/quail>

²<https://github.com/EmilKalbaliyev/Latent-Skill-Selection>

Llama model family

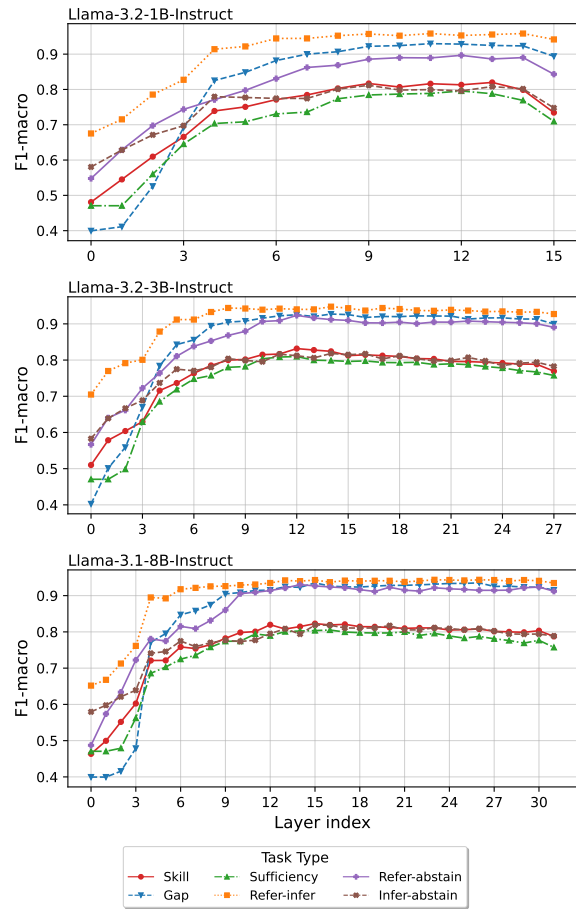


Figure 4: Layer-wise skill probing results on the *validation* set of the QuAIL dataset for the Llama model family.

Model	Skill Performance					Abstention						#
						FAR for skills		Skill Confusion				
	All	Balanced	Refer	Infer	Abstain	Refer	Infer	Refer	Infer	Abstain		
Llama-3.2-1B-Instruct	49.31	44.65	61.46	46.67	25.83	4.97	22.17	9.89	73.08	17.03	364	
Llama-3.2-3B-Instruct	72.55	66.20	83.43	71.83	43.33	3.73	12.75	9.51	53.87	36.62	284	
Llama-3.1-8B-Instruct	76.57	70.63	86.05	76.25	49.58	3.31	12.25	8.28	50.69	41.03	290	
Qwen2.5-1.5B-Instruct	64.23	66.32	78.45	55.50	65.00	9.53	33.17	11.08	63.88	25.04	623	
Qwen2.5-3B-Instruct	62.66	67.09	72.93	54.17	74.17	20.44	38.25	18.85	58.47	22.68	785	
Qwen2.5-7B-Instruct	71.90	72.30	85.50	64.75	66.67	6.63	28.25	8.78	61.97	29.25	547	
gemma-3-1b-it	52.26	41.45	60.77	56.08	7.50	1.66	7.58	9.92	75.21	14.88	121	
gemma-3-4b-it	72.00	62.58	85.50	72.25	30.00	1.52	9.08	5.73	56.77	37.50	192	
gemma-3-12b-it	74.35	70.97	88.67	69.67	54.58	3.31	20.50	5.99	61.35	32.67	401	

Table 2: LLM prompting results on the *validation* set of the QuAIL dataset that contains 724 *Refer*, 1200 *Infer*, and 240 *Abstain* questions. We report overall, skill-balanced, and skill-wise accuracy, following [Kalbaliyev and Sirts \(2025\)](#). We additionally report the false abstention rate (FAR), the percentage of falsely abstained questions belonging to the {*Refer*, *Infer*} skills, and skill confusion during abstention, i.e., the skill distribution of questions on which the LLMs abstain and the total number (#) of abstentions per model.

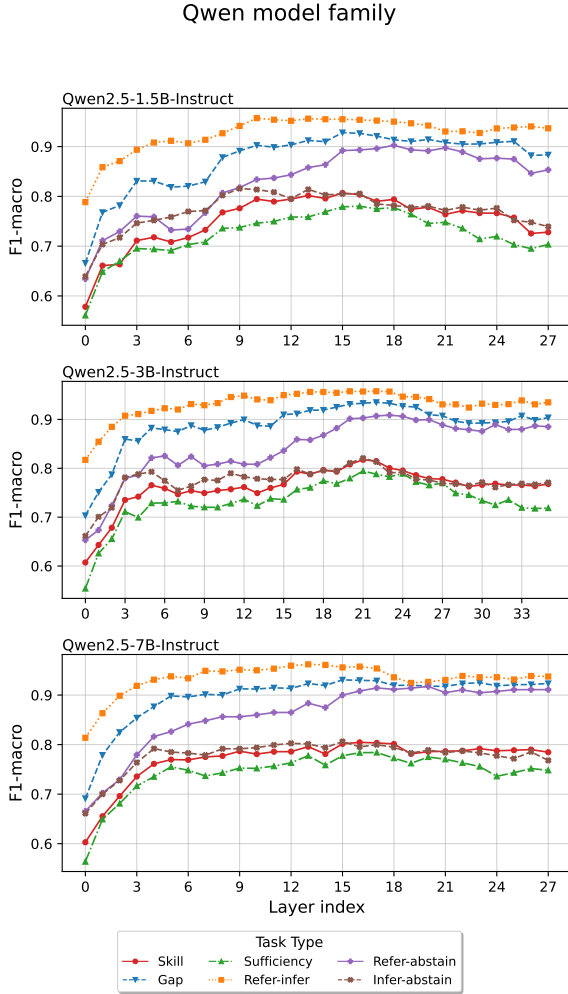


Figure 5: Layer-wise skill probing results on the *validation* set of the QuAIL dataset for the Qwen model family.

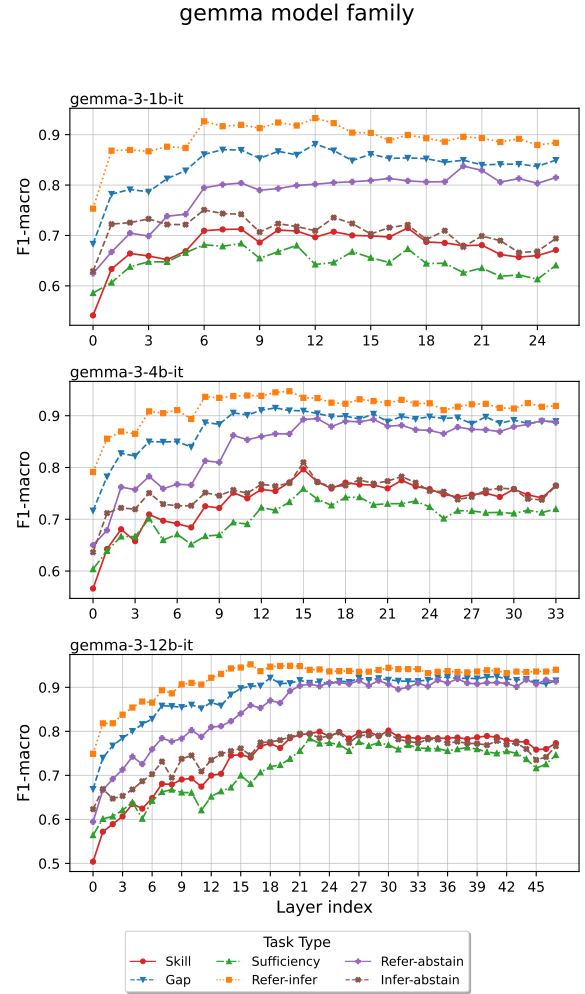


Figure 6: Layer-wise skill probing results on the *validation* set of the QuAIL dataset for the Gemma model family.