Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs

Anonymous ACL submission

Abstract

We observe a novel phenomenon, *contextual entrainment*, across a wide range of language models (LMs) and prompt settings, providing a new mechanistic perspective on how LMs become distracted by "irrelevant" contextual information in the input prompt. Specifically, LMs assign significantly higher logits (or probabilities) to any tokens that have previously appeared in the context prompt, even for random tokens. This suggests that contextual entrainment is a *mechanistic* phenomenon, occurring independently of the relevance or semantic relation of the tokens to the question or the rest of the sentence. We find statistically significant evidence that the magnitude of contextual entrainment is influenced by semantic factors. Counterfactual prompts have a greater effect compared to factual ones, suggesting that while contextual entrainment is a mechanistic phenomenon, it is modulated by semantic factors.

We hypothesize that a cluster of attention heads — the *entrainment heads* — corresponds to contextual entrainment. Using a novel entrainment head discovery method based on differentiable masking, we identify these heads across various settings. When we "turn off" these heads, i.e., set their output to zero, the effect of contextual entrainment is significantly attenuated, causing the model to generate output that capitulates to what it would produce if no distracting context were provided. Our discovery of contextual entrainment, along with our investigation into LM distraction via the entrainment heads, marks a key step towards mechanistic analysis and mitigation of the distraction problem.[1]

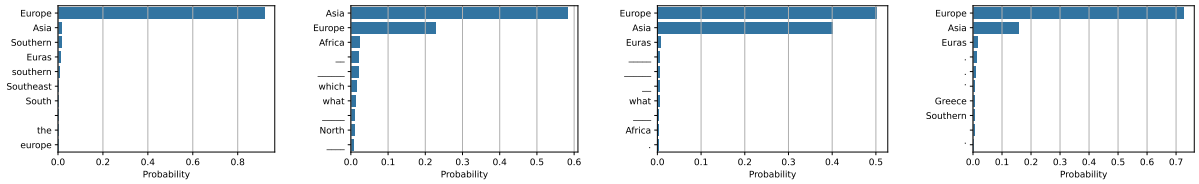
1 Introduction

Language models (LMs), especially large language models (LLMs), can sophisticatedly utilise contextual information provided in prompts to a surprising degree. Brown et al. (2020) was among the first to identify this capability and coin the term *in-context learning* (ICL) to describe this capability. Subsequent work has demonstrated that LMs can process and process and utilise contextual information provided in prompts across various settings.

Nonetheless, LMs can also misuse contextual information in prompts (Figure 1). Shi et al. (2023) experimented with inserting distracting, irrelevant information into grade-school maths problems and found that it successfully diverted the model from reaching the correct answer. Their work shed light on a fundamental issue within LMs, which they termed *distraction*. Since then, distraction has been recognised as one of the most challenging and widespread issues for RAG (Yoran et al., 2023; Cuconasu et al., 2024; Wu et al., 2024), prompting the development of distraction mitigation strategies. Notably, Yoran et al. (2023) proposed leveraging an NLI model to remove irrelevant context from the prompt as a solution to this problem.

Distraction, however, is a phenomenon in LMs that is easy to grasp but difficult to define precisely. Most prior work defines distraction using the term "(ir)relevant," framed in RAG and information retrieval terms; i.e., whether the context prompt contains the information needed to answer the question correctly. While this provides an adequate general description of the problem, we identify challenges when examining it in greater detail. First, *relevance* is too broad a concept. Consider the following context prompts: (1) `Messi is a football player`, (2) `Japan is in Asia`, (3) `Greece is in Asia`, and (4) `Colorless green ideas sleep furiously`. According to the earlier definition, all of these are 'irrelevant' since they do not contain the information needed to correctly answer the question `Greece is located on the continent of` __. However, it is evident that they differ drastically in how they might influence an LM's response, and therefore, a more precise definition and more fine-grained taxonomy of distraction is needed. Moreover, we find evidence that

[1] The code and data of this work will be publicly available online. *GitHub URL withdrawn for submission.*



`Greece is located on the continent of`

(a) When no context is provided to the model, it can confidently generate the correct response.

`CONTEXT: Iraq is located on the continent of Asia. PROMPT: Greece is located on the continent of`

(b) However, when context from a related topic is provided, the model may become distracted and produce incorrect responses.

`CONTEXT: Asia is the largest continent in the world by both land area and population. PROMPT: Greece is located on the continent of`

(c) The model continues to exhibit distractions across different formats and paraphrasing of contextual information.

`CONTEXT: Asia. PROMPT: Greece is located on the continent of`

(d) Just a single token, *Asia*, can distract the model. This indicates that the phenomenon of LLM distraction is more superficial and fundamental than previously suggested.

Figure 1: LLMs can be distracted by various types of context. Each sub-figure illustrates `Llama-3.1-8B`'s output probability of the top ten tokens given the input prompt. The model inputs are displayed in `green`.

these "irrelevant" context prompts can benefit the LMs' performance. Although they do not contain the exact answer, they may provide useful implicit information about the question.

Regardless of the debate on the precise definition of distraction, we observe a phenomenon in LMs related to how they use and misuse information from the context prompt. In particular, we identify *contextual entrainment* — LMs consistently assign significantly higher probabilities (or logits) to any tokens that have appeared earlier in the context prompt, regardless of their relevance or semantic relation to the question or the prompt. Simply put: *llama see, llama do*. If a token appears in the context prompt, even a randomly token, the model assigns it a higher probability or logit. Our experiments show that various LMs exhibit contextual entrainment across a wide range of configurations.

At first glance, this phenomenon shares some similarities with the inductive pattern-repeating phenomenon identified by (Elhage et al., 2021; Olsson et al., 2022), the differences are far more substantial. First, contextual entrainment does not require the reappearance of a prefix as a trigger, unlike inductive pattern-repeating, which depends on first encountering a pattern `[A][B]` and then seeing the prefix `[A]` again to generate `[B]`. Instead, contextual entrainment occurs when a token has previously appeared in the context. Second, unlike the inductive pattern-repeating phenomenon, which is largely independent of semantic factors and token statistics (Olsson et al., 2022), the magnitude of the effect of contextual entrainment is influenced by semantic factors. In particular, we find that counterfactual prompts have a significantly greater effect. Thus, we identify contextual entrainment as a novel phenomenon that plays a crucial role in LMs' use and misuse of contextual information in prompts. It may be a major factor contributing to the distraction problem.

Lastly, in contrast to the previous discussion on how this phenomenon differs from inductive pattern-repeating, here we identify a similarity. Similar to the induction heads identified by Olsson et al. (2022), we find that 3–10% of the attention heads are associated with contextual entrainment, which we refer to as *entrainment heads*. When these entrainment heads are disabled, i.e., their output is set to zero, the effect of contextual entrainment is drastically suppressed and the LM generates outputs similar to that produced when no context prompts are given. This provides an interesting insight into the mechanism governing the contextual entrainment phenomenon, and we hope our work can serve as a starting point for investigating the problem of distraction mechanistically.

Contributions In this paper, we identify **contextual entrainment**, a novel phenomenon that plays a crucial role in LM distraction (§3). We provide evidence that the phenomenon can be considered mechanistic and occurs commonly across various LMs and settings (§3.2). However, it is also influenced by semantic factors: counterfactual context more effectively induce contextual entrainment (§3.2). Finally, we identify **entrainment heads** (§4) using a novel method based on differentiable masking which, when "turned off," the effect of contextual entrainment is drastically suppressed.

2 Related Work

Distraction LMs are known to be susceptible to distractions caused by contextual information in prompts. For instance, Shi et al. (2023) found

Context Setting	Context Prompt	Query Prompt	🥺	🙂
Distract	`On the inside, bananas are white.`	`What color are mangoes on the inside? They are`	`white`	`orange`
Irrelevant	`The capital of Canada is Ottawa.`	`What color are mangoes on the inside? They are`	`Ottawa`	`orange`
Random	`Promotion`	`Greece is located on the continent of`	`Promotion`	`orange`
Distraction over Counterfactual Context			😈	🥺	🙂
Counterfactual	`On the inside, bananas are green.`	`What color are mangoes on the inside? They are`	`green`	`white`	`orange`

Table 1: Prompt Setup. The emojis represent the target tokens: 😈: counterfactual; 🥺: distracting; 🙂: correct.

that while LMs can accurately solve grade-school maths problems, they may fail when provided with additional information in the prompt. They, however, did not explore the mechanisms underlying these distractions. More research has since confirmed that LMs can be easily distracted (Yoran et al., 2023; Wu et al., 2024; Cuconasu et al., 2024, *inter alia*). This problem has received remarkable attention in the RAG community due to its apparent connection to retrieval robustness. Since retrievers cannot always retrieve perfectly relevant documents, the LLM within a RAG system should be as resistant to distraction as possible.

Mechanistic Interpretability & Induction Heads Our research setup shares similarities with efforts to understand ICL, but, as discussed earlier, the differences are more significant. Elhage et al. (2021); Olsson et al. (2022) successfully identified an inductive pattern-repeating phenomenon, a crucial step toward understanding how LMs perform ICL. They observed that if an LM has encountered a sequence of tokens — even for random tokens — it will repeat the sequence if it appears again in the prompt. For instance, given the input `Category 40 ids node struction ... Category 40 ids node`, the model predicts `struction` as the most probable next token. Therefore, they concluded that LMs possess some inductive capability and "are not memorising a fixed table of n-gram statistics." They also identified certain attention heads as *induction heads* in two-layer toy transformer models, which they claim perform pattern completion. More recently, Crosbie and Shutova (2024) identified induction heads in real-world LMs such as `Llama-3-8B`.

3 Context Entrainment

We present experiments that confirm the existence of the contextual entrainment phenomenon in this section. We find that contextual entrainment is both a mechanistic phenomenon but influenced by semantic factors. The phenomenon is observed in all prompt settings, including those with completely randomly sampled tokens, suggesting that its existence is independent of semantic factors (i.e., contextual entrainment is mechanistic); however, the exact magnitude of its impact depends on semantic factors, as demonstrated by the significantly greater effect of counterfactual context.

3.1 Experimental Setup

Prompts & Data Table 1 shows how we constructed our prompts using facts from the LRE dataset (Hernandez et al., 2024), which contains facts in the triplet format: $\langle$*source*, *target*, *relation*$\rangle$ or $\langle s, t, r\rangle$. For example, $\langle$*Canada*, *Ottawa*, *capital*$\rangle$ corresponds to the fact that Canada's capital is Ottawa. Each fact in the dataset contains several prompt templates that we leverage to construct the context and query portion of the prompts.[2]

We present the model with a context and a query in the prompt, separated by a single space character (e.g., `<context> <query>`). Given a query generated from a fact $\langle s, t, r\rangle$, there are four context prompt settings: **distraction**, where facts $\langle s', t', r\rangle$ are sampled from the same relation type r but differ in source s' and target t'; **irrelevant**, where facts $\langle s', t', r'\rangle$ are sampled from a completely different relation type r' without domain or range overlap; **random**, where the context consists of a single randomly chosen token; and **counterfactual**, where the target in the fact $\langle s', t'_{\text{cf}}, r\rangle$ is replaced by another target sampled from the same relation. The random tokens are sampled from the Brown corpus (Francis and Kucera, 1979). For larger relations that yield more than 100,000 combinations, we cap the size at 100,000 through random sampling.

Language Models We experiment with `GPT2 XL` (Radford et al., 2019) and 4 LLaMA models (Touvron et al., 2023): `Llama-3.1-8B`, `Llama-3.1-8B-Instruct`, `Llama-2-7b-hf`, and `Llama-2-13b-hf`.

3.2 Experiment Results

Figure 2 presents the experimental results for three types of distracting contexts: distracting context from a related topic (Distract), irrelevant topic (Irrelevant), and random token (Random). The figure shows the averaged results across all LRE relations.

[2]Appendix A presents more details of the LRE dataset.

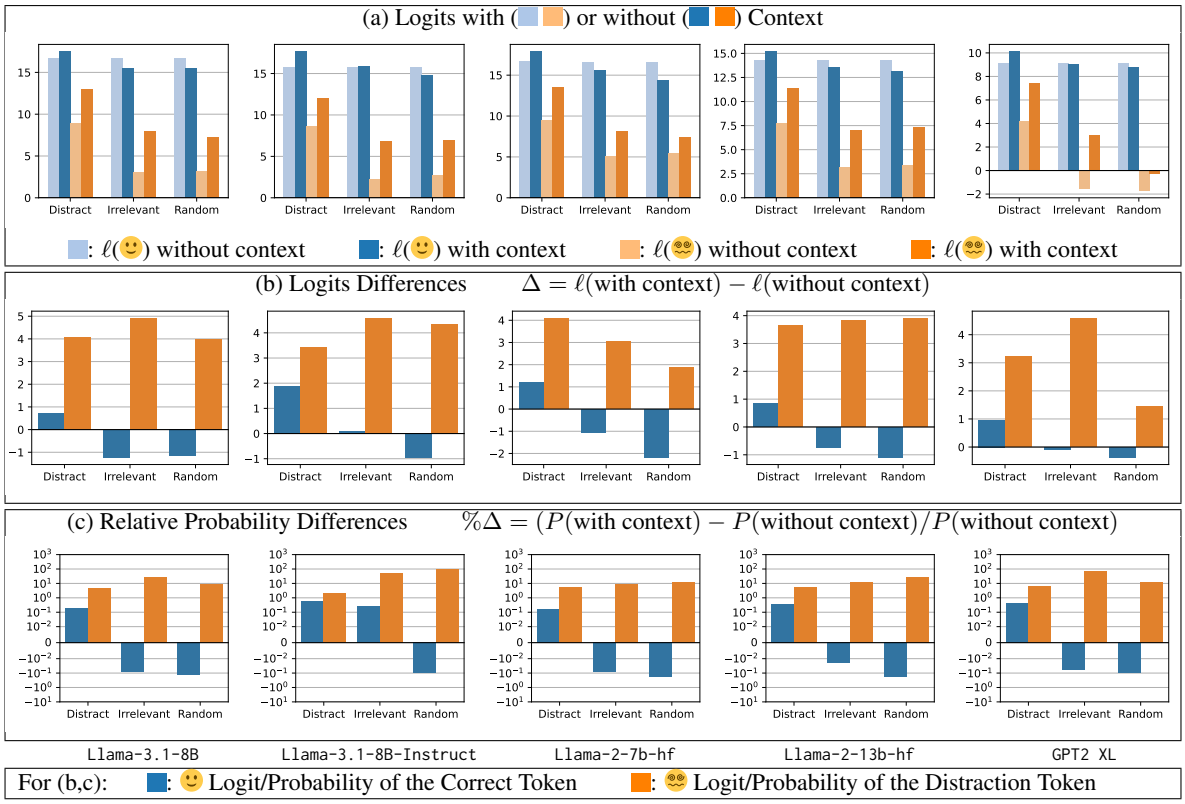


Figure 2: Logit and probability values change consistently after the context prompt is provided to the model. The LM assigns significantly higher logits and probabilities to tokens that appear in the context prompts. All shifts in probabilities and logits are statistically significant, with $p < 0.0001$ according to paired t-tests.

The full list of experimental results, which supports the same conclusion, can be found in Appendix B.

Finding 1: Contextual Entrainment: LMs assign higher logits and probabilities to tokens appearing in the context. When a model is given distracting context, there is a significant increase in the logits and probabilities of the corresponding distracting tokens. For example, when asked the question `Greece is located in _`, distracting context prompts such as `Japan is in Asia`, `Bananas are yellow`, or even a single randomly sampled token, `Promotion`, can cause the model to assign higher logits to the distracting tokens: `Asia`, `yellow`, and `Promotion`, respectively. When normalised with softmax, logit increases translate into higher probabilities. Notably, the model typically assigns very low probabilities (10^{-5} to 10^{-3}) to these tokens, but with distracting context, their probabilities can increase by a factor of 10 to 100. Paired Student's (1908) t-tests confirm that these increases are statistically significant across all LMs, regardless of their size, family, or instruction-tuning status.

Finding 2: "Distracting" context prompts can be beneficial when relevant. While the probabilities and logits of the distracting tokens ( ) consistently increase, the direction of change for the correct token ( ) varies based on topic relevance. Except for the instruction-tuned `Llama-3.1-8B-Instruct` model, There is a small but statistically significant decrease in the correct answer token's logit when context information from irrelevant or random context prompts are provided.

While prior work typically groups "irrelevant context" into a single category and considers it detrimental, our findings suggest the need for a more nuanced classification. Although distracting context may not contain the exact correct answer, the implicit hints it provides can be beneficial, increasing the likelihood of the model generating the correct response. Figure 3 illustrates an example where "distracting" contextual information proves helpful, particularly in cases of question ambiguity. For instance, in the question `In Argentina, people speak the language of` from the *country language* relation in the LRE dataset, the term `language` can be interpreted metaphorically, leading to responses like `love` or `football`. However, the distracting context `In Russia, the primary language is Russian` can guide the model towards the correct answer.

Discussion: Contextual Entrainment — A Novel Mechanistic Phenomenon Thus, a new perspec-

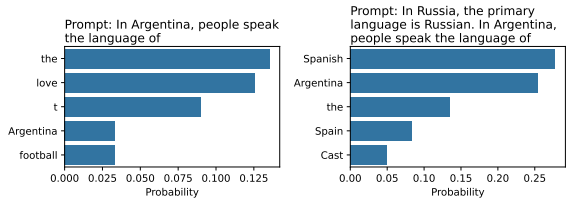


Figure 3: Providing "distracting" yet relevant context can be beneficial. For instance, when a question is ambiguous, such context may help clarify its intended meaning or guide interpretation. For example, "distracting" context provided, `Llama3.1-8B`'s top responses to the prompt `In Argentina, people speak the language of` shifted from `love` and `football` to the language-related tokens: `Spanish`, `Spain` and `Cast` (the first wordpiece of "Castilian Spanish").

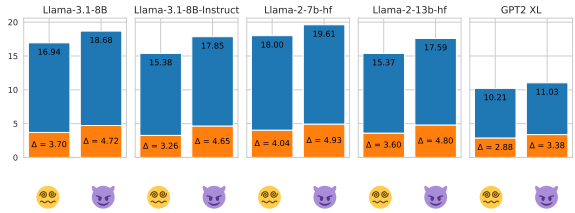


Figure 4: Counterfactual context prompts consistently cause greater distraction than factual context prompts.

tive emerges for analysing the phenomenon of distraction from a mechanistic angle, which we term *Contextual Entrainment*. Specifically, the model assigns a higher probability to tokens that appear within the context prompt. The fact that the model assigns higher probabilities to completely random tokens underscores the *mechanistic* nature of this phenomenon, as no linguistic or factual factors can plausibly account for the increase in logits and probabilities of the random tokens.

Our coinage of the term *Contextual Entrainment* does not imply any connection to human cognitive or psycholinguistic phenomena, such as brain entrainment (Poeppel and Assaneo, 2020; Pérez et al., 2022) or lexical entrainment (Garrod and Anderson, 1987; Brennan and Clark, 1996), nor does it suggest that LMs in any way replicate human brains or cognition. Rather, we use this term because *entrainment* most accurately describes the phenomenon we have observed. It refers solely to the output patterns exhibited by LMs in response to contextual input, without making any claims about the underlying cognitive mechanisms or their resemblance to human cognitive processes.

3.3 Counterfactual Experiment Results

Figure 4 shows the results with counterfactual context prompts. Using these counterfactual context prompts results in a significantly greater impact compared to previously identified factual context prompts. This suggests that, while we previously established that contextual entrainment is a "mechanistic" phenomenon, it is still subject to semantic factors in determining its magnitude of impact.

Finding 3: Counterfactual context prompts consistently cause greater distraction than factual context prompts. We present the model with two types of distracting context prompts: a factual (😊) one (`Japan is in Asia`) and a counterfactual (😈) one (`Japan is in Africa`). After the context prompt, we query the LM with a question (`Greece is located in` __) and observe how the context prompt changes the logits and probability of the 😈 (counterfactual token - `Africa`), the 😊 (distracting token - `Asia`), and the 🙂 (correct token - `Europe`). The 😊 setting, serving as a control group, is identical to the distraction prompt setting in Section 3.2.

The amount of distraction — in other words, the magnitude of contextual entrainment — is greater for counterfactual context prompts (😈) than factual prompts (😊). This is because the absolute logits of the 😊 token when 😊 prompts are provided are significantly lower than those of the 😈 token when 😈 prompts are provided (height of blue bars in Figure 4). Moreover, the extent of change is greater for counterfactual prompts. With a 😊 prompt, the 😊 token's logits increases smaller compared to no context, while a 😈 prompt causes a much larger increase, showing that counterfactual prompts create stronger distractions and greater shifts in the model's output (orange bar height in Figure 4).

Discussion: Contextual Entrainment — A mechanistic phenomenon affected by semantic factors. We have established that the presence of contextual entrainment is independent of semantic factors, given that it occurs even with random tokens. However, in this subsection, we also find that this "mechanistic" phenomenon is nevertheless modulated by semantic factors. In particular, counterfactual prompts induce a greater effect on contextual entrainment than factual context prompts.

The mechanism through which LMs utilise information from prompts is not yet fully understood. There is an ongoing debate regarding whether this capability arises from mere memorisation (Golchin et al., 2024) or from the implementation of an algorithm within the LMs weights and parameters during pre-training (Olsson et al., 2022; Lindner et al., 2023). Our findings suggest that this may not be a strict dichotomy; rather, it could be a compositional phenomenon in which both processes

operate concurrently.

Furthermore, the fact that counterfactual prompts can cause greater effects in contextual entrainment suggests that current models are more prone to distraction from counterfactual context prompts. This highlights the potential threat of dis- and misinformation.

4 Entrainment Heads

Recent research presents the argument that attention heads play the crucial role in controlling the LMs' utilisation of context (Wang et al., 2022; Meng et al., 2022; Jin et al., 2024; Crosbie and Shutova, 2024; Yu et al., 2024a, *inter alia*). Notably, Jin et al. (2024) studied a similar phenomenon, termed *knowledge conflicts*, which examines how models react when information from the context prompt contradicts the information acquired during pre-training. They use the terms *internal memory* and *external context* to refer to these two types of information. Furthermore, they identify two types of attention heads — memory heads and context heads — that correspond to the LM's utilisation of these distinct sources of information.

Knowledge conflict appears very similar to our counterfactual experiment; however, our research differs in several places. First and foremost, our prompt setting does not present a conflict. While we might both include a piece of counterfactual information in the context (e.g., `The capital city of Germany is Moscow`), we will ask the model to answer a question unrelated to either Germany or Moscow (e.g., we would query the model with `The capital of Nigeria is the city of` __); whereas knowledge conflict would query the model with a question directly related to Germany or Moscow (e.g., `The capital of Germany/Russia is the city of` __). Second, counterfactual experiments are part of our investigation into the contextual entrainment phenomenon, whereas researchers who study information conflict focus solely on scenarios where such conflicts arise. Third, while identifying this novel phenomenon of contextual entrainment, we seek to understand the process by which LMs utilise information from the context prompt. In contrast, studies on information conflict focus more on the applicational aspect, where the desideratum is to find a way to ensure that LLMs can effectively resolve conflicting information and generate outputs that align with the intended factuality or coherence of the given context. Nevertheless, there are several aspects in which our research can mutually inform and benefit from one another.

In particular, Jin et al. (2024) found that attention heads play a key role in utilizing contextual information. As we will show later in this section, our experimental results further confirm this finding. However, their method is limited to investigating each attention head in isolation and do not consider the interaction between attention heads. Moreover, an increasing number of studies (Niu et al., 2024; Yu et al., 2024b; Bhaskar et al., 2024) have identified issues with this individual approach, as it disregards the intricate structures of transformer LMs, and have advocated for a more holistic analysis of the entire computational "circuit." Inspired by this line of research, we adapt the differentiable masking based approach proposed by Yu et al. (2024b); Bhaskar et al. (2024) to identify the set of attention heads responsible for contextual entrainment. Our approach yields better results than the method proposed by Jin et al. (2024), suggesting that a circuit of attention heads — the "entrainment heads" — may have been formed in the model to process contextual information.

4.1 Entrainment Heads Discovery

Inspired by Yu et al. (2024b), we propose an automatic method to identify attention heads corresponding to contextual entrainment. We "turn off" specific heads by setting their contribution to the residual stream (Elhage et al., 2021)[3] to zero. To achieve this, we introduce a binary mask m_j for each head h_j, which selectively activates or deactivates heads ($\sum_{h_j \in H_i} m_j h_j(x_i)$). This mask is made differentiable by converting the sampled variable s_i from a Gumbel-sigmoid distribution using the straight-through estimator (Bengio et al., 2013):

$$s_i = \sigma\left(\frac{l_i - \log \frac{\log \mathcal{U}_1}{\log \mathcal{U}_2}}{\tau}\right); m_i = [\mathbb{1}_{s_i > \frac{1}{2}} - s_i]_{\text{detach}} + s_i, \quad (1)$$

where $\tau \in (0, \infty)$ is a temperature hyperparameter, l_i is a learnable logit of the sigmoid distribution $\sigma(\cdot)$, and $\mathcal{U}_1, \mathcal{U}_2 \sim \text{Uniform}(0, 1)$ are random variables drawn from a uniform distribution.

We can then apply gradient descent to a dataset to identify the optimal combinations of attention heads to disable in order to suppress contextual entrainment. Our objective is to determine the set of attention heads that contribute the most to contextual entrainment while minimising the number of heads used, using the following loss function:

[3] We briefly review residual stream in Appendix C.

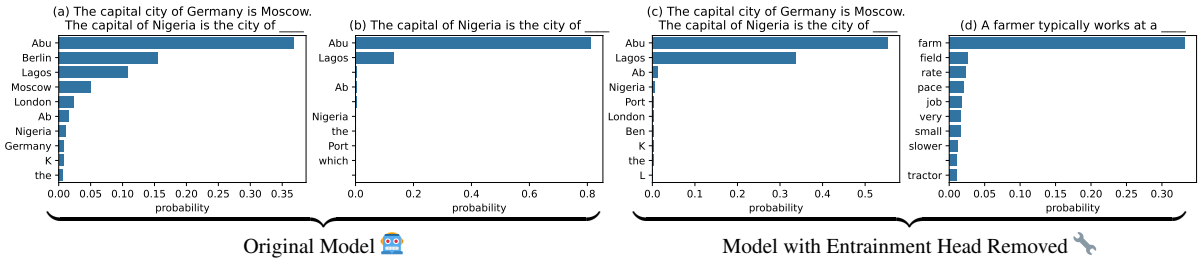


Figure 5: Effects of "Removing" the Entrainment Heads. The three figures show the top 10 token probabilities for their respective settings. (a,b) In the original model, when a piece of counterfactual information is presented, the model assigns higher probabilities to the distractions: `Berlin` and `Moscow`. (c) After setting the output of the identified entrainment heads to zero, however, the effect of contextual entrainment is drastically attenuated. (d) This operation of removing the entrainment heads has only a small impact on other capabilities; the model can still correctly answer questions in other domains.

$$\mathcal{L} = \underbrace{\ell(🙂) - \ell(😳)}_{\text{Logits } \Delta} + \underbrace{\lambda \cdot \frac{1}{|H|} \sum_{i=1}^{|m|} \sigma(l_i)}_{\text{Sparsity Loss}} . \qquad (2)$$

Experimental Setup We conduct our entrainment head experiments using `Llama-3.1-8B`, which has 1,024 attention heads (32 layers × 32 heads). Each LRE relation is split into training (80%), development (10%), and test (10%) sets. Entrainment heads are identified using the training set over 500 epochs,[4] selecting the epoch with the best effect and fewest heads.[5] All results are reported on the test set, which the model has neither seen nor used for hyperparameter search or checkpoint selection.

4.2 Experiment Results & Analysis

We first present our findings through the case study using the country–capital city relation. The remaining relations support the same findings, we will present them collectively at the end of this section.

"Turning off" the entrainment heads drastically reduce contextual entrainment. Our algorithm identified 36 entrainment heads for the country–capital city relation. When "turning these entrainment heads off," i.e., setting their output to zero, the model attenuates the effect of contextual entrainment, as illustrated in Figure 5. Normally, `Llama-3.1-8B` can be distracted by the counterfactual context `The capital city of Germany is Moscow`, confirming our previous findings in §3.3. However, after "turning off" the entrainment heads, the contextual entrainment effect is substantially attenuated. The rankings of the tokens `Berlin` and `Moscow` dropped from 2nd and 4th to 53rd and 68th, respectively. Additionally, the difference in logits between the correct token 🙂 (`Abu`)[6] and the distracting tokens 😳 (`Berlin` and `Moscow`) increased from 0.86 and 2.00 to 7.54 and 7.79, respectively. This attenuation is not a fluke. Table 2 shows that removing the entrainment heads significantly shifts the logits difference, probability difference, and the ranks of the 😳 tokens towards the values observed when no distracting context is provided, across the entire country–capital city relation test set.

Measure	🤖 No 😳	🤖 With 😳	🔧 No 😳	🔧 With 😳
ℓ(🙂)	19.51	20.68	19.49	21.21
ℓ(😳)	8.75	12.99	7.87	8.01
$\Delta = \ell$(😳) - ℓ(🙂)	10.76	7.69	11.62	13.20
Avg. 🙂 Token Rank	1.00	1.00	1.00	1.00
Avg. 😳 Token Rank*	1756.7	37.5	1707.3	1289.6

Table 2: Effects of "Removing" the Entrainment Heads across the Entire Country–Capital Relation City Test Set. Removing the entrainment heads caused a significant effect across logits delta and the ranks of the 😳 tokens, making them capitulate the situation when no distracting context is provided. *: $p < 6.9 \times 10^{-54}$ according to paired t-tests conditions between 🤖 and 🔧.

Table 4 demonstrates that our differentiable-masking-based entrainment head discovery method is applicable to other relations in the LRE dataset, highlighting the generalisability of our approach. We observe an increase in logit differences with the same scale to our country–capital city case study, further supporting the findings of our case study.[7]

Removing the entrainment heads has merely a small effect on other LM capabilities. While removing the entrainment heads significantly impacts contextual entrainment, it has only a negligible to small effect on other LM capabilities. First, we use other relations from the LRE dataset

[4]We use the AdamW optimiser (Loshchilov and Hutter, 2019) with $\lambda = 1.0$, $\tau = 1.0$, and a learning rate of 1.0.

[5]Specifically, we use the epoch with the maximum logit difference + number of heads × 0.1.

[6]The first wordpiece of *Abuja*, the capital city of Nigeria.

[7]The full result of every LRE relation will be publicly available online. *URL withdrawn for submission.*

Relation	Strict Acc. 🤖	Strict Acc. 🔧	Credulous Acc. 🤖	Credulous Acc. 🔧
company hq	83.5%	90.0%	88.0%	90.0%
country capital city	100.0%	100.0%	100.0%	100.0%
country currency	83.7%	100.0%	100.0%	100.0%
country language	85.7%	100.0%	100.0%	100.0%
country largest city	100.0%	100.0%	100.0%	100.0%
food from country	92.0%	98.5%	100.0%	98.5%
fruit inside color	77.0%	100.0%	98.0%	100.0%
fruit outside color	38.0%	84.0%	82.0%	84.0%
landmark in country	89.5%	91.0%	95.0%	91.0%
landmark on continent	88.5%	83.0%	97.0%	83.0%
product by company	95.0%	96.0%	98.0%	96.0%
star constellation name	84.7%	89.3%	92.3%	89.3%
task done by tool	78.0%	91.0%	93.5%	91.0%
task person type	78.5%	80.0%	80.0%	80.0%
work location	60.5%	75.0%	75.0%	75.0%
arithmetic 0-shot	100.0%	100.0%	100.0%	100.0%
spelling correction 1-shot	73.6%	72.0%	78.6%	76.8%
spelling correction 2-shot	94.6%	91.6%	97.0%	94.8%
spelling correction 5-shot	99.0%	98.4%	100.0%	100.0%
translation 1-shot	74.4%	73.0%	78.4%	76.8%
translation 2-shot	94.0%	93.0%	97.0%	96.2%
translation 5-shot	98.6%	97.2%	99.6%	99.4%

Table 3: Removing the entrainment heads of the country–capital city relation has a small to negligible effect on other LM capabilities. This table compares the strict (answer in top-3) and credulous (answer in top-10) accuracy of the original model (🤖) and the model with country–capital city entrainment heads removed (🔧). Removing these heads has a negligible effect on the LM's performance across other relations, with no obvious differences between 🤖 and 🔧.

Relation	# Heads	Density	ℓ(🙂) - ℓ(😊) 🤖	⇒	🔧
company hq	90	8.8%	3.94	⇒	14.68
country capital city	36	3.5%	7.69	⇒	13.20
country currency	42	4.1%	4.73	⇒	11.67
country language	30	2.9%	6.20	⇒	8.95
country largest city	33	3.2%	8.68	⇒	13.35
food from country	38	3.7%	3.98	⇒	9.95
fruit inside color	56	5.5%	0.97	⇒	11.16
fruit outside color	80	7.8%	2.14	⇒	13.82
landmark in country	59	5.8%	3.93	⇒	9.68
landmark on continent	52	5.1%	2.51	⇒	9.14
product by company	110	10.7%	3.62	⇒	16.47
star constellation name	72	7.0%	1.07	⇒	8.87
task done by tool	66	6.4%	4.70	⇒	12.31
task person type	41	4.0%	6.51	⇒	12.47
work location	68	6.6%	3.17	⇒	12.68

Table 4: Entrainment Head Identified across All LRE Relations. A small set of attention heads (3.2% to 10.7%) can substantially increase the gap between the logits of 🙂 and 😊 tokens, i.e., attenuate contextual entrainment. This may suggest that these heads play a crucial role in the prescence of contextual entrainment and can have broader implication in understanding how LMs utilise context information from prompts.

to evaluate whether the LM can still interpret the query and recall factual information, as well as perform ICL. In Table 3 shows the performance of the original model (🤖) and the modified model (🔧) on all other relations without distracting context, demonstrating that the model can still perform factual recall. We report both strict (the correct answer appears within the top-3 predicted tokens) and credulous (top-10) accuracy, as multiple correct answers may exist. Relying solely on whether the gold-standard token is the most probable leads to unstable results (Appendix D). Moreover, we experiment with the three ICL tasks identified by Brown et al. (2020): arithmetic, spelling correction, and translation (Appendix E). After removing the entrainment heads (🔧), the model exhibits only a small performance decrease (0.2∼3%) and continues demonstrate strong ICL capabilities with high accuracy in the same ballpark with 🤖.

This finding supports our hypothesis that this "circuit" of entrainment heads collectively corresponds to contextual entrainment rather than other capabilities and phenomena, such as factual recall, and is not strongly related to how LMs process and utilise contextual information or perform ICL more broadly. Thus, contextual entrainment and its connection to entrainment heads provides a novel perspective for understanding distraction. While current mitigation strategies (Yoran et al., 2023; Cuconasu et al., 2024; Wu et al., 2024) focus on methods external to the model — either modifying the context prompt or prompting the model to self-correct through reasoning — our findings suggest there could be a way to mitigate distraction by directly modifying or monitoring the internal mechanisms of LMs when performing RAG.

5 Conclusion

Llama see, llama do. We observe and confirm *contextual entrainment*, a novel phenomenon. If a token has appeared previously in the prompt, the model assigns a higher logit to that token, even for random tokens. Thus, a novel mechanistic effect may be at play in governing how LMs process and utilise information from the prompt—an effect that is analogous to but distinct from previously identified phenomena, such as the inductive pattern repetition effect observed by Olsson et al.'s (2022). However, we also discover that contextual entrainment is influenced by semantic factors. This finding highlights the potential threat of dis- and misinformation, which may be more severe than mere mistakes generated by the model. It also suggests that there may not be a strict dichotomy between mechanistic and statistical interpretations of LMs. Our identification of the entrainment heads suggests that interpretability techniques could provide crucial insights for real-world applications, such as the study of distraction.

6 Limitations

We did not conduct our experiments using larger LMs such as `Llama-3.1-70B` and `Llama-3.1-405B` due to resource limitations. However, we have used `Llama-3.1-8B`, `Llama-3.1-8B-Instruct`, and `Llama-2-13B-hf`, as these models are sufficiently large and powerful for our experiments. Moreover, we observed no differences in findings between larger and smaller models, such as `GPT-2 XL` and `Llama-2-13B-hf`, suggesting that our results are not significantly affected by model scale within this range. We encourage others to reproduce our work using larger models to further validate our findings.

Our experimental setup is rigorous. However, since RAG is most relevant to the problem of distraction, we conducted experiments using only the LRE dataset in this setting. We did not use standard RAG datasets (e.g., SimpleQA (Wei et al., 2024)), as they are difficult to control and compare fairly. Nonetheless, our experiment with random token inputs provides strong evidence — if such a setup yields successful results, then more structured approaches are unlikely to fail. Once again, we encourage others to reproduce our results using these datasets to further validate our findings.

Finally, while demonstrating two novel and insightful findings — the contextual entrainment phenomenon and the identification of entrainment heads — we do not propose an application to mitigate the distraction problem. We believe our findings serve as foundational steps toward addressing this issue. Given the depth of contributions presented in this work, we leave such applications for future research.

References

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *Preprint*, arXiv:1308.3432.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. 2024. Finding Transformer Circuits with Edge Pruning. *Preprint*, arXiv:2406.16778.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Joy Crosbie and Ekaterina Shutova. 2024. Induction Heads as an Essential Mechanism for Pattern Matching in In-context Learning. *Preprint*, arXiv:2407.07011.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Shahriar Golchin, Mihai Surdeanu, Steven Bethard, Eduardo Blanco, and Ellen Riloff. 2024. Memorization in In-Context Learning. *Preprint*, arXiv:2408.11546.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models.

In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

David Lindner, Janos Kramar, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. 2023. Tracr: Compiled transformers as a laboratory for interpretability. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR 2019*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In *The Twelfth International Conference on Learning Representations*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Alejandro Pérez, Matthew H. Davis, Robin A.A. Ince, Hanna Zhang, Zhanao Fu, Melanie Lamarca, Matthew A. Lambon Ralph, and Philip J. Monahan. 2022. Timing of brain entrainment to the speech envelope during speaking, listening and self-listening. *Cognition*, 224:105051.

David Poeppel and M. Florencia Assaneo. 2020. Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21(6):322–334.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, page 24.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31210–31227. PMLR.

Student. 1908. The Probable Error of a Mean. *Biometrika*, 6(1):1–25.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How Easily do Irrelevant Inputs Skew the Responses of Large Language Models? In *First Conference on Language Modeling*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *The Twelfth International Conference on Learning Representations*.

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024a. Mechanistic Understanding and Mitigation of Language Model Non-Factual Hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.

Lei Yu, Jingcheng Niu, Zining Zhu, and Gerald Penn. 2024b. Functional Faithfulness in the Wild: Circuit Discovery with Differentiable Computation Graph Pruning. *Preprint*, arXiv:2407.03779.

A LRE Dataset

We construct our experimental prompts using commonsense and factual data from the LRE dataset (Hernandez et al., 2024). This dataset comprises 47 relations with over 10,000 instances, spanning four categories: factual associations, commonsense knowledge, implicit biases, and linguistic knowledge. The dataset was created by filtering out ambiguous or noisy triples from PARAREL (Elazar et al., 2021), ensuring that each relation is well-defined and distinct. Additionally, the dataset was expanded by incorporating examples from structured knowledge sources, increasing coverage across various relation types.

However, during our experimentation, we identify several issues with certain relations. Some relations are too obscure for all of our chosen models to complete, resulting in highly noisy logits and probabilities for the correct tokens in some models. For example, the *superhero archnemesis* relation

Relation	# Samples	Context Templates	Query Templates
company hq	674	The headquarters of {} is in the city of Where are the headquarters of {}? It is in the city of	{} is headquartered in the city of The headquarters of {} are in the city of
country capital city	24	The capital of {} is the city of What is the capital of {}? It is the city of	The capital city of {} is The capital of {} is
country currency	30	What is the official currency of {}? It is called the {}'s official currency is called the The name of {}'s currency is the	The official currency of {} is the {}'s official currency is the
country language	24	{}, where most people speak In {}, people speak the language of People in {} speak the language of	People in {} speak The language used in {} is In {}, the primary language is
country largest city	24	What is the largest city in {}? It is the city of The largest city in {} is the city of	The largest city in {} is The biggest city in {} is
food from country	30	What is the country of origin for {}? It originates from {} originates from the country of	{} originates from {} is from the country of
fruit inside color	36	What color are {} on the inside? They are	On the inside, {} are
fruit outside color	30	What color are {} on the outside? They are the color of	On the outside, {} are
landmark in country	836	What country is {} in? It is in {} is in the country of	{} is in the country of
landmark on continent	947	What continent is {} on? It is on {} is on the continent of	{} is on the continent of
product by company	522	Which company developed {}? It was developed by	{} was created by {} is a product of
star constellation name	362	What is the name of the constellation that {} is part of? It is part of {} is part of the constellation named What is the name of the constellation that {} belongs to? It belongs to	{} is part of the constellation named
task person type	32	The task of {} would be best performed by someone with the role of a The professional role most suited to handle {} is a	{} is best suited for someone with the role of a
task done by tool	52	What tool is used for {}? Usually, you need a To accomplish {}, you need a tool called a	The tool used for {} is called a
work location	38	A {} typically works at a You can usually find a {} working in a	A {} typically works at a

Table 5: Selected LRE Relations.

Setting	ℓ(🙂)			ℓ(🧐)			P(🙂)			P(🧐)		
	No CTX	With CTX	Δ	No CTX	With CTX	Δ	No CTX	With CTX	Δ	No CTX	With CTX	Δ
Llama-3.1-8B												
Distraction	16.72	17.47	0.75	8.94	13.02	4.08	0.39	0.47	0.08	4.83e-03	0.03	0.02
Irrelevant	16.68	15.45	-1.23	3.05	7.96	4.92	0.38	0.35	-0.03	8.44e-05	2.51e-03	2.42e-03
Random	16.69	15.52	-1.17	3.19	7.19	4.01	0.38	0.34	-0.05	3.41e-04	3.53e-03	3.19e-03
Llama-3.1-8B-Instruct												
Distraction	15.75	17.63	1.88	8.64	12.07	3.43	0.27	0.43	0.16	5.13e-03	0.02	0.01
Irrelevant	15.74	15.84	0.11	2.19	6.78	4.58	0.27	0.34	0.07	1.31e-05	7.45e-04	7.32e-04
Random	15.74	14.78	-0.95	2.66	7.00	4.33	0.27	0.24	-0.03	3.54e-05	3.55e-03	3.52e-03
Llama-2-13b-hf												
Distraction	14.33	15.17	0.84	7.76	11.39	3.63	0.32	0.44	0.12	6.35e-03	0.04	0.04
Irrelevant	14.28	13.56	-0.72	3.14	6.98	3.84	0.32	0.31	-6.20e-03	2.53e-04	3.35e-03	3.10e-03
Random	14.29	13.19	-1.10	3.41	7.32	3.91	0.32	0.26	-0.06	3.58e-04	9.73e-03	9.37e-03
Llama-2-7b-hf												
Distraction	16.67	17.87	1.19	9.44	13.54	4.10	0.40	0.47	0.07	5.67e-03	0.04	0.03
Irrelevant	16.60	15.54	-1.06	5.06	8.11	3.05	0.40	0.36	-0.03	3.57e-04	3.74e-03	3.39e-03
Random	16.61	14.43	-2.18	5.50	7.39	1.89	0.40	0.32	-0.08	4.27e-04	5.53e-03	5.10e-03
GPT2 XL												
Distraction	9.17	10.13	0.96	4.18	7.41	3.23	0.18	0.26	0.08	7.76e-03	0.05	0.05
Irrelevant	9.16	9.06	-0.11	-1.55	3.04	4.59	0.18	0.17	-0.01	6.86e-05	5.28e-03	5.21e-03
Random	9.16	8.79	-0.37	-1.72	-0.28	1.44	0.18	0.16	-0.02	1.84e-04	2.31e-03	2.13e-03

Table 6: Average Logits and Probabilities acorss All LRE Relations and Prompt Settings.

includes instances such as ⟨*Martian Manhunter*, *Despero*, *superhero archnemesis*⟩, which may not be well-represented across models. Additionally, some relations in the linguistics and bias domains are not particularly relevant to the LM distraction setting. As a result, we select 15 relations, with their statistics listed in Table 5.

B Context Entrainment Experiment Supplementary Results

Table 6 shows the results across all relations to generate Figure 2. There is a small amount of variance for the logits of the correct token (ℓ(🙂) No CTX) because of we cap the amount of samples to 100,000 by random sampling, as described in Section 3.1.

Table 7 shows the breakdown for each LRE relation for the `Llama-3.1-8B` model.[8]

C Background: The Residual Stream

Here, we review the concept of the *residual stream* (Elhage et al., 2021), which provides the relevant background for our entrainment head discovery method. The basic idea is to view the computation of transformer LMs as maintaining a communication channel between model components via a shared vector space: the residual stream. Instead of operating independently, attention heads and MLP modules in each layer continuously pass and modify information through residual connections. Starting with the word embedding x_0, each layer (residual block) modifies it to become x_i, and the final result, x_{-1}, is converted into the output probability distribution by the unembed module. For each residual block at layer i, the output of the previous layer is x_{i-1}. Let H_i denote the set of all attention heads in the layer. The block's output, x_i, is then computed as described in Equation (3):

$$\begin{aligned} x_i^{\text{mid}} &= x_{i-1} + \sum_{h \in H_i} h(x_{i-1}); \\ x_i &= x_i^{\text{mid}} + \text{MLP}(x_i^{\text{mid}}). \end{aligned} \tag{3}$$

D Evaluation Method

We report both strict (the correct answer appears within the top-3 predicted tokens) and credulous (top-10) accuracy, as multiple correct answers may exist. Relying solely on whether the gold-standard token is the most probable leads to unstable results. For example, the LRE dataset includes the example `On the outside, apples are __`, where `red` is the

[8]Results from the rest of the models will be publicly available online. *URL withdrawn for submission*.

Relation	Setting	ℓ(🙂) No CTX	ℓ(🙂) With CTX	Δ	ℓ(🤓) No CTX	ℓ(🤓) With CTX	Δ	P(🙂) No CTX	P(🙂) With CTX	Δ	P(🤓) No CTX	P(🤓) With CTX	Δ
city in country	dstr	19.56	20.73	1.17	9.84	14.37	4.53	0.75	0.81	0.05	6.38e-03	0.03	0.02
	irr	19.49	19.53	0.04	5.54	9.80	4.27	0.75	0.77	0.03	5.49e-05	2.38e-03	2.32e-03
	random	19.52	16.96	-2.56	3.68	7.47	3.79	0.75	0.63	-0.12	2.33e-05	1.95e-03	1.93e-03
company hq	dstr	16.59	16.93	0.34	7.63	13.31	5.68	0.45	0.46	6.30e-03	1.50e-03	0.06	0.06
	irr	16.66	15.43	-1.23	4.34	9.64	5.30	0.46	0.36	-0.09	1.22e-04	6.17e-03	6.05e-03
	random	16.63	15.86	-0.77	2.74	5.59	2.85	0.45	0.41	-0.04	1.37e-04	3.26e-03	3.12e-03
country capital city	dstr	18.42	20.03	1.60	7.48	11.95	4.47	0.83	0.91	0.08	1.47e-04	1.12e-03	9.73e-04
	irr	18.43	17.33	-1.10	3.73	9.07	5.35	0.82	0.74	-0.08	2.18e-05	1.42e-03	1.40e-03
	random	18.44	17.29	-1.16	3.59	8.68	5.09	0.82	0.74	-0.09	3.28e-05	1.98e-03	1.95e-03
country currency	dstr	17.76	18.29	0.53	8.08	12.59	4.51	0.19	0.51	0.32	4.07e-04	7.58e-03	7.17e-03
	irr	17.72	15.25	-2.48	2.47	6.46	3.99	0.19	0.18	-0.01	3.72e-04	2.00e-03	1.63e-03
	random	17.73	16.41	-1.33	3.30	6.53	3.24	0.19	0.18	-8.41e-03	6.85e-05	1.20e-03	1.13e-03
country language	dstr	14.79	18.32	3.53	8.29	12.14	3.85	0.23	0.56	0.34	8.41e-03	0.01	2.97e-03
	irr	14.79	17.50	2.71	3.04	8.10	5.05	0.23	0.56	0.34	8.68e-05	4.16e-03	4.07e-03
	random	14.72	13.66	-1.07	3.26	8.64	5.38	0.22	0.18	-0.04	8.84e-04	0.01	0.01
country largest city	dstr	18.11	19.68	1.58	7.38	11.16	3.78	0.67	0.79	0.12	1.22e-04	8.22e-04	7.00e-04
	irr	18.13	17.15	-0.98	3.50	8.18	4.69	0.67	0.60	-0.08	1.58e-05	8.34e-04	8.19e-04
	random	18.13	17.53	-0.60	3.73	9.40	5.67	0.67	0.64	-0.04	1.29e-04	3.13e-03	3.00e-03
food from country	dstr	18.14	18.06	-0.08	8.99	12.73	3.73	0.54	0.55	6.08e-03	1.70e-03	0.01	0.01
	irr	18.12	16.25	-1.87	4.03	9.89	5.86	0.54	0.41	-0.13	3.28e-04	6.26e-03	5.93e-03
	random	18.10	16.56	-1.54	3.40	6.25	2.85	0.53	0.43	-0.10	1.09e-04	6.43e-04	5.34e-04
fruit inside color	dstr	15.78	15.57	-0.21	13.35	14.58	1.23	0.12	0.17	0.05	0.02	0.07	0.05
	irr	15.77	12.82	-2.95	1.54	6.28	4.74	0.12	0.10	-0.02	7.85e-07	1.63e-04	1.62e-04
	random	15.77	14.78	-0.99	3.02	6.86	3.84	0.12	0.14	0.01	1.06e-03	1.31e-03	2.47e-04
fruit outside color	dstr	14.57	14.53	-0.03	11.47	12.05	0.58	0.06	0.08	0.02	9.13e-03	0.01	2.67e-03
	irr	14.51	12.97	-1.55	3.40	8.52	5.12	0.06	0.05	-5.39e-03	4.31e-06	2.00e-03	1.99e-03
	random	14.52	12.95	-1.57	3.58	7.92	4.35	0.06	0.04	-0.02	1.77e-03	6.01e-03	4.24e-03
landmark in country	dstr	17.40	17.60	0.19	8.45	13.83	5.37	0.51	0.49	-0.02	1.23e-03	0.04	0.04
	irr	17.44	15.82	-1.62	3.97	9.36	5.40	0.52	0.45	-0.07	1.93e-04	6.79e-03	6.59e-03
	random	17.42	15.79	-1.62	3.30	6.55	3.25	0.51	0.43	-0.08	1.61e-04	8.17e-04	6.56e-04
landmark on continent	dstr	17.59	17.46	-0.13	11.90	15.64	3.74	0.36	0.28	-0.08	8.79e-03	0.07	0.06
	irr	17.06	15.59	-1.47	4.15	8.50	4.36	0.32	0.24	-0.09	2.39e-05	8.56e-04	8.32e-04
	random	17.07	16.16	-0.91	3.43	6.48	3.04	0.33	0.29	-0.04	1.68e-04	5.98e-04	4.30e-04
product by company	dstr	15.99	15.03	-0.96	6.83	11.13	4.31	0.49	0.47	-0.02	1.95e-03	0.03	0.03
	irr	16.03	14.29	-1.74	2.67	5.43	2.77	0.50	0.41	-0.08	7.22e-05	5.18e-04	4.45e-04
	random	16.03	15.29	-0.74	2.28	4.74	2.46	0.50	0.46	-0.05	5.94e-04	1.30e-03	7.09e-04
star constellation name	dstr	17.57	16.33	-1.24	11.83	15.43	3.60	0.28	0.26	-0.02	4.51e-03	0.04	0.03
	irr	17.56	14.94	-2.62	2.88	7.87	4.98	0.28	0.20	-0.07	2.74e-05	1.34e-03	1.31e-03
	random	17.58	16.48	-1.10	3.40	7.27	3.87	0.28	0.21	-0.07	1.91e-04	2.13e-03	1.93e-03
task done by tool	dstr	15.82	16.56	0.74	6.65	12.38	5.73	0.28	0.34	0.06	1.60e-03	0.03	0.03
	irr	15.82	13.67	-2.15	0.83	5.32	4.50	0.28	0.19	-0.09	5.52e-06	5.57e-04	5.51e-04
	random	15.85	14.72	-1.13	2.90	7.51	4.60	0.28	0.24	-0.04	4.21e-05	6.53e-03	6.49e-03
task person type	dstr	14.71	16.92	2.21	6.70	11.91	5.21	0.21	0.39	0.18	9.63e-04	0.01	0.01
	irr	14.70	14.32	-0.38	0.99	6.74	5.75	0.21	0.20	-6.79e-03	8.89e-06	1.70e-03	1.69e-03
	random	14.73	13.90	-0.84	2.87	8.11	5.24	0.21	0.17	-0.04	3.44e-05	7.84e-03	7.80e-03
work location	dstr	14.77	17.51	2.74	8.18	13.08	4.90	0.22	0.39	0.18	2.85e-03	0.04	0.03
	irr	14.70	14.34	-0.35	1.68	8.23	6.55	0.22	0.20	-0.02	1.31e-05	2.96e-03	2.94e-03
	random	14.74	14.01	-0.73	2.53	7.09	4.56	0.22	0.18	-0.04	4.47e-05	2.89e-03	2.85e-03

Table 7: Average Logits and Probabilities for each LRE Relation. LM: `Llama-3.1-8B`.

Relation	Exact Acc. (Top-1) 🤖	Exact Acc. (Top-1) 🔧	Strict Acc. (Top-3) 🤖	Strict Acc. (Top-3) 🔧	Credulous Acc. (Top-10) 🤖	Credulous Acc. (Top-10) 🔧
company hq	71.0%	63.5%	83.5%	90.0%	88.0%	90.0%
country capital city	94.0%	94.0%	100.0%	100.0%	100.0%	100.0%
country currency	18.0%	19.7%	83.7%	100.0%	100.0%	100.0%
country language	37.0%	60.3%	85.7%	100.0%	100.0%	100.0%
country largest city	97.0%	97.0%	100.0%	100.0%	100.0%	100.0%
food from country	78.0%	79.0%	92.0%	98.5%	100.0%	98.5%
fruit inside color	49.0%	50.0%	77.0%	100.0%	98.0%	100.0%
fruit outside color	0.0%	0.0%	38.0%	84.0%	82.0%	84.0%
landmark in country	72.0%	54.5%	89.5%	91.0%	95.0%	91.0%
landmark on continent	41.5%	38.5%	88.5%	83.0%	97.0%	83.0%
product by company	86.0%	81.0%	95.0%	96.0%	98.0%	96.0%
star constellation name	22.3%	26.0%	84.7%	89.3%	92.3%	89.3%
task done by tool	58.0%	58.5%	78.0%	91.0%	93.5%	91.0%
task person type	67.0%	53.5%	78.5%	80.0%	80.0%	80.0%
work location	49.0%	49.0%	60.5%	75.0%	75.0%	75.0%
arithmetic 0-shot	95.8%	97.9%	100.0%	100.0%	100.0%	100.0%
spelling correction 1-shot	58.2%	54.8%	73.6%	72.0%	78.6%	76.8%
spelling correction 2-shot	77.6%	74.0%	94.6%	91.6%	97.0%	94.8%
spelling correction 5-shot	87.4%	86.2%	99.0%	98.4%	100.0%	100.0%
translation 1-shot	58.4%	57.0%	74.4%	73.0%	78.4%	76.8%
translation 2-shot	74.8%	73.8%	94.0%	93.0%	97.0%	96.2%
translation 5-shot	85.4%	84.8%	98.6%	97.2%	99.6%	99.4%

Table 8: Removing the entrainment heads of the country–capital city relation has a small to negligible effect on other LM capabilities with exact (🙂 must be top-1 response) included. The exact accuracy metric is highly unstable and therefore lacks reference value.

only correct answer. In some cases, however, the LM selects `green` as the most probable next token, which is also a valid response.

Table 8 shows Table 3's results with the exact (🙂 must be top-1 response) metrics included. The exact accuracy metric is highly unstable due to the aforementioned reasons and therefore lacks reference value.

E ICL Tasks

We employ the three ICL tasks identified by Brown et al. (2020) to evaluate the effect of removing the entrainment heads on the LM's overall capability. In particular, Brown et al. (2020) proposed three tasks: arithmetic, spelling correction, and translation. Figure 6 shows Brown et al. (2020)'s illustration of these tasks. However, Brown et al. (2020) did not release the full dataset used for evaluation, so we recreate it.

Arithmetic We randomly sampled 1000 prompts from all possible two digit summations. E.g., `23 + 18 = 41`.

Spelling Correction We obtain the spelling correction data by prompting ChatGPT. Specifically,

(a) Prompt to Obtain the Spelling Correction Data

```
Give me 200 simple, random English words with
1 letter scrambled. For example:
gaot => goat
sakne => snake
brid => bird
fsih => fish
dcuk => duck
cmihp => chimp
organize the results in a Python list:
[('gaot', 'goat'), ('sakne', 'snake'), ...]
```

(b) Prompt to Obtain the Translation Data

```
Give me 200 simple, random English words with
their French translations. For example:
thanks => merci
hello => bonjour
mint => menthe
wall => mur
otter => loutre
bread => pain
Avoid using French accent marks or letters
that does exist in English
organize the results in a Python list:
[('thanks', 'merci'), ('hello', 'bonjour'),
...]
```

Table 9: Prompts Used to Obtain ICL Evaluation Data.

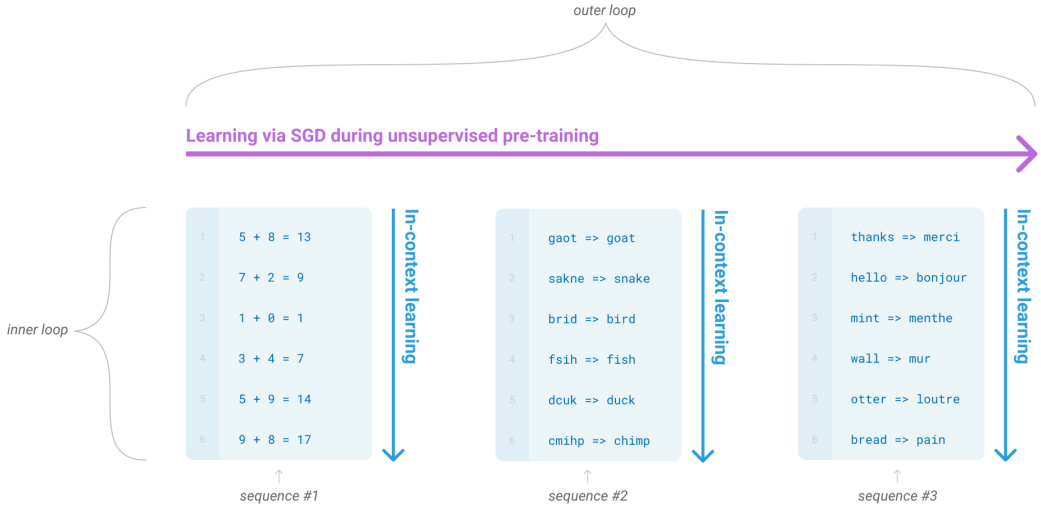


Figure 6: Brown et al.'s (2020) illustration of the three ICL tasks: arithmetic, spelling correction and translation.

we use **ChatGPT-o1**[9] with the prompt shown in Table 9(a). Table 10 lists the generated spelling correction pairs. We then randomly sample 1,000 instances under three different settings (1-shot, 2-shot, and 5-shot) from these 200 pairs.

Translation Similarly, for translation, we also sampled English to French word pairs by prompting **ChatGPT-o1**. The prompt used is shown in Table 9(a). Table 11 lists the generated spelling correction pairs. We then randomly sample 1,000 instances under three different settings (1-shot, 2-shot, and 5-shot) from these 200 pairs.

[9] https://openai.com/index/openai-o1-system-card/

gaot ⇒ goat, sakne ⇒ snake, brid ⇒ bird, fsih ⇒ fish, dcuk ⇒ duck, cmihp ⇒ chimp, hosre ⇒ horse, tiegr ⇒ tiger, zbera ⇒ zebra, muose ⇒ mouse, lino ⇒ lion, bera ⇒ bear, wlof ⇒ wolf, fxo ⇒ fox, dree ⇒ deer, frgo ⇒ frog, girafef ⇒ giraffe, doneky ⇒ donkey, bunyn ⇒ bunny, sehep ⇒ sheep, whela ⇒ whale, shrak ⇒ shark, eagel ⇒ eagle, corw ⇒ crow, sawn ⇒ swan, gosoe ⇒ goose, pengiun ⇒ penguin, ostirch ⇒ ostrich, moneky ⇒ monkey, kaola ⇒ koala, haed ⇒ head, hnad ⇒ hand, foto ⇒ foot, lge ⇒ leg, amr ⇒ arm, era ⇒ ear, yee ⇒ eye, lpi ⇒ lip, teo ⇒ toe, hiar ⇒ hair, aplpe ⇒ apple, baanna ⇒ banana, ornage ⇒ orange, maong ⇒ mango, garpe ⇒ grape, pecah ⇒ peach, pera ⇒ pear, plmu ⇒ plum, kwii ⇒ kiwi, leomn ⇒ lemon, carrto ⇒ carrot, pottao ⇒ potato, onoin ⇒ onion, tomtao ⇒ tomato, ltetuce ⇒ lettuce, rdaish ⇒ radish, spianch ⇒ spinach, cucubmer ⇒ cucumber, ppeper ⇒ pepper, celeyr ⇒ celery, rde ⇒ red, bleu ⇒ blue, geren ⇒ green, yellwo ⇒ yellow, puprle ⇒ purple, balck ⇒ black, wihte ⇒ white, pnik ⇒ pink, borwn ⇒ brown, gary ⇒ gray, franec ⇒ france, sapin ⇒ spain, chnia ⇒ china, inida ⇒ india, itlay ⇒ italy, jaapn ⇒ japan, caanda ⇒ canada, barzil ⇒ brazil, geramny ⇒ germany, russai ⇒ russia, teaxs ⇒ texas, manie ⇒ maine, oiho ⇒ ohio, iwoa ⇒ iowa, uath ⇒ utah, nevaad ⇒ nevada, aalska ⇒ alaska, haawii ⇒ hawaii, flordia ⇒ florida, gerogia ⇒ georgia, tabel ⇒ table, cahir ⇒ chair, phoen ⇒ phone, clcok ⇒ clock, wacth ⇒ watch, lihgt ⇒ light, doro ⇒ door, windwo ⇒ window, sopon ⇒ spoon, frok ⇒ fork, rnu ⇒ run, jmup ⇒ jump, wlak ⇒ walk, tlak ⇒ talk, raed ⇒ read, wrtie ⇒ write, eta ⇒ eat, slepe ⇒ sleep, drvie ⇒ drive, siwm ⇒ swim, bgi ⇒ big, smlal ⇒ small, fsat ⇒ fast, solw ⇒ slow, hto ⇒ hot, clod ⇒ cold, nwe ⇒ new, odl ⇒ old, hpapy ⇒ happy, sda ⇒ sad, cta ⇒ cat, dgo ⇒ dog, cpu ⇒ cup, pne ⇒ pen, snu ⇒ sun, mono ⇒ moon, satr ⇒ star, teer ⇒ tree, rokc ⇒ rock, blal ⇒ ball, oepn ⇒ open, clsoe ⇒ close, puhs ⇒ push, plul ⇒ pull, lfit ⇒ lift, dorp ⇒ drop, crary ⇒ carry, hodl ⇒ hold, thorw ⇒ throw, cacth ⇒ catch, doctro ⇒ doctor, laweyr ⇒ lawyer, teacehr ⇒ teacher, nusre ⇒ nurse, drievr ⇒ driver, artsit ⇒ artist, sinegr ⇒ singer, writre ⇒ writer, chfe ⇒ chef, pliot ⇒ pilot, freind ⇒ friend, enmey ⇒ enemy, hosue ⇒ house, hoem ⇒ home, famliy ⇒ family, moeny ⇒ money, waetr ⇒ water, frie ⇒ fire, earht ⇒ earth, widn ⇒ wind, yse ⇒ yes, on ⇒ no, pu ⇒ up, dwon ⇒ down, lfet ⇒ left, rigth ⇒ right, ni ⇒ in, otu ⇒ out, dya ⇒ day, ngiht ⇒ night, ciyt ⇒ city, tonw ⇒ town, roda ⇒ road, steret ⇒ street, sohp ⇒ shop, sotre ⇒ store, bnak ⇒ bank, csah ⇒ cash, hosiptal ⇒ hospital, clinci ⇒ clinic, questoin ⇒ question, anwser ⇒ answer, probelm ⇒ problem, solutoin ⇒ solution, loev ⇒ love, haet ⇒ hate, peaec ⇒ peace, wra ⇒ war, truht ⇒ truth, lei ⇒ lie, muisc ⇒ music, moive ⇒ movie, boko ⇒ book, paeg ⇒ page, paepr ⇒ paper, pecnil ⇒ pencil, deks ⇒ desk, soaf ⇒ sofa, pillwo ⇒ pillow, blnaket ⇒ blanket.

Table 10: 200 Spelling Correction Pairs Used for ICL Capability Evaluation.

thanks ⇒ merci, hello ⇒ bonjour, mint ⇒ menthe, wall ⇒ mur, otter ⇒ loutre, bread ⇒ pain, water ⇒ eau, friend ⇒ ami, love ⇒ amour, cat ⇒ chat, dog ⇒ chien, house ⇒ maison, horse ⇒ cheval, cow ⇒ vache, cheese ⇒ fromage, family ⇒ famille, black ⇒ noir, white ⇒ blanc, red ⇒ rouge, green ⇒ vert, blue ⇒ bleu, boy ⇒ garcon, girl ⇒ fille, night ⇒ nuit, day ⇒ jour, morning ⇒ matin, evening ⇒ soir, sun ⇒ soleil, moon ⇒ lune, star ⇒ etoile, sky ⇒ ciel, flower ⇒ fleur, car ⇒ voiture, city ⇒ ville, country ⇒ pays, beach ⇒ plage, forest ⇒ foret, river ⇒ riviere, mountain ⇒ montagne, desert ⇒ desert, island ⇒ ile, table ⇒ table, chair ⇒ chaise, window ⇒ fenetre, door ⇒ porte, book ⇒ livre, pen ⇒ stylo, pencil ⇒ crayon, letter ⇒ lettre, store ⇒ magasin, restaurant ⇒ restaurant, coffee ⇒ cafe, tea ⇒ the, juice ⇒ jus, milk ⇒ lait, egg ⇒ oeuf, butter ⇒ beurre, sugar ⇒ sucre, salt ⇒ sel, pepper ⇒ poivre, chicken ⇒ poulet, beef ⇒ boeuf, fish ⇒ poisson, bird ⇒ oiseau, snake ⇒ serpent, frog ⇒ grenouille, turtle ⇒ tortue, rabbit ⇒ lapin, pig ⇒ cochon, sheep ⇒ mouton, goat ⇒ chevre, fox ⇒ renard, wolf ⇒ loup, lion ⇒ lion, tiger ⇒ tigre, bear ⇒ ours, phone ⇒ telephone, computer ⇒ ordinateur, keyboard ⇒ clavier, screen ⇒ ecran, mouse ⇒ souris, camera ⇒ camera, photo ⇒ photo, movie ⇒ film, music ⇒ musique, song ⇒ chanson, dance ⇒ danse, poem ⇒ poeme, library ⇒ bibliotheque, museum ⇒ musee, school ⇒ ecole, university ⇒ universite, teacher ⇒ professeur, student ⇒ etudiant, office ⇒ bureau, job ⇒ travail, money ⇒ argent, bank ⇒ banque, street ⇒ rue, road ⇒ route, building ⇒ batiment, tall ⇒ grand, small ⇒ petit, short ⇒ court, big ⇒ gros, new ⇒ nouveau, old ⇒ vieux, happy ⇒ heureux, sad ⇒ triste, angry ⇒ fache, tired ⇒ fatigue, busy ⇒ occupe, free ⇒ libre, open ⇒ ouvert, closed ⇒ ferme, expensive ⇒ couteux, cheap ⇒ , yes ⇒ oui, no ⇒ non, maybe ⇒ , never ⇒ jamais, always ⇒ toujours, often ⇒ souvent, sometimes ⇒ parfois, rarely ⇒ rarement, early ⇒ tot, late ⇒ tard, now ⇒ maintenant, soon ⇒ bientot, yesterday ⇒ hier, today ⇒ aujourd), tomorrow ⇒ demain, hour ⇒ heure, minute ⇒ minute, second ⇒ seconde, time ⇒ temps, moment ⇒ moment, week ⇒ semaine, month ⇒ mois, year ⇒ annee, monday ⇒ lundi, tuesday ⇒ mardi, wednesday ⇒ mercredi, thursday ⇒ jeudi, friday ⇒ vendredi, saturday ⇒ samedi, sunday ⇒ dimanche, spring ⇒ printemps, summer ⇒ ete, autumn ⇒ automne, winter ⇒ hiver, police ⇒ police, fire ⇒ feu, help ⇒ aide, problem ⇒ probleme, question ⇒ question, answer ⇒ reponse, truth ⇒ verite, lie ⇒ mensonge, idea ⇒ idee, important ⇒ important, interesting ⇒ interessant, possible ⇒ possible, impossible ⇒ impossible, difficult ⇒ difficile, easy ⇒ facile, strong ⇒ fort, weak ⇒ faible, light ⇒ lumiere, dark ⇒ sombre, direction ⇒ direction, left ⇒ gauche, right ⇒ droite, straight ⇒ tout), back ⇒ arriere, up ⇒ haut, down ⇒ bas, in ⇒ dans, out ⇒ dehors, on ⇒ sur, under ⇒ sous, behind ⇒ derriere, next ⇒ prochain, near ⇒ pres, far ⇒ loin, between ⇒ entre, each ⇒ chacun, all ⇒ tous, some ⇒ quelques, none ⇒ aucun, every ⇒ chaque, anyway ⇒ de toute facon example ⇒ exemple, reason ⇒ raison, mistake ⇒ erreur, gift ⇒ cadeau, party ⇒ fete, plan ⇒ plan, goal ⇒ objectif, success ⇒ succes.

Table 11: 200 Translation Pairs (English ⇒ French) Used for ICL Capability Evaluation.