Towards Robust Knowledge Representations in Multilingual LLMs for `Equivalence` and `Inheritance` based Consistent Reasoning

Gaurav Arora
Amazon
gaurvar@amazon.com

Srujana Merugu
Amazon
smerugu@amazon.com

Shreya Jain
IIT Jammu*
2020uee0135@iitjammu.ac.in

Vaibhav Saxena
Amazon
saxenvai@amazon.com

Abstract

Reasoning and linguistic skills form the cornerstone of human intelligence, facilitating problem-solving and decision-making. Recent advances in Large Language Models (LLMs) have led to impressive linguistic capabilities and emergent reasoning behaviors, fueling widespread adoption across application domains. However, LLMs still struggle with complex reasoning tasks, highlighting their systemic limitations. In this work, we focus on evaluating whether LLMs have the requisite representations to reason using two foundational relationships: "equivalence" and "inheritance". We introduce novel tasks and benchmarks spanning six languages and observe that current SOTA LLMs often produce conflicting answers to the same questions across languages in 17.3-57.5% of cases and violate inheritance constraints in up to 37.2% cases. To enhance consistency across languages, we propose novel "Compositional Representations" where tokens are represented as composition of equivalent tokens across languages, with resulting conflict reduction (up to -4.7%) indicating benefits of shared LLM representations.

1 Introduction

Reasoning is the capacity to employ logic and analyze relationships among entities to extrapolate from known evidence to derive new insights. Language significantly bolsters this process by supplying the necessary structure and vocabulary for encoding complex ideas, thus facilitating hypothesis generation and evaluation. The intricate connection between linguistic and reasoning capabilities is a hallmark of human intelligence, enabling abstract thinking, problem-solving, and decision-making.

Recent advancements in LLMs such as ChatGPT (OpenAI, 2022) and Claude (Anthropic, 2023c) showcase their exceptional language generation capabilities and their potential to boost performance

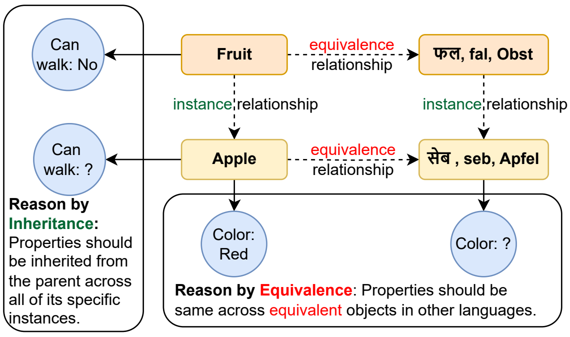


Figure 1: An example of Reasoning by "Equivalence" and Reasoning by "Inheritance" based on *existence* of equivalence/inheritance relationship between concepts.

across diverse Natural Language Processing (NLP) tasks (Ahuja et al., 2023), with multiple studies also pointing to emergent reasoning abilities at scale (Wei et al., 2022). However, LLMs continue to face challenges with complex reasoning tasks such as planning and problem-solving, indicating that their expansive modeling capacity and extensive training regime might enable them to mask deeper systemic shortcomings through superficial reasoning. As LLMs increasingly permeate applications catering to multilingual users with complex needs, gaining a deeper understanding of their functioning becomes imperative, as it could uncover systemic gaps and pave the way for superior models.

Robust reasoning hinges on the availability of powerful constructs such as entity-relation (ER) graphs and rules to interpret relationships. For instance, an ER graph with entities A, B, and C, where "A is the father of B" and "C is the wife of A" allows us to deduce that C is likely the mother of B, based on interpretation of the relationships "father" and "wife". While there are myriad relationships underpinning reasoning such as “cause and effect“ and ”comparison“, in our current work, we focus on "equivalence" and "inheritance" due to their predominance in enhancing the efficiency of

[*] Contributed to this work during her internship at Amazon

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7744–7762
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

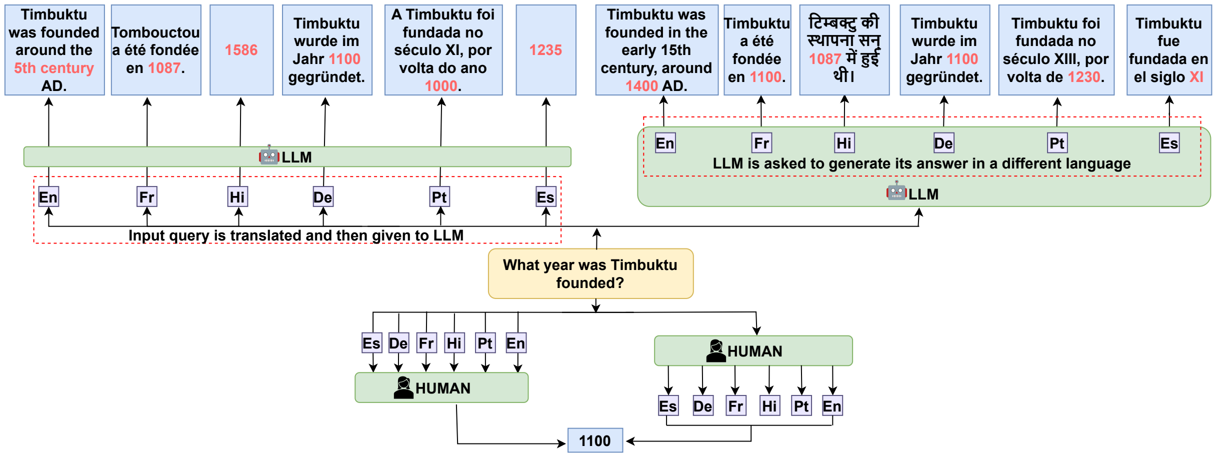


Figure 2: An example of LLM (Claude-v1 Instant) lacking "equivalence relationship between equivalent concepts across languages" due to tight coupling of knowledge representation and language expression unlike in Humans.

logical inference through property transfer, which is also reflected in their adoption as core constructs of knowledge representation and programming languages (Minsky et al., 1974). Fig 1 illustrates these relationships showing how humans create the necessary representations of "Apple" to reason across equivalent objects ("seb" : "Apple" in Hindi (transliterated), "Apfel" : "Apple" in German, etc) independent of the language/script of expression (Reasoning by Equivalence), and inherit properties from the abstract concept "Fruit" across all of its specific instances ("Apple", "Orange" etc) (Reasoning by Inheritance). Here, we expect representations of equivalent objects to be similar, while that of inherited objects satisfy transitivity. These representations are crucial for efficient learning, knowledge sharing, and updation of beliefs.

Typically, humans create abstractions using denotational semantics, i.e., a word's meaning is defined by objects it describes, which is the favored approach in logical theory. In contrast, LLMs use distributional semantics, i.e., a word's meaning stems from the training data context, which can be problematic when the data has gaps, such as infrequent connections between equivalent words across languages. Hence, despite impressive performance on NLP tasks(Ahuja et al., 2023), it is unclear if LLMs create the necessary representations within and across languages to support consistent reasoning across equivalent and inherited objects.

Contributions. In this work, we focus on whether LLMs have the requisite representations to reason by equivalence and inheritance across languages and make the below contributions.

1. We introduce a novel task and parallel benchmark datasets of factoid QA to evaluate "Reasoning by Equivalence" in LLMs and assess the performance of multiple SOTA LLMs on this task across 6 languages (English, French, Spanish, German, Portuguese and Hindi). On our benchmarks, LLMs generate conflicting answers across languages in 17.3-57.5% of cases indicating a significant gap. We also perform a controlled experiment to identify factors promoting consistency across languages and find a strong positive correlation with similarity in script and typology.

2. We present another task and associated new benchmark to evaluate "Reasoning by Inheritance" in LLMs and study the proficiency of multiple SOTA LLMs across six languages. Our results indicate that LLM answers violate inheritance constraints in up to 37.2% cases across these languages with most violations observed in Hindi.

3. We also propose a novel method for constructing "Compositional Representations" in LLMs by representing tokens as composition of other equivalent tokens in vocabulary, which grants the model access to (otherwise) distant representations of equivalent objects across languages, thereby facilitating improved knowledge sharing and reduction in conflicts with gains up to 4.7% compared to baselines.

To the best of our knowledge, this is the first quantitative study of LLM reasoning via equivalence and inheritance across languages. We will share the benchmarks as a community resource and to ensure reproducibility after organization approval. Note that even when the desired equivalence and inheritance relationships hold and properties transfer correctly (our current focus), there may be gaps in LLM's multi-step reasoning process due to other factors as we discuss in detail in Appendix A.1.

2 Related Work

Multilingual NLP. LLMs like ChatGPT (OpenAI, 2022), GPT-4 (OpenAI et al., 2024), Claude (Anthropic, 2023c), BLOOMZ (Muennighoff et al., 2023), XGLM (Lin et al., 2022) have shown impressive performance on standard multilingual NLP tasks and benchmarks (Ahuja et al., 2023; Zhao et al., 2023; Enis and Hopkins, 2024; Ahuja et al., 2024). Despite extensive evaluations and the existence of parallel multilingual datasets such as MLQA and XQUAD (Ahuja et al., 2024), to the best of our knowledge, there is no prior work or tailored benchmarks for assessing LLMs' ability to reason by equivalence and inheritance across multiple languages. Further, there is a chance of public benchmarks with duplicated knowledge across languages being included in LLM training data, rendering reasoning related assessments unreliable. Our study is the first to create controlled benchmarks and evaluate LLMs on these reasoning tasks to identify gaps and potential contributing factors.

Reasoning in LLMs. Reasoning abilities of LLMs have been studied for problem solving, decision making, and critical thinking (Huang and Chang, 2023; Wei et al., 2022; Bubeck et al., 2023). Prior work has also looked at evaluating ability of encoder only models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) to understand ontological knowledge (Wu et al., 2023). In this work, we focus on reasoning based on two foundational relationships: equivalence and inheritance and evaluate popular LLMs on these dimensions across multiple languages. In recent years, there have been advances (Aspis et al., 2022; Lazzari et al., 2024; Marconato et al., 2023) in neuro-symbolic architectures that combine symbolic and sub-symbolic components to enable efficient computation of symbolic representations and deductive reasoning. However, these works do not present a detailed analysis of representations of equivalent or related entities and these methods also entail much higher computational costs, limiting their adoption.

Representation Learning in NLP. Improving distributed representations led to significant performance improvements in past (Liu et al., 2021; Devlin et al., 2018; Mikolov et al., 2013). Prior work on adapting attention mechanisms to bridge gaps across disparate but related inputs such as translated/transliterated data has led to improved multilingual representations (Conneau et al., 2020; Khanuja et al., 2021; Arora et al., 2023). In our current work, we bridge the gap between distant representation spaces of various languages by adapting the attention mechanism to better utilize the token-language mapping.

3 Reasoning by Equivalence

Reasoning by equivalence is a core building block that enables efficient and scalable reasoning across contexts, with the efficiency being determined by the size of the equivalence classes (sets of equivalent objects). Construction of these "equivalence classes", i.e., "abstract concepts" from specific contexts and reusing these abstract concepts flexibly beyond the specific contexts (An et al., 2023; Mitchell, 2021; Kumar et al., 2023; Giunchiglia and Walsh, 1992; Hull, 1920), is a natural human skill. The human ability to acquire knowledge from one language (e.g., "apple is red") and construct representations shared across languages as in Fig 1 is a prime example. Similar to multilingual humans, LLMs also see large amount of multilingual data during pre-training (Blevins and Zettlemoyer, 2022). For instance, pretraining data of GPT-3 and BLOOM spanned 119 and 46 languages (Brown et al., 2020; Scao et al., 2023) respectively. In this section, we evaluate if SOTA LLMs also have the ability to reason by equivalence across languages given their impressive multilingual capabilities, and if this ability is due to shared representations (i.e. existence of equivalence relationship) or duplication of knowledge across languages. Note that the existence of "equivalence relationships encoded in LLM representations" is a fundamental prerequisite for complex reasoning even though it does not guarantee that LLMs can effectively leverage it for multi-step logical reasoning due to gaps in LLM's inference mechanism based on associative attention (see Appendix A.1).

3.1 How good are LLMs at exhibiting "Reasoning by Equivalence"?

We evaluate if LLMs exhibit "Reasoning by Equivalence" specifically across languages by estimating how dependent LLM's answers are on the language of input/output expression. A high dependency indicates tight coupling of the knowledge representation with the language and points to lack of shared abstractions and limited ability to reason

	En-Fr	En-Es	En-De	En-Pt	En-Hi	Avg
Claude v1 Instant	31.41	32.66	30.81	31.99	56.31	36.64
Claude v2	22.20	22.54	21.75	22.89	47.02	27.28
Claude v3 Sonnet	**20.99**	**21.23**	**20.60**	**21.49**	**43.51**	**25.56**
BLOOMZ-7B	37.23	34.67	50.88	33.95	52.69	41.88
XGLM-7.5B	42.99	42.49	38.74	41.74	57.53	44.70

Table 1: Conflict rate between LLM answers to parallel En-XX questions, XX = [Fr, Es, De, Pt, Hi].

	En-Fr	En-Es	En-De	En-Pt	En-Hi	Avg
Claude v1 Instant	29.91	30.44	29.72	30.28	42.92	32.65
Claude v2	20.45	21.13	20.52	21.38	**28.14**	22.32
Claude v3 Sonnet	**17.89**	**18.06**	**17.34**	**17.99**	28.61	**19.97**

Table 2: Conflict rate between LLM responses in [En, Fr, Es, De, Pt, Hi] for input questions in En.

by equivalence with knowledge duplication. Fig 2 shows an example where Claude-v1 instant's answers are significantly dependent on the language of input/output expression. Below, we outline our methodology, dataset, metrics and results.

3.1.1 Methodology

Let $R = \{r_1, \cdots r_L\}$ be a set of languages with $[X^{r_1} X^{r_2}, ... X^{r_L}]$ denoting parallel questions across $r_i \in R$. We perform two assessments.

Dependency on **I**nput **L**anguage **(DIL)**. Given L parallel questions $[X^{r_1}, X^{r_2}, ... X^{r_L}]$; we generate LLM answers[1] Y^{r_i} for each X^{r_i} independently. Choosing an anchor language $r_{\text{anchor}} \in R$, we check for each $r \in R \setminus \{r_{\text{anchor}}\}$ if $Y^{r_{\text{anchor}}}$ and Y^r conflict with each other .

Dependency on **O**utput **L**anguage **(DOL)**. Given r_{anchor}, we input $X^{r_{\text{anchor}}}$ into the LLM and generate answers Y^r for $X^{r_{\text{anchor}}}$ for all $r \in R$ independently. Then, we check for each $r \in R \setminus \{r_{\text{anchor}}\}$ if $Y^{r_{\text{anchor}}}$ and Y^r conflict with each other.

Here, two answers are called conflicting if they contain contradictory information and not merely if there are different or one of them is non-informative. For our experiments, we consider English (En), French (Fr), Spanish (Es), German (De), Portuguese (Pt) and Hindi (Hi) languages, i.e. $R = \{\text{En}, \text{Fr}, \text{Es}, \text{De}, \text{Pt}, \text{Hi}\}$ and $r_{\text{anchor}} = \text{En}$. Since authors in (Lin et al., 2022; Ahuja et al., 2023) show that English instructions in the prompt perform better than instructions written in the native language for non-English languages, we tune the English instructions in prompt separately for each LLM and then keep these consistent for that LLM across all DIL and DOL experiments.

[1]Temperature=0 across the paper for deterministic outputs.

3.1.2 Dataset and Metrics

Dataset. To ensure feasibility of automated evaluation via LLM-based judges and reduce variations due to subjective interpretation and cultural variations, our evaluation focused primarily on objective factual/attribute-based questions on entities. We prepare En factual questions dataset consisting of 88,334 questions on well known named entities and translate the dataset to Fr, Es, De, Pt and Hi using AWS Translate (AWS, 2017b). See Appendix A.5.1 for more details on the dataset and Fig 15 for a few sample questions.

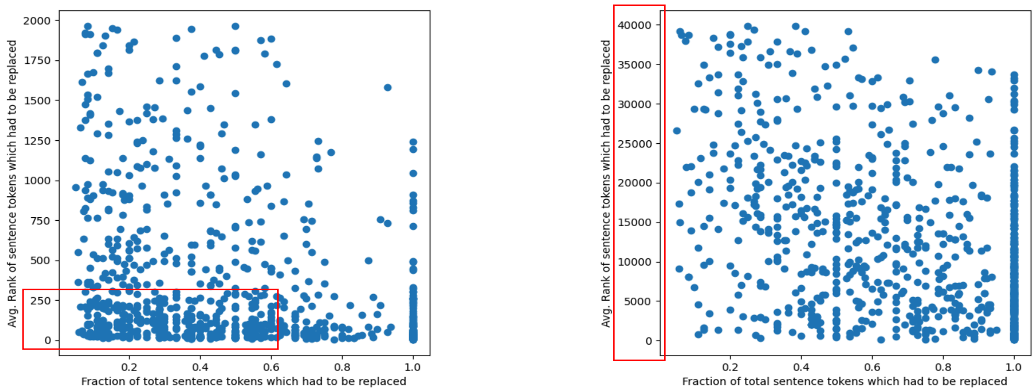
Metrics. We compute the conflicts among answers generated by LLM for different input/output expression languages.[2] We define conflict rate between (r_i, r_j) language pair as fraction of total answer pairs which are conflicting, i.e. $\text{ConflictRate}(r_i, r_j) = \frac{\Sigma_{k=1}^{|D|} \text{J}(Y_k^{r_i}, Y_k^{r_j})}{|D|}$, where $|D|$ is dataset size and J returns 1 if $(Y_k^{r_i}, Y_k^{r_j})$ are conflicting, else returns 0. We use Claude v3 Sonnet as the judge J with prompt shown in Fig 13 in Appendix A.4. Table 5 in Appendix A.4 shows that the average precision of our judge is >95%.

3.1.3 Analysis and Results

Table 1 shows conflict rate for various LLMs for DIL task.[3] We can see that conflict rate reduces with increase in model strength. Open source models lag behind closed source models by a significant margin with 25-44% average conflict rate across

[2]Since we care about consistent and common knowledge representation in LLMs for equivalent concepts, we only assess conflicting LLM responses to equivalent questions and not worry about factual accuracy of responses.

[3]GPT-3.5 had similar results as Claude v3 Sonnet but we could not add those results due to organization policy.



(a) German

(b) Hindi

Figure 3: Avg. Rank and Fraction of tokens which had to be replaced in German and Hindi with parallel English token to achieve consistent answer as English.

different LLMs and languages. Since we establish from results in Table 1 that LLMs are highly dependent upon input expression language, and that this dependency is consistent across varied open-source and closed-source LLMs of varied size, we evaluate only Claude family models for DOL task as shown in Table 2. We observe similar trends as average conflict rate of 19-32%. Both these results show that knowledge representation is tightly coupled with expression/language in LLMs, indicating a lack of right abstractions and limited knowledge sharing across languages in LLMs. Fig 16 and Fig 17 in Appendix A.5.2 show sample conflicting answers from various LLMs to equivalent questions from DIL and DOL tasks respectively.

3.2 Factors affecting LLMs ability to exhibit "Reasoning by Equivalence"

To better understand knowledge transfer and source of conflicts across languages in LLMs, we perform a controlled experiment wherein we create synthetic QnA data with non-existent named entities that LLM does not have any prior knowledge on and train it on synthetic data in one language and test for its transfer in other languages.

3.2.1 Controlled Experiment

Dataset. We create synthetic data of non-existent named entities and hallucinated articles about those entities using Claude v1-instant. We also generate factual questions about synthetic named entities which can be answered only from hallucinated articles. We only keep those questions for which Claude's answers with and without the article conflict with each other to ensure any LLM is unlikely to have any prior knowledge of our synthetic data. Our synthetic QnA dataset has 2063 synthetic named entities and has 32016 QnA pairs. As we perform the controlled experiment with En, De, Hi and HiEn (transliterated Hindi), and our synthetic question set is in En, we also translate it to De and Hi using AWS Translate, and transliterate to HiEn using IndicTrans (Bhat et al., 2015).

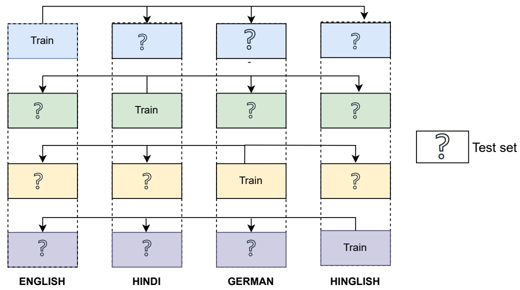


Figure 4: Setup for the controlled experiment. We train on unique 25% of the data for each language and test on its parallel data in other three languages.

Experiment Setup. We train XGLM-4.5B on unique 25% of the synthetic data for each language and test on its parallel data in other three languages. Specifically, we train on concatenated $D_1^{\text{en}}, D_2^{\text{hi}}, D_3^{\text{de}}, D_4^{\text{hien}}$ data where D_s^r denotes synthetic data in r^{th} language from s^{th} quarter of D^r. For $s = 1$, En is the anchor language and we evaluate knowledge transfer to $\text{Hi}, \text{De}, \text{HiEn}$ by looking at conflicts between LLM answer to En question from D_1^{en} and LLM answer to parallel $\text{Hi}, \text{De}, \text{HiEn}$ questions from $D_1^{\text{hi}}, D_1^{\text{de}}, D_1^{\text{hien}}$ respectively. As shown in Fig 4, same procedure is followed for $s \in \{2, 3, 4\}$ where $\text{Hi}, \text{De}, \text{HiEn}$ are anchor languages respectively.

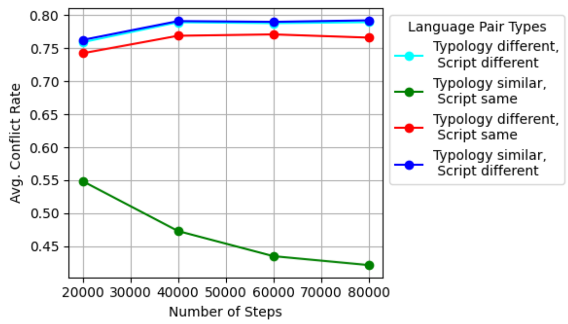


Figure 5: Conflict rate for different language pair types.

3.2.2 Results and Analysis

Fig 5 shows there is significantly higher knowledge transfer between languages with similar typology and same script, as compared to the pairs where either typology or script is different (see Fig 18 in Appendix A.5.3 for conflict rate of all language pairs individually). For instance, conflict rate for En-De is much lower than that of En-HiEn, which in turn is lower than that of En-Hi.

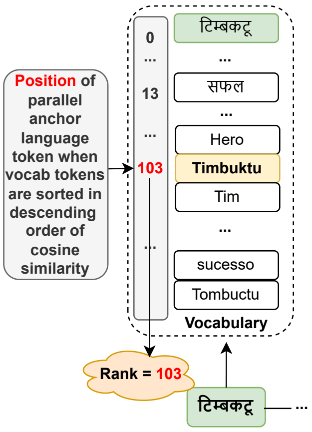


Figure 6: Illustration of Rank.

We define **Rank** of a token x in a non-anchor question as the *position* at which its parallel token from anchor question occurs if we sort all tokens in the vocabulary by cosine similarity with x in descending order. Rank of a token captures the relative proximity to the anchor language's parallel token in the embedding space. To assess what it would take to get a non-conflicting answer, we replace tokens in the non-anchor question with their parallel anchor question tokens in the descending order of rank, i.e., farthest (non-anchor,anchor) parallel tokens are replaced first, till we obtain a non-conflicting answer. Fig 3a shows dense region of De tokens with small ranks which had to be replaced by their parallel anchor question tokens to reach a non-conflicting consistent answer. This shows that having close enough representations for equivalent tokens in different languages might also not be enough, and they have to be same for LLM to learn consistent knowledge. This limitation stems from the representation space LLMs operate in since we project discrete symbols in the continuous embedding space. On the other hand, for Hi in Fig 3b, average rank of replaced tokens is significantly higher relative to De, likely due to the typology and script differences in case of En – Hi, which points to the need for near similar representations to share knowledge.

These results show that consistency of LLMs for distant languages is likely to stem from duplication of knowledge across languages in the training data, whereas for languages with similar typology and script, knowledge sharing occurs due to similar representation of equivalent tokens and information propagation. The effect of "duplicate knowledge in LLM training" is also reflected in conflict rate numbers of Tables 1 and 2 which are lower than those in Table 4. Results in Tables 1 and 2 are based on factual questions on well-known real entities, while the Table 4 results are from the controlled experiment with questions about synthetically created non-existent named entities. For the real entities in Tables 1 and 2, LLMs may have been pre-trained on duplicate information about the same entity expressed in multiple languages leading to consistent responses even when the LLM representations of the corresponding entities are significantly different (i.e., no equivalence relationship). However, for the synthetic entities in Table 4, LLMs are unlikely to have seen duplicate knowledge across languages during pre-training and has to learn about them (in one language) during the controlled training process resulting in significantly higher "conflict rates". See Appendix A.2 for more discussion on the effect of duplication of knowledge across languages in LLM's training data.

4 Reasoning by Inheritance

Reasoning by Inheritance is also a key building block of common sense and logical reasoning in humans enabled by concept abstractions. Humans identify common patterns amongst instances of the same type and create abstract concepts to reason consistently across all specific instances by inheriting properties from the abstract concept as shown in Fig 1. In this section, we investigate if SOTA LLMs can use their ontological knowledge and inherit properties from abstract concepts consistently across various specific instances of the abstract concept within multiple languages.

We evaluate if an LLM exhibits "Reasoning by

	En	Fr	Es	De	Pt	Hi	Avg
Claude v1 Instant	3.36	18.33	17.16	9.12	15.05	36.5	16.59
Claude v2	3.37	4.55	11.69	4.96	6.78	18.53	8.31
Claude v3 Sonnet	**0.13**	**4.43**	**7.27**	**1.86**	**5.41**	**14.69**	**5.63**
BLOOMZ-7B	4.78	8.76	7.49	8.08	10.4	23.27	10.46
XGLM-7.5B	36.61	30.87	35.65	36.5	37.27	33.59	35.08

Table 3: Conflict rate of LLM answers on inheritance-based questions on common concepts in En, Fr, Es, De, Pt, Hi.

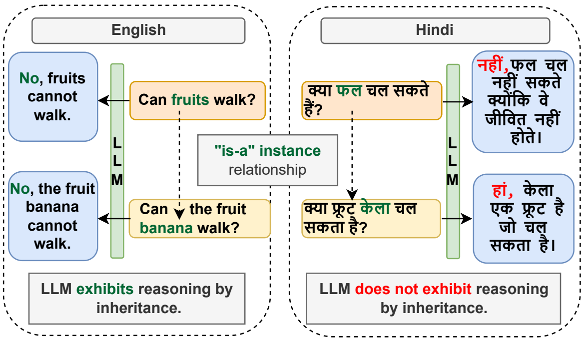


Figure 7: An example of LLM (Claude v1 Instant) exhibiting reasoning by inheritance in En but not in Hi.

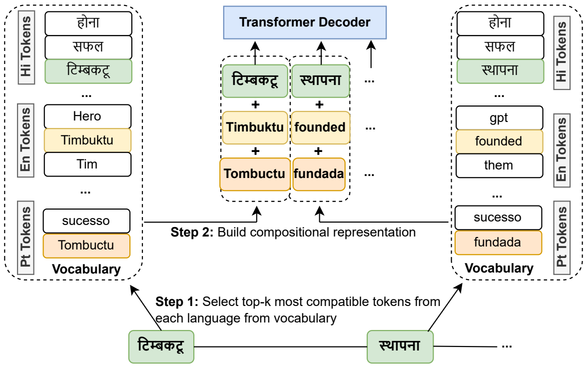


Figure 8: CoRe Illustration: LLM can access distant equivalent representations to permit knowledge sharing.

Inheritance" within a language by checking if specific instances of an abstract parent concept inherit properties of the parent without conflicts. Fig 7 shows an example wherein Claude-v1 instant does not exhibit reasoning by inheritance in Hi due to the lack of the right abstractions.

Dataset. We prepare a set of 35 abstract concepts and 2396 well known named entities which are specific instances of those abstract concepts. For each one of the abstract concept, we hand-curate set of properties that all of its specific instances should inherit. We create templatized questions from those properties to prepare En dataset and translate it to Fr, Es, De, Pt and Hi using AWS Translate (AWS, 2017b). Fig 19 in Appendix A.6.1 shows a sample of our dataset.

Methodology. We evaluate the consistency of LLMs in inheriting and applying ontological knowledge consistently across specific instances of abstract concepts by directly asking in native language if the specific instance has the property of abstract concept and checking if the answer for abstract concept and specific instance are conflicting.

Metrics. We compute conflict rate as in the Section 3 but focus on comparing LLM answers on an inheritable property for an abstract parent concept and that of its specific instances.

Results. Table 3 shows conflict rate for questions requiring reasoning by inheritance across six languages.[3] We notice that conflict rate is low for En for most models but quite high for Hi, which likely has low representation in the model training corpus. This indicates that the ability to reason by inheritance likely depends on the amount of specific language data during training LLMs, which in turn points to gaps in inductive biases in LLMs. Fig 20 in Appendix A.6.2 shows few LLM responses that violate inheritance constraints with conflicting answers for parent and child concepts.

5 Compositional Representation (CoRe)

We now consider mechanisms to mitigate lack of knowledge consistency across languages which emerged as a problem in the prior sections. Our analysis points to two key observations:

1. "Identical" representations for equivalent concepts ensures perfect knowledge transfer while "distant" representations lead to separate copies of knowledge being learned.
2. Languages from different families (such as En and Hi) have distant LLM representations for equivalent concepts resulting in low knowledge sharing between them and high inconsistency unless duplicate information is fed in both languages.

Based on these observations, we propose CoRe with the aim of bridging distant representation spaces to enable greater knowledge sharing. It hinges on the key idea that representing a concept via composition of representations of all equivalent concepts across languages would enable LLM to maintain consistent knowledge for that concept across languages as shown in Fig. 8. This is in contrast to the current LLM models (Fig. 9) where a sentence input into the transformer decoder in LLMs does not have access to representations of equivalent tokens from distant languages and can only utilize the localised knowledge in that language's representation space.

5.1 Methodology

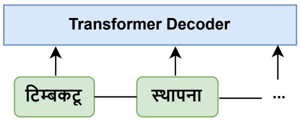


Figure 9: Default mechanism in Transformers.

Attention (Vaswani et al., 2017) is an essential mechanism of transformer architecture that converts an input sequence into a latent encoding using representational vectors formed from the input, i.e., queries, keys and values to determine the importance of each portion of input while decoding. Typically, in transformers, each token pays attention to other nearby tokens in the input sequence. Since this process can miss out on equivalent tokens in other languages, especially in the absence of a parallel multi-lingual corpus, we modify the learning approach to consider all tokens in the vocabulary as candidates for attention. Specifically our methodology consists of two steps: (a) proximal token selection, and (b) construction of compositional representation, which we describe below.

Step 1: Proximal token selection. For a given token, we first select the top-n proximal tokens across each language based on compatibility of the existing representations. Formally, let $X = [x_i]_{i=1}^{N}$ and $Z = [z_i]_{i=1}^{N}$ denote the sequence of input tokens and the associated embedding representations of size d_o. Let $U = [u_j]_{j=1}^{M}$ be all the vocabulary tokens and $B = [b_j]_{j=1}^{M}$ be the associated embedding representations of size d_o. Further, let U^r be the vocabulary tokens associated with the language $r \in R$, where R denotes the entire set of languages being considered. Let $Q = [q_i]_{i=1}^{N}$, $K = [k_j]_{j=1}^{M}$ and $V = [v_j]_{j=1}^{M}$ be the sequences of query, key and value vectors of dimensions d_k, d_k and d_v respectively, given by $q_i = z_i^T W_Q$, $k_j = b_j^T W_K$ and $v_j = b_j^T W_V$ where $W_Q \in R^{d_o \times d_k}$, $W_K \in R^{d_o \times d_k}$ and $W_V \in R^{d_o \times d_v}$ are the learned projection matrices. Let $C = \frac{QK^T}{\sqrt{d_k}} \in R^{N \times M}$ be the compatibility matrix between Q and K.

For each input token x_i, we identify the top n proximal or most-compatible tokens from each language $r \in R$ as per the compatibility values:

$$U_{sel}^r(i) = \{u_j | C_{ij} \in \text{top-}n(\{C_{ij} | u_j \in U^r\}),$$

where top-$n(\cdot)$ denotes the largest n values of the input set. Note that in addition to using the organic representations for estimating compatibility, we could also use additional cues from domain ontologies or dictionaries to construct this proximal token set. Further, to ensure computational efficiency, instead of considering the entire set of vocabulary, a smaller candidate pool of tokens per language can be chosen using K-NN based on existing token embeddings at each stage.

Step 2: Construction of compositional representation. The next step is to build a compositional representation from all the selected proximal tokens similar to regular attention mechanism. To ensure only the selected proximal tokens contribute, we define $f(C) = [f_{ij}]$ where,

$$f_{ij} = \begin{cases} 0 & \text{if } u_j \in U_{sel}^r(i) \ \ \forall r \in R \\ -\infty & \text{otherwise.} \end{cases} \quad (1)$$

We augment representation of Z being input to the decoder layer as $Z' = softmax(C + f(C))V$. Since R includes all the languages including ones distant from that of the input sequence X, Z' becomes a composition of equivalent tokens from all the languages yielding more consistent responses.

5.2 Experiments and Results

CoRe augments the transformer architecture and can be used while pretraining LLMs. However, since pretraining entails additional compute cost, it is preferable to use CoRe with existing pre-trained models. For our experiments, we augment pre-trained XGLM-4.5B model with CoRe.

Dataset, Downstream Task and Implementation Details. Since we want to examine if CoRe helps improve consistency among distant languages, we use the same dataset and setup as the controlled experiment in Sec 3.2.1. We augment XGLM architecture and add CoRe to it. We initialize learnable projection matrices W_Q, W_K, W_V by identity matrix to ensure stable continual training and choose $n \in \{5, 10, 15\}$ for selecting the top-n tokens for our experiment. In our experiments, for

	n	en-hi	en-de	en-hien	hi-en	hi-de	hi-hien	de-en	de-hi	de-hien	hien-en	hien-hi	hien-de
XGLM-4.5B	-	74.4	58.6	73.2	72.6	73.6	72.4	57.1	75	74.3	70.5	72.1	71.8
XGLM-4.5B + CoRe	5	73.5	55.6	**71.9**	71.9	72.7	71.9	55.7	74	**73.3**	**68.4**	70.4	69.5
	10	**72.8**	**53.9**	73.1	**71**	**72.2**	72.2	**53.8**	**73.9**	73.9	69.8	**70**	**69.1**
	15	73.2	56.8	72.3	71.3	73.3	**71.8**	56.8	75.5	73.5	68.7	71	69.2

Table 4: Effect of CoRe on conflict rate for different language pairs.

each one of the languages we are working with (En, Hi, De, HiEn), we consider the top-n most compatible tokens from all vocabulary tokens of that language and En. This considers En as the anchor language and helps build a "bridge" between distant representation spaces of other languages and En. We did this for more efficient experimentation but there is no constraint in CoRe on the set of languages from which we can choose top-n proximal tokens. We identify the language(s) a vocabulary token can belong to beforehand using language detector from AWS Comprehend (AWS, 2017a) and store the asset for repeated use during forward pass. We use Pytorch (Paszke et al., 2019) and Huggingface Transformers library (Wolf et al., 2020) for implementation. We train baseline XGLM-4.5B and XGLM-4.5B+CoRe for 20k steps on p3dn.24xlarge machine with 8 GPUs with learning rate of 1e-05 and linear learning rate scheduler with 80040 max steps, gradient accumulation steps of 2 and per device training batch size of 1.

Results. Table 4 shows that CoRe consistently reduces conflict rate across 12 language pairs, with gains up to 4.7%. Variation in conflict rate with different values of n suggests that keeping n static is not ideal as it might add noise to representations in some cases. Distribution of cosine similarity of random 1K parallel En-De words in Fig 10 shows that representations from CoRe for equivalent words are closer with similar behavior observed for other language pairs. Fig 21 in Appendix A.7.1 shows sample questions where XGLM-4.5B+CoRe provides more consistent answers compared to XGLM-4.5B (baseline). To evaluate CoRe's impact on a downstream task we did a small scale experiment for NLI task on XNLI (Conneau et al., 2018) dataset and observed 14% reduction in inconsistency across predictions for parallel En-De NLI data points without hurting NLI performance, see Appendix A.7.2 for more details.

Efficiency of CoRe. In our experiments, adding CoRe increased the training time by ~50% but that is without using FlashAttention (Dao et al., 2022) for CoRe which is expected to be ~2x faster.

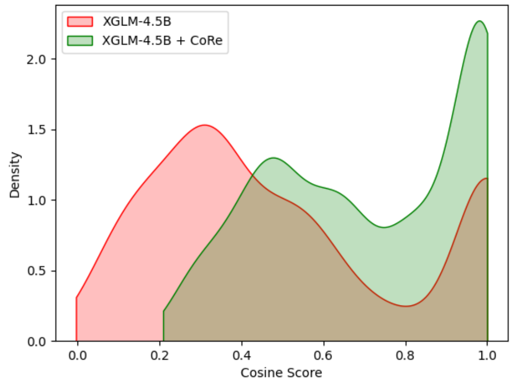


Figure 10: Distribution of cosine score of random 1K parallel En-De words.

We were unable to experiment using FlashAttention because we only had access to V100 machine. Since compatibility matrix is constructed by dot product of Q and K matrices as described in Sec 5.1, this is a computationally intensive operation which can be made more efficient by considering a smaller candidate pool of tokens per language of interest using a more efficient method like K-NN before constructing compatibility matrix. We will be experimenting with these alternatives which can improve computational efficiency of CoRe as part of our future work.

6 Conclusion

We introduce "Reasoning by Equivalence" and "Reasoning by Inheritance" tasks and evaluate popular LLMs to highlight the lack of consistent representation across languages. This systemic gap manifests in inefficient learning, limited knowledge sharing, and over-reliance on extensive data and computational resources, pointing to the need for better representations. We also perform controlled experiments to identify the influencing factors and propose CoRe to bridge the gap between distant language representations which leads to 4.7% boost in performance. We hope our work spurs further research on gaining richer understanding of LLM-based reasoning across languages.

Limitations

Our current work has a few limitations, which we discuss below.

Scope of Relationships: Our study focuses on LLMs' ability to reason based on two foundational relationships: "equivalence" and "inheritance." Future research could broaden this scope to include other key relationships such as "cause and effect," "comparison," and "mereological" relationships. Our proposed CoRe approach is also applicable only to symmetric relationships such as "equivalence," but there is a possibility of extending to asymmetric and transitive relationships using hyperbolic representations.

Evaluation Focus: Current experiments primarily targeted objective, factoid-based questions, chosen for their clarity and the feasibility of automated evaluation via LLM-based judges. This approach facilitated a less ambiguous assessment of LLM reasoning capabilities and the benefits of our CoRe-based mitigation. However, reasoning tasks do encompass subjective, long-form generation tasks such as summarization and problem solving, which can be explored in future, since that could entail access to expensive human-in-the-loop evaluations.

LLM Architectures: Our current work focuses exclusively on transformer-based autoregressive generative LLMs,which include widely used models such as Claude, XGLM, GPT-4, and LLama. Recent advances in neuro-symbolic methods offer alternative architectures and training methods that enhance reasoning abilities albeit at a higher computational costs that limits their adoption in real-world application. Our research identifies specific gaps in the popular LLMs, highlighting the need to integrate ideas from neuro-symbolic research.

Ethics Statement

Our research aims to identify and address gaps in the reasoning abilities of widely used LLMs, particularly for low-resource languages used by a large population of the world. To ensure the validity of our findings, we created a new parallel factoid QA datasets and conducted controlled experiments to prevent data duplication across languages in LLM training data from influencing LLM reasoning performance. The datasets used have no associated privacy or intellectual property concerns, and we plan to open-source them post-review to adhere to double-blind protocol and ensure reproducibility. Evaluation was performed using automated LLMs, with prompts detailed in the appendix for transparency. Beyond the common ethical considerations of using generative language models, our work did not involve any additional ethical issues.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. *Preprint*, arXiv:2303.12528.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *Preprint*, arXiv:2311.07463.

Shengnan An, Zeqi Lin, Bei Chen, Qiang Fu, Nanning Zheng, and Jian-Guang Lou. 2023. Does deep learning learn to abstract? a systematic probing framework. *Preprint*, arXiv:2302.11978.

Anthropic. 2023a. Anthropic support. `https://support.anthropic.com/en/articles/7996845-what-are-some-things-i-can-use-claude-for`. Accessed: 14-10-2024.

Anthropic. 2023b. Claude 2 model card. `https://www-cdn.anthropic.com/5c49cc247484cecf107c699baf29250302e5da70/claude-2-model-card.pdf`. Accessed: 14-10-2024.

Anthropic. 2023c. Welcome to claude. `https://docs.anthropic.com/claude/docs/intro-to-claude`. Accessed: 30-04-2024.

Anthropic. 2024. Claude 3 model card. `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`. Accessed: 14-10-2024.

Gaurav Arora, Srujana Merugu, and Vivek Sembium. 2023. CoMix: Guide transformers to code-mix using POS structure and phonetics. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7985–8002, Toronto, Canada. Association for Computational Linguistics.

Yaniv Aspis, Krysia Broda, Jorge Lobo, and Alessandra Russo. 2022. Embed2Sym - Scalable Neuro-Symbolic Reasoning via Clustered Embeddings. In *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning*, pages 421–431.

AWS. 2017a. Aws comprehend. `https://aws.amazon.com/comprehend/`. Accessed: 14-10-2024.

AWS. 2017b. Aws translate. https://docs.aws.amazon.com/translate/. Accessed: 30-04-2024.

AWS. 2023. Aws bedrock claude. https://aws.amazon.com/bedrock/claude/. Accessed: 14-10-2024.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *CoRR*, abs/2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

Fausto Giunchiglia and Toby Walsh. 1992. A theory of abstraction. *Artificial Intelligence*, 57(2):323–389.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. *Preprint*, arXiv:2212.10403.

Clark L Hull. 1920. Quantitative aspects of the evolution of concepts: An experimental study. *Psychological Monographs*, 28(1):1–85.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Sreejan Kumar, Ishita Dasgupta, Nathaniel D. Daw, Jonathan D. Cohen, and Thomas L. Griffiths. 2023. Disentangling abstraction from statistical pattern matching in human and machine learning. *Preprint*, arXiv:2204.01437.

Nicolas Lazzari, Stefano De Giorgis, Aldo Gangemi, and Valentina Presutti. 2024. Sandra – a neuro-symbolic reasoner based on descriptions and situations. *Preprint*, arXiv:2402.00591.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2021. Representation learning for natural language processing. *CoRR*, abs/2102.03732.

Emanuele Marconato, Gianpaolo Bontempo, Elisa Ficarra, Simone Calderara, Andrea Passerini, and Stefano Teso. 2023. Neuro-symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal. *Preprint*, arXiv:2302.01242.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Marvin Minsky et al. 1974. A framework for representing knowledge.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/index/chatgpt`. Accessed: 30-04-2024.

OpenAI, Josh Achiam, and Steven Adler et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Greg Rushing. 2020. Aws machine learning blog. `https://aws.amazon.com/blogs/machine-learning/amazon-translate-ranked-as-1-machine-translation-provider-by-intento`. Accessed: 11-10-2024.

Teven Le Scao, Angela Fan, and Christopher Akiki et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023. Do PLMs know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3080–3101, Toronto, Canada. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Appendix

A.1 Existence of Equivalence/Inheritance Relationships vs Using them in Multi Step Reasoning

It is important to distinguish between **(a)** the existence of "equivalence" and "inheritance" relationships among concepts in LLM representations, and **(b)** the LLM's ability to actually utilize these relationships (encoded in the representations) for multi step reasoning. The former can be considered a prerequisite for the latter.

By the "existence of an equivalence relationship" between A and B, we mean that the LLM's representations and their direct use during inference allows properties of A and B to be transferred to each other or reconciled for conflicts. Similarly, an "inheritance relationship" between A and B would imply that A inherits properties of B.

Consider the question `"Does the river Kaveri flow in the same continent as the river Seine?"`. We could answer this correctly as

```
1. The Kaveri river flows in India
2. The Seine river flows in France
3. India is part of the continent Asia
4. France is part of the continent Europe
5. Kaveri = कावेरी
6. India = भारत
7. flows in = में बहती
8. Seine = सीन
9. France = फ्रांस
10. Continent = महाद्वीप
```

Figure 11: An example of prior information which can be combined to answer the question "Does the river Kaveri flow in the same continent as the river Seine?" in both English and Hindi. Equivalence relationships are in bold.

"No" using different sets of prior information with varying levels of reasoning applied. For example, in **Scenario 1**, we might directly utilise the information "Kaveri and Seine flow in different continents". Alternatively, in **Scenario 2**, we might need to combine multiple pieces of information (expressed as entity-attribute or entity relationship predicates) as shown in Fig 11 to arrive at the answer. Scenario 2 demonstrates the ability to generalize from a small set of training data with limited compute effort to address a broader range of questions. To arrive at the correct answer, we need both the basic building block relationships as well as the capability to effectively combine them as part of LLM inference.

In our current study, we are primarily focused on evaluating whether the basic equivalence/inheritance relationships exist in LLMs by assessing simple property transfer and conflict resolution on questions related to equivalent entities or parent-child entities. It is possible that even when the desired relationships hold and properties transfer correctly, there may be gaps in the overall multi-step reasoning process, resulting in an inaccurate LLM response. This is because the existence of "abstraction" or "equivalence/inheritance relationships encoded in LLM representations" does not necessarily mean that the LLM would always effectively leverage this information for its reasoning since the LLM's inference mechanism, which relies on associative attention, differs from logical operations. We do not yet evaluate this larger capability (item b) because there are multiple factors involved, and there is likely a gap in item (a) itself.

```
EN: Which continent is Kaveri located in?

HI: कावेरी किस महाद्वीप में स्थित है?
```

Figure 12: An example parallel question from MLAMA (Kassner et al., 2021) dataset.

A.2 Impact of Prior Information seen during Training

Consider a pair of parallel questions from the MLAMA (Kassner et al., 2021) dataset ("Which continent is Kaveri located in ?") as shown in Fig 12. When an LLM provides the same response "Asia" without conflicts for the two parallel questions in En and Hi, it could be due to one of the following two reasons:

1. **Existence of Equivalence relationship/Common Abstraction**: The LLM is aware of equivalence between the En and Hi versions of entities "Kaveri" and "India" in Fig 11 beyond just the basic language constructs.

2. **Duplication of Information**: The LLM could have seen two parallel, aligned pieces of information that "Kaveri is located in Asia" in both En and Hi even though there LLM's representations of the corresponding equivalent entities are highly divergent. In the latter case, while the LLM can answer this specific question well without conflicts, that behavior might not generalize well to other questions about "Kaveri" in Hi that were not part of the training data.

Hence, conflict rate estimates could lead to misleading conclusions about the existence of equivalence and inheritance relationships for well-known real entities due to the bias introduced by prior duplicate (or even contradictory) information in the LLM training data across languages. This **effect of "duplicate knowledge in LLM training"** is also reflected in our results, where the conflict rate numbers in Table 4 are higher than those in Tables 1 and 2. We expect the effect observed in Tables 1 and 2 to be more pronounced in public datasets like MLAMA, as they may be part of the LLM's training data, either directly or indirectly. For instance, we generated responses from Claude v1 instant for parallel En – De 100 randomly sampled questions from MLAMA dataset and observed ∼15% conflict

En-Fr	En-Es	En-De	En-Pt	En-Hi	Avg
96%	97%	94%	95%	94%	95.20%

Table 5: Precision of Claude v3 Sonnet as the judge in identifying conflicting answers across language pairs

rate which is much lower than the average results in Tables 1 and 2 (∼30%).

A.3 Accounting for LLMs' proficiency in different languages

We evaluate LLMs only on the languages for which there is official documentation of support for that language or there is prior work demonstrating good performance on NLP tasks of that language. For example, Anthropic claims support for English, Spanish, Portuguese, French, German and multiple other languages (Anthropic, 2023a; AWS, 2023), and also showcases Claude's multilingual capabilities on these languages. Hindi is also mentioned in Claude model cards (Anthropic, 2023b, 2024). In our study, the LLMs we evaluate on a language are proficient with respect to the linguistic patterns and the common vocabulary of that language, which is different from the knowledge (factual information) aspects. Further in our evaluation, we focused on well-formed objective factual questions to avoid variations due to subjective interpretations and cultural nuances. This ensures that the knowledge consistency or alternatively the conflict rate in responses across languages primarily depends on (a) the knowledge duplication in the training data and (b) the effectiveness of knowledge transfer, i.e., equivalent entity representations across the languages. The point we wish to highlight in our work is that LLM training and representations should be designed so as to enable efficient knowledge transfer within and across languages.

A.4 Judge Precision

Table 5 shows precision of Claude v3 Sonnet as the judge in identifying conflicting answers across language pairs obtained by annotating random sample of 100 answer pairs for 5 language pairs each. Prompt for Claude v3 Sonnet judge is shown in Fig 13.

A.5 Reasoning by Equivalence

A.5.1 Data Preparation

We hand-curated 51 (parent) abstract concepts manually (e.g., monuments, actors, cities, etc.) that primarily correspond to common nouns. See Fig 14 for full list of the 51 abstract concepts. Then, we created 3641 named entities that are specific instances of these (parent) abstract concepts, e.g., Taj Mahal is a specific instance of monument) using Claude with human review to weed out non-existent ones. For each one of the 3641 named entities, we prepared a set of questions which have objective or factual answers using Claude with the following prompt.

Give different unambiguous complete questions about "{entity}" which have specific factual answers.

We built this prompt after multiple iterations of analysis of generated questions for a small sample of entities. The dimensions of evaluation were: (i) The generated question should be complete and unambiguous, i.e. it should be clear which entity the question is about and what attribute/fact is being asked, (ii) The answer to the generated question should be an unambiguous factual response. From the final set of generated 88,334 questions, we randomly sampled and manually annotated 500 generated questions of which 96.2% were complete and unambiguous, and 99.4% had an unambiguous factual answer. The annotations were done by a professional English speaker.

Our original question set is in En which we also translate to Fr, Es, De, Pt and Hi using AWS Translate (AWS, 2017b). We translated using AWS translate which is one of the best commercial translation services (Rushing, 2020). AWS translate is expected to translate questions while preserving their meaning. We rely on AWS translate to translate the concept into most natural variant in case multiple variants are possible. To get an estimate of lower bound of AWS translate's performance on our datasets, we consider the En-Hi translation task since Hi being a non-Latin language with different lexical representation is much more divergent from En. We enlisted a professional bilingual Hindi and English speaker to annotate a random sample of 200 En-Hi question pairs on the translation accuracy and observe 97.5% accuracy. This high translation accuracy is likely due to the nature of our English questions, which are well-formed and unambiguous. Fig 15 shows few sample questions from our dataset.

A.5.2 Analysis

Fig 16 and Fig 17 show sample conflicting answers from various LLMs to equivalent questions from DIL and DOL tasks respectively.

```python
def create_prompt(question, response1, response2):
    prompt_data = f"""Human: Below given are two answers to the question '{question}'.

      First Answer: {response1}

      Second Answer: {response2}

      Are above answers contradictory to each other. First share your reasoning in one line briefly within <thinking><thinking> tags. Be precise.
Finally, within <response></response> tags provide final answer with a Yes or No only.

      Assistant:"""
    return prompt_data
```

Figure 13: Prompt for Claude v3 Sonnet to use it as the judge.

{'Monument', 'Australian Open', 'Thomas Cup', 'US Open', 'Vegetables', 'Bikes', 'Organizations', 'Football Asian Cup', 'Songs', 'Beaches', 'Actress', 'All England Open', 'Fruits', 'Sudirman Cup', 'Football African Cup of Nations', 'Museum', 'French Open', 'National Parks', 'Singer', 'Books', 'City', 'Movie', 'State', 'Poets', 'Uber Cup', 'Cars', 'Country', 'Badminton Player', 'Animals', 'ICC Cricket World Cup years', 'Movie Directors', 'Cricket Player', 'Football Copa America', 'World Championships', 'Politicians', 'Actor', 'Football Player', 'Lakes', 'Football CONCACAF Gold Cup', 'World Junior Championships', 'UEFA European Championship', 'ICC Champions Trophy', 'Asia Cup', 'ICC Test Championship', 'ICC World Twenty20', 'Flowers', 'Tennis Player', 'Authors', 'Fibers', 'FIFA World Cup', 'Wimbledon'}

Figure 14: List of 51 (parent) abstract concepts.

Parent Concept	Child Concept	Question (En)	Question (Fr)	Question (Hi)	Question (De)	Question (Pt)	Question (Es)
Monument	Great Wall of China	How long is the Great Wall of China?	Quelle est la longueur de la Grande Muraille de Chine ?	चीन की महान दीवार कितनी लंबी है?	Wie lang ist die Chinesische Mauer?	Quanto tempo dura a Grande Muralha da China?	¿Qué longitud tiene la Gran Muralla China?
Movie	Swades	Who starred as the lead actor in the movie Swades?	Qui a joué le rôle principal dans le film Swades ?	स्वदेस फिल्म में मुख्य अभिनेता के रूप में किसने अभिनय किया था?	Wer spielte die Hauptrolle im Film Swades?	Quem estrelou como ator principal no filme Swades?	¿Quién interpretó al actor principal en la película Swades?
Cricket Player	Kieron Pollard	What is the highest individual score by Kieron Pollard in a T20 international match?	Quel est le meilleur score individuel de Kieron Pollard lors d'un match international du T20 ?	एक T20 अंतरराष्ट्रीय मैच में कीरोन पोलार्ड का सर्वोच्च व्यक्तिगत स्कोर क्या है?	Was ist die höchste Einzelpunktzahl von Kieron Pollard in einem T20-Länderspiel?	Qual é a maior pontuação individual de Kieron Pollard em uma partida internacional T20?	¿Cuál es la puntuación individual más alta de Kieron Pollard en un partido internacional del T20?
French Open	2006	Who did Justine Henin beat in the 2006 French Open women's singles final?	Qui a battu Justine Henin lors de la finale du simple féminin des Internationaux de France 2006 ?	2006 के फ्रेंच ओपन महिला एकल फाइनल में जस्टिन हेनिन ने किसे हराया था?	Wen hat Justine Henin 2006 im Finale der French Open im Dameneinzel besiegt?	Quem Justine Henin venceu na final individual feminina do Aberto da França de 2006?	¿A quién derrotó Justine Henin en la final individual femenina del Abierto de Francia 2006?
Country	Germany	What is the capital of Germany?	Quelle est la capitale de l'Allemagne ?	जर्मनी की राजधानी क्या है?	Was ist die Hauptstadt von Deutschland?	Qual é a capital da Alemanha?	¿Cuál es la capital de Alemania?

Figure 15: Sample questions from our question bank which we use to evaluate LLMs for reasoning by equivalence.

A.5.3 Additional Details on Controlled Experiment

For computing the rank of a multi-token word, we compute the rank of each one of its tokens and consider the minimum rank amongst them as the rank of multi-token word. Figure 18 shows conflict rate of all (anchor, non-anchor) language pairs with number of training steps.

A.6 Reasoning by Inheritance

A.6.1 Dataset

Fig 19 shows sample questions about abstract concepts and their specific instances from our dataset.

A.6.2 Qualitative Analysis

Fig 20 shows a few sample errors from various LLMs wherein they violate inheritance constraints by giving conflicting answers for parent and child concept, and across children of the same type.

A.7 Compositional Representation (CoRe)

A.7.1 Qualitative Analysis

Fig 21 shows few sample anchor and non-anchor language questions with their answers from XGLM-4.5B (baseline) and XGLM-4.5B+CoRe.

A.7.2 Effect of CoRe on a Downstream Task

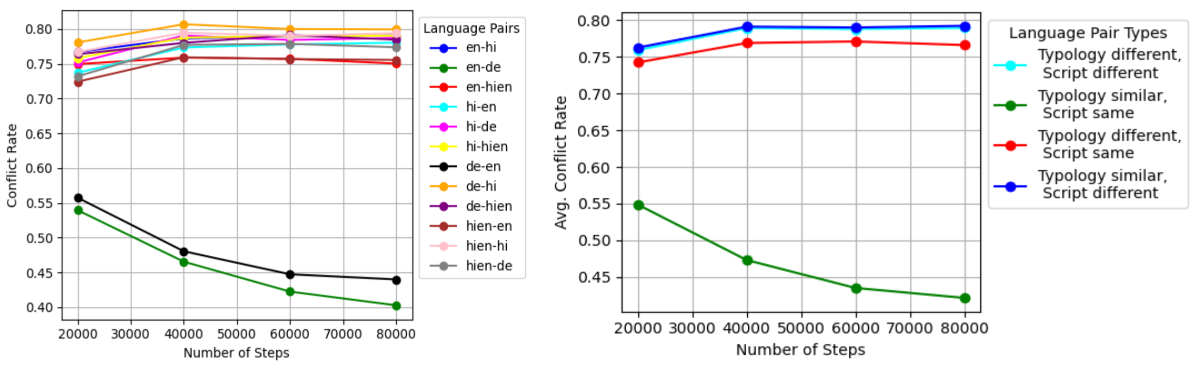
The main focus of our work in "Reasoning by Equivalence" was to assess if LLMs exhibit consistent knowledge across languages, and we proposed CoRe to improve consistency. However, to evaluate CoRe's impact on a downstream task we performed a small experiment for NLI task on XNLI (Conneau et al., 2018) dataset (on first 100 examples from validation set) with the same zero shot setup as XGLM paper (Lin et al., 2022) and observed comparable performance with and without CoRe (46-47 for En, 39-40 for De). However, we observed a significant reduction in inconsistency across predictions for parallel $En - De$ NLI data points (28% inconsistency in predictions without CoRe vs 14% inconsistency with CoRe, wherein model's prediction is said to be inconsistent if it predicts different labels for parallel datapoints across languages (En and De in this case)). We anticipate that incorporating CoRe during pre-training or continual training with larger dataset could yield further improvement in downstream task performance as well which we could not do because of compute constraints. We consider those experiments and further evaluation of CoRe across various different tasks

Parent Concept	Child Concept	Lang.	Question	Sample answers from different LLMs			
Monument	Great Wall of China	En	How long is the Great Wall of China?	BLOOMZ -7B	2400 km	Claude v2	The Great Wall of China is 21,196 km (13,171 miles) long.
		Fr	Quelle est la longueur de la Grande Muraille de Chine ?		2400 km		La Grande Muraille de Chine fait environ 21 196 km de long.
		Hi	चीन की महान दीवार कितनी लंबी है?		3,000 km		चीन की महान दीवार की कुल लंबाई लगभग 21,196 किलोमीटर है।
		De	Wie lang ist die Chinesische Mauer?		1949		Die Chinesische Mauer ist etwa 21.196 Kilometer lang.
		Pt	Quanto tempo dura a Grande Muralha da China?		over 2400 years		A Grande Muralha da China tem uma extensão de aproximadamente 21.196 km.
		Es	¿Qué longitud tiene la Gran Muralla China?		18,000 km		La Gran Muralla China tiene una longitud de aproximadamente 21.196 kilómetros.
Movie	Swades	En	Who starred as the lead actor in the movie Swades?	XGLM-7.5 B	Amitabh Bachchan.	Claude v1 Instant	Shah Rukh Khan
		Fr	Qui a joué le rôle principal dans le film Swades ?		Amitabh Bachchan.		Shah Rukh Khan
		Hi	स्वदेस फिल्म में मुख्य अभिनेता के रूप में किसने अभिनय किया था?		दिलीप कुमार.		अमिताभ बच्चन
		De	Wer spielte die Hauptrolle im Film Swades?		Swades.		Shah Rukh Khan
		Pt	Quem estrelou como ator principal no filme Swades?		Amitabh Bachchan.		Shah Rukh Khan
		Es	¿Quién interpretó al actor principal en la película Swades?		Amitabh Bachchan.		Shah Rukh Khan
Cricket Player	Kieron Pollard	En	What is the highest individual score by Kieron Pollard in a T20 international match?	Claude v3 Sonnet	83 not out against Sri Lanka in 2021	Claude v2	75 runs
		Fr	Quel est le meilleur score individuel de Kieron Pollard lors d'un match international du T20 ?		83 non out contre le Sri Lanka en 2021		119
		Hi	एक T20 अंतरराष्ट्रीय मैच में कीरोन पोलार्ड का सर्वोच्च व्यक्तिगत स्कोर क्या है?		109 रन (नाबाद)		56
		De	Was ist die höchste Einzelpunktzahl von Kieron Pollard in einem T20-Länderspiel?		83 Runs gegen Sri Lanka im Jahr 2010		75
		Pt	Qual é a maior pontuação individual de Kieron Pollard em uma partida internacional T20?		89 não eliminado contra Sri Lanka em 2010		104 não fora
		Es	¿Cuál es la puntuación individual más alta de Kieron Pollard en un partido internacional del T20?		83 no out contra Argentina en 2022		104
French Open	2006	En	Who did Justine Henin beat in the 2006 French Open women's singles final?	Claude v1 Instant	Kim Clijsters	Claude v3 Sonnet	Justine Henin beat Svetlana Kuznetsova in the 2006 French Open women's singles final.
		Hi	2006 के फ्रेंच ओपन महिला एकल फाइनल में जस्टिन हेनिन ने किसे हराया था?		2006 के फ्रेंच ओपन महिला एकल फाइनल में जस्टिन हेनिन ने निकोल वाइडिसोवा को हराया था।		जस्टिन हेनिन ने 2006 के फ्रेंच ओपन महिला एकल फाइनल में रूसी खिलाड़ी स्वेतलाना कुज्नेत्सोवा को हराया था।
		De	Wen hat Justine Henin 2006 im Finale der French Open im Dameneinzel besiegt?		Amélie Mauresmo		Justine Henin besiegte Swetlana Kusnezowa im Finale der French Open 2006 im Dameneinzel.
		Pt	Quem Justine Henin venceu na final individual feminina do Aberto da França de 2006?		Justine Henin venceu a italiana Amélie Mauresmo na final individual feminina do Aberto da França de 2006.		Justine Henin venceu Svetlana Kuznetsova na final individual feminina do Aberto da França de 2006.
		Es	¿A quién derrotó Justine Henin en la final individual femenina del Abierto de Francia 2006?		A Amélie Mauresmo		Justine Henin derrotó a Svetlana Kuznetsova en la final individual femenina del Abierto de Francia 2006.
Country	Germany	En	What is the capital of Germany?	XGLM-7.5 B	Berlin.	BLOOMZ -7B	Berlin
		Fr	Quelle est la capitale de l'Allemagne ?		Berlin.		Berlin
		Hi	जर्मनी की राजधानी क्या है?		बर्लिन.		Berlin
		De	Was ist die Hauptstadt von Deutschland?		Berlin.		Berlin
		Pt	Qual é a capital da Alemanha?		Berlin.		Berlin
		Es	¿Cuál es la capital de Alemania?		Berlin.		Berlin

Figure 16: Sample errors from different LLMs on DIL task. Red and Green colored cells highlight conflicting and non-conflicting answers with the anchor language (En) answer, respectively.

Parent Concept	Child Concept	Question	Lang.	Sample answers from different LLMs			
Monument	Great Wall of China	How long is the Great Wall of China?	En	Claude v1 Instant	The Great Wall of China is approximately 13,170 miles (21,196 kilometers) long.	Claude v2	The Great Wall of China is 21,196 km (13,171 miles) long.
			Fr		La Grande Muraille de Chine mesure environ 21 000 kilomètres de long.		La Grande Muraille de Chine fait environ 21 196 km de long.
			Hi		चीन की महान दीवार की लंबाई लगभग 21,196 किलोमीटर है।		ग्रेट वॉल ऑफ चीन की लंबाई लगभग 21,196 किलोमीटर है।
			De		Die Chinesische Mauer ist etwa 21.196 Kilometer lang.		Die Große Mauer ist über 21.000 Kilometer lang.
			Pt		A Grande Muralha da China mede aproximadamente 21 mil quilômetros de comprimento.		A Grande Muralha da China tem uma extensão de 21.196 quilômetros.
			Es		La Gran Muralla China mide aproximadamente 21,196 kilómetros de largo.		La Gran Muralla China mide 21,196 kilómetros de largo.
Movie	Jo Jeeta Wohi Sikandar	Who directed the music for the movie Jo Jeeta Wohi Sikandar?	En	Claude v1 Instant	A.R. Rahman	Claude v3 Sonnet	Jatin-Lalit composed the music for the 1992 Hindi film Jo Jeeta Wohi Sikandar.
			Fr		A. R. Rahman		La musique du film Jo Jeeta Wohi Sikandar a été composée par Uttam Singh.
			Hi		फिल्म 'जो जीता वोही सिकंदर' के लिए संगीत का निर्देशन अनु मलिक ने किया था।		जो जीता वही सिकंदर फिल्म की संगीत निर्देशन विष्णु शर्मा ने किया था।
			De		Der indische Filmkomponist Anu Malik komponierte die Musik für den Film Jo Jeeta Wohi Sikandar.		Die Musik für den Film Jo Jeeta Wohi Sikandar wurde von Bappi Lahiri komponiert.
			Pt		Anees Bazmee		A música do filme Jo Jeeta Wohi Sikandar foi composta por Uttam Singh.
			Es		A. R. Rahman dirigió la música para la película Jo Jeeta Wohi Sikandar.		La música de la película Jo Jeeta Wohi Sikandar fue compuesta por Jatin-Lalit.
Cricket Player	Kieron Pollard	What is the highest individual score by Kieron Pollard in a T20 international match?	En	Claude v1 Instant	75 runs against Sri Lanka in 2012	Claude v3 Sonnet	83 not out against Sri Lanka in 2021
			Fr		127		Le score individuel le plus élevé de Kieron Pollard dans un match international T20 est 83 non out contre l'Argentine en 2022.
			Hi		कियरोन पोलार्ड ने टी20 इंटरनेशनल मैच में सबसे ज्यादा अकेले रन बनाए हैं 75 रन।		किरोन पोलार्ड द्वारा एक टी20 अंतरराष्ट्रीय मैच में सर्वाधिक व्यक्तिगत स्कोर 75 रन है।
			De		127 gegen Sri Lanka im Jahr 2019		Kieron Pollards höchste individuelle Punktzahl in einem T20-Länderspiel ist 83 Runs gegen Sri Lanka im Jahr 2021.
			Pt		127 contra o Sri Lanka em 2012		O maior score individual de Kieron Pollard em uma partida internacional de T20 é 89 não eliminado contra a Índia em 2016.
			Es		127		El puntaje individual más alto de Kieron Pollard en un partido internacional de T20 es 83 no out contra Argentina en 2022.
French Open	2006	Who did Justine Henin beat in the 2006 French Open women's singles final?	En	Claude v1 Instant	Svetlana Kuznetsova	Claude v3 Sonnet	Justine Henin beat Svetlana Kuznetsova in the 2006 French Open women's singles final.
			Hi		जस्टिन हेनिन ने 2006 फ्रेंच ओपन महिला एकल फाइनल में मारिया शारापोवा को हराया था।		जस्टिन हेनिन ने 2006 फ्रेंच ओपन महिला एकल फाइनल में रूसी खिलाड़ी स्वेतलाना कुज्नेत्सोवा को हराया था।
			De		Justine Henin besiegte Katarina Srebotnik im Damen-Einzel-Finale der French Open 2006.		Justine Henin besiegte Swetlana Kusnezowa im Finale der French Open 2006 im Dameneinzel.
			Pt		Maria Sharapova		Justine Henin venceu Svetlana Kuznetsova na final de simples feminino do Aberto da França de 2006.
			Es		A Kim Clijsters en la final femenina individual del Abierto de Francia 2006.		Justine Henin venció a Svetlana Kuznetsova en la final de individuales femenina del Abierto de Francia 2006.

Figure 17: Sample errors from different LLMs on DOL task. Red and Green colored cells highlight conflicting and non-conflicting answers with the anchor language (En) answer, respectively.



(a) For different language pairs

(b) Averaged for different language pair types

Figure 18: Figure showing how conflict rate changed during training in controlled experiment.

Parent Concept	Child Concept	Language	Question (Parent Concept)	Question (Child Concept)
Fruit	Banana	De	Kann die Frucht laufen?	Kann die Fruchtbanane laufen?
Fruit	Banana	Hi	क्या फ्रूट चल सकता है?	क्या फ्रूट केला चल सकता है?
Fruit	Lychee	Hi	क्या फ्रूट चल सकता है?	क्या फ्रूट लीची चल सकती है?
Books	Dhalgren	En	Can the Book eat?	Can the Book Dhalgren eat?
Books	Dhalgren	Hi	क्या किताब खा सकती है?	क्या बुक धालग्रेन खा सकता है?
Books	The Great Gatsby	Hi	क्या किताब खा सकती है?	क्या किताब द ग्रेट गैट्सबी खा सकती है?
Monument	Airavatesvara Temple (Tamil Nadu)	En	Can the Monument smell other things?	Can the Monument Airavatesvara Temple (Tamil Nadu) smell other things?
Monument	Airavatesvara Temple (Tamil Nadu)	De	Kann das Monument andere Dinge riechen?	Kann das Monument Airavatesvara-Tempel (Tamil Nadu) andere Dinge riechen?
Monument	Great Wall of China (China)	De	Kann das Monument andere Dinge riechen?	Kann das Monument Chinesische Mauer (China) andere Dinge riechen?
Mountain Ranges	Shivaliks	Pt	A Cordilheira pode voar?	Os Shivaliks da Cordilheira podem voar?
Mountain Ranges	Shivaliks	Es	¿Puede volar la Cordillera?	¿Pueden volar los Shivaliks de la Cordillera?
Mountain Ranges	Himalaya	Es	¿Puede volar la Cordillera?	¿Puede volar la Cordillera del Himalaya?

Figure 19: Dataset sample of questions about abstract concepts and their specific instances which we use to evaluate LLMs for reasoning by inheritance.

Parent Concept	Child Concept	Language	Question (Parent Concept)	Question (Child Concept)	Sample answers from different LLMs			Comment
					Model	To Parent Concept Question	To Child Concept Question	
Fruit	Banana	De	Kann die Frucht laufen?	Kann die Fruchtbanane laufen?	Claude v2	Nein, eine Frucht kann nicht laufen.	Nein, die Fruchtbanane kann nicht laufen, da sie eine Frucht ist und keine Beine zum Laufen hat.	LLM is able to **exhibit reasoning by inheritance for "banana" in De, but not in Hi**. In Hi though, it is able to **reason correctly for "Lychee", a different specific instance of "Fruit", but not for "banana"**.
Fruit	Banana	Hi	क्या फ्रूट चल सकता है?	क्या फ्रूट केला चल सकता है?		नहीं, फ्रूट चल नहीं सकता।	हां, फ्रूट केला चल सकता है।	
Fruit	Lychee	Hi	क्या फ्रूट चल सकता है?	क्या फ्रूट लीची चल सकती है?		नहीं, फ्रूट चल नहीं सकता।	नहीं, फ्रूट लीची चल नहीं सकती।	
Books	Dhalgren	En	Can the Book eat?	Can the Book Dhalgren eat?	Claude v2	No, the Book cannot eat because it is an inanimate object.	No, the Book Dhalgren cannot eat since it is an inanimate object.	LLM is able to **exhibit reasoning by inheritance for "Dhalgren" in En, but not in Hi**. In Hi though, it is able to **reason correctly for "The Great Gatsby", a different specific instance of "Book", but not for "Dhalgren"**.
Books	Dhalgren	Hi	क्या किताब खा सकती है?	क्या बुक धालग्रेन खा सकता है?		नहीं, किताब खाना नहीं खा सकती।	हां, बुक धालग्रेन खा सकता है।	
Books	The Great Gatsby	Hi	क्या किताब खा सकती है?	क्या किताब द ग्रेट गैट्सबी खा सकती है?		नहीं, किताब खाना नहीं खा सकती।	नहीं, किताब नहीं खा सकती	
Monument	Airavatesvara Temple (Tamil Nadu)	En	Can the Monument smell other things?	Can the Monument Airavatesvara Temple (Tamil Nadu) smell other things?	BLOOMZ-7B	Yes	Yes	LLM is able to **exhibit reasoning by inheritance for "Airavatesvara Temple (Tamil Nadu)" in En, but not in De**. In De though, it is able to **reason correctly for "Great Wall of China (China)", a different specific instance of "Monument", but not for "Airavatesvara Temple (Tamil Nadu)"**. Answers are also not consistent across children of the same type, within and across languages.
Monument	Airavatesvara Temple (Tamil Nadu)	De	Kann das Monument andere Dinge riechen?	Kann das Monument Airavatesvara-Tempel (Tamil Nadu) andere Dinge riechen?		no	Yes	
Monument	Great Wall of China (China)	De	Kann das Monument andere Dinge riechen?	Kann das Monument Chinesische Mauer (China) andere Dinge riechen?		no	no	
Mountain Ranges	Shivaliks	Pt	A Cordilheira pode voar?	Os Shivaliks da Cordilheira podem voar?	XGLM-7.5B	A Cordilheira pode voar.	Os Shivaliks da Cordilheira podem voar.	LLM is able to **exhibit reasoning by inheritance for "Shivaliks" in Pt, but not in Es**. In Es though, it is able to **reason correctly for "Himalaya", a different specific instance of "Mountain Ranges", but not for "Shivaliks"**.
Mountain Ranges	Shivaliks	Es	¿Puede volar la Cordillera?	¿Pueden volar los Shivaliks de la Cordillera?		Yes, it can fly.	No.	
Mountain Ranges	Himalaya	Es	¿Puede volar la Cordillera?	¿Puede volar la Cordillera del Himalaya?		Yes, it can fly.	Yes, it can.	

Figure 20: Sample errors from various LLMs wherein they violate inheritance constraints. Red and Green colored cells highlight conflicting and non-conflicting answers with the parent concept question's answer, respectively.

on NLP benchmarks as part of future work.

Anchor		Non-Anchor		XGLM-4.5B (Baseline)		XGLM-4.5B + CoRe	
Lang	**Anchor Question**	**Lang**	**Non-Anchor Question**	**Anchor Answer**	**Non-Anchor Answer**	**Anchor Answer**	**Non-Anchor Answer**
en	When was Mukesh Singh born?	de	Wann wurde Mukesh Singh geboren?	March 15, 1990.	Mukesh Singh wurde am 15. März 1985 in Pokhara, Nepal, geboren.	March 15, 1990.	March 15, 1990.
en	How tall is Mohammad Zaheer?	de	Wie groß ist Mohammad Zaheer?	6 feet 2 inches	6 Fuß 7 Zoll	6 feet 5 inches	6 Fuß 5 Zoll
en	Who did Isla Kirkland take over from as CEO of Zetabyte Corporation in 2020?	de	Von wem hat Isla Kirkland 2020 die Position des CEO der Zetabyte Corporation übernommen?	Daniel Jones	Amanda Zhou	Daniel Jones	Daniel Jones
en	In which Indian city was Madhuri Deshpande born?	de	In welcher indischen Stadt wurde Madhuri Deshpande geboren?	Hyderabad	Vijayawada	Hyderabad	Hyderabad
en	What is the headquarters location of WattGrid Technologies?	hi	WattGrid Technologies का मुख्यालय क्या है?	Shanghai	बीजिंग, चीन	Shanghai	शंघाई, चीन
en	What is the height in meters of the tallest building called Maxima Tower in Ultima?	hi	अल्टिमा में मैक्सिमा टॉवर नामक सबसे ऊंची इमारत की मीटर में ऊंचाई कितनी है?	250 meters	450 मीटर	250 meters	250 मीटर
en	In what city and country was Sanjay Kapoor born?	hien	sanjay kapur kaa janm kis shahar or desh main hua tha?	Mumbai, India	sanjay kapur kaa janm 15 march 1985 ko kapurkapur, bharat main hua tha.	Mumbai, India	sanjay kapur kaa janm 15 march 1985 ko mumbai, maharashtra main hua tha.
en	Which team did Nepal, led by Bikram Khatri, defeat in the 2020 ACC Eastern Region T20 tournament final?	hien	2020 ACC purvi kshetra T20 toornament ke final main bikram khatri ke netritv vaali nepal ne kis team ko haraaya?	Malaysia	australia	Malaysia	malaysia
hi	संदीप सिंह कितने साल के थे जब उन्होंने नेपाली राष्ट्रीय टीम के लिए पदार्पण किया था?	en	How old was Sandeep Singh when he made his debut for the Nepali national team?	16	19 years old	16	16
hi	जॉन डेविडसन किस घरेलू क्रिकेट टीम का प्रतिनिधित्व करते हैं?	en	Which domestic cricket team does John Davidson represent?	न्यू साउथ वेल्स	Queensland	न्यू साउथ वेल्स	New South Wales
hien	srilanka ke ubharate hue ballebaaji khilaadi ramesh tendulkar ne apane kariyar main ab tak kitne list e match khele hai?	hi	श्रीलंका के उभरते हुए बल्लेबाजी खिलाड़ी रमेश तेंदुलकर ने अपने करियर में अब तक कितने लिस्ट ए मैच खेले हैं?	30	250	25	25
hien	jhang vei ne apane test padaarpan par koun sa jersey number pahana tha?	hi	झांग वेई ने अपने टेस्ट पदार्पण पर कौन सा जर्सी नंबर पहना था?	10	19	10	10

Figure 21: Sample anchor and non-anchor language questions wherein we get consistent answers after training with CoRe.