The Emergence of the Pragmatic Dimension in Instructed-LMs

Davide Mazzaccara, Raffaella Bernardi
CIMEC, University of Trento; Free University of Bozen-Bolzano
davide.mazzaccara@unitn.it, Raffaella.Bernardi@unibz.it

Abstract

Instruction-tuning fundamentally transforms how language models process linguistic input and interact with the user. Through the lens of speech act theory, we investigate whether instruction-tuning causes models to shift from prioritizing syntactical form to pragmatic intent. We create a controlled dataset of 400 sentences systematically varying along two dimensions: syntactical structure (declarative vs. interrogative) and communicative intent (assertive vs. request). Using Principal Component Analysis on hidden state representations from Qwen2.5 (1.5B-7B) and models from two other families (Gemma3-1B, and Llama3.2-3B), we reveal a consistent pattern: base models cluster sentences by syntactical form, while instruction-tuned models reorganize representations around pragmatic intent. This syntactic-to-pragmatic shift occurs in middle layers, with declarative requests and interrogative requests—maximally separated in base models—becoming the most similar categories after instruction-tuning. The phenomenon explains how instruction-tuned models correctly interpret indirect speech acts, treating polite declaratives like "I'd appreciate corrections" as functionally equivalent to direct interrogatives. Our findings demonstrate that instruction-tuning teaches models to prioritize the communicative dimension over surface form, a fundamental reorganization consistent across model scales and architectures.

Keywords: syntax, pragmatics, instruction-tuning

1. Introduction

Speech acts are actions performed through language (Searle, 1969). Depending on the communicative intent of the speaker uttering its words in some linguistic form, it can perform different types of actions. For instance, an assertive speech act commits the speaker to what it is has been said (e.g., *The door is on the right*), while a directive speech act is a request for an action/information (e.g., *Could you tell me the capital of Italy?*). The English language has different syntactical categories to express these speech acts: declaratives, interrogatives, exclamatives and imperative among others. Each syntactical type is characteristically used to perform a specific kind of speech act. Declaratives, for instance, tend to express assertives (statements), whereas interrogatives request for information. However, this correlation is not perfect and misalignments between surface form and communicative intent give rise to indirect speech acts. Declarative sentences like the following are interpreted as polite requests in speech acts theory:

I would appreciate it if you could tell me the capital of Italy.

Similarly, rhetorical questions have the form of interrogatives but are equivalent to assertives for the communicative intent:

Who would be against peace?
($\approx$ *Nobody is against peace*)

Transformer-based Language Models (LMs) represent a huge step towards machines capable of reproducing human language. Since their first steps (Devlin et al., 2019; Radford et al., 2019), they demonstrated robust understanding of different linguistic phenomena from pre-training (Warstadt et al., 2020). However, it is due to instruction-tuning that LMs have found their path to undisputed success in language applications (Brown et al., 2020). In pre-training, LMs statistically learns how language sequences are structured and how to complete prompts with tokens in line with the rules of grammar (approximately morphology, syntax and semantics). After this, instruction-tuning (Ouyang et al., 2022) is the fine-tuning phase where LMs statistically learn how not only to complete a sentence but also to fulfil the intent of the user involved in the conversation (pragmatics). In other words, they learn to communicate. Since humans communicate with language technology mostly to solve tasks, by design the training focuses on a specific type of directive speech acts, i.e., instructions.

Through the lens of direct and indirect speech acts, our research question to investigate instruction-tuning in LMs is the following:

RQ: *Does a pragmatic dimension emerge from Instruction-Tuning in LMs and where?*

2. Related Works

Having LMs demonstrated near-saturating performances on formal language tasks, new interest for pragmatics is emerging in the NLP community (Ma et al., 2025) and beyond (Gubelmann, 2024). During instruction-tuning LMs learn to encode and satisfy the user intent, thus most NLP studies look

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4967–4973
11–16 May 2026.

for pragmatic abilities comparing base and instruct versions of the same LM (Wu et al., 2024; He et al., 2025). Zhao et al. (2024), among them, observe that first LMs' layers tend to be similar between the base and the instruct models, encoding general syntactic abilities. In middle and final layers, however, the instructed LMs' representations appear to be more focused on the task to perform.

A recent line of NLP studies identify linear concepts in LMs' latent space (Marks and Tegmark, 2024). With a focus on instruction-tuning, Heo et al. (2025) observes how LMs internally know when they are successfully following an instruction, highlighting the role of prompt engineering. Stolfo et al. (2025) extracts instructional steering vectors, showing their effectiveness and demonstrating these can be transferred from instruction-tuned models to base models.

3. Experiment

3.1. Dataset and Models

This study hypothesises that base LMs are focused on the surface (syntactical) form of input sequences, whereas instruct LMs on the communicative (i.e., pragmatic) intent behind this linguistic form. To test this hypothesis, we create a dataset where each data point is distinguishable according to both syntactical and pragmatic dimensions. The resulting dataset would allow us to understand which dimension is prioritised by the model before and after instruction-tuning. Regarding syntactical form, this analysis is limited to declarative and interrogative sentences; regarding the intent, to statements and requests.

The first step is the creation of a seed of instructions, distributed across 10 tasks: part-of-speech tagging, translation, arithmetic and counting operations, paraphrasing and simplification, phonetic analysis (e.g., list vowels), text summarisation, spelling correction, word relations (e.g., syonyms/antonyms), format conversion (e.g., JSON, XML, etc.), and concept explanation. For each task, 10 sentences are created following the same biclausal structure: one clause for the instruction (e.g., 'Translate in Italian'), and one for the object of the instruction (e.g., the sentence to translate). Secondly, these instructions are modified in two directions: syntactical form, resulting in declarative and interrogative sentences, and communicative intent, resulting in requests or statements. The resulting total of 400 sentences falls into these four categories: declaratives with request intent and declaratives with assertive intent; interrogatives with request intent and interrogatives with assertive intent. An example of the configuration is reported in Table 1. The two steps for dataset creation are performed by Claude Opus 4.1, with manual supervision to ensure syntactic diversity among the same-group sentences and clarity in the communicative intent.

	Assertive	**Request**
Declarative	'Teacher' and 'concepts' are the nouns in 'The teacher explained difficult concepts'	Finding the nouns in 'The teacher explained difficult concepts' would be beneficial
Interrogative	Isn't it obvious that 'The teacher explained difficult concepts' contains nouns?	Could you extract the nouns from 'The teacher explained difficult concepts'?

Table 1: **Syntactical Forms and Speech Acts**. From the instruction 'Extract the nouns from 'The teacher explained difficult concepts" are derived one *declarative* and one *interrogative* with assertive intent; and one *assertive* and one *requests*.

The reference model for this study is Qwen2.5, with parameters ranging from 1.5B to 7B. Along this model, Gemma-3-1B and Llama3.2-3B are tested. For all these models, both the base and the instructed versions are analysed.

3.2. Method

To test our hypothesis about base LMs focusing on the syntactical level and instruct LMs on the pragmatical one, we need to extract and visualise meaningful representations for each sentence of our dataset.

As model representations, we extract hidden states for the final token position of each sequence, encoding the contextual representation of the entire sequence (Marks and Tegmark, 2024; Bertolazzi et al., 2025). To find emergent dimensions of variation between the two models, we applied Principal Component Analysis (PCA) on the hidden states. The hidden states for all four classes of sentences are fitted in the PCA transformation jointly, ensuring the resulting 2D projections capture the primary axes of variation. This procedure is performed layer by layer, allowing for tracking the emergence of the clusters. As a proxy of similarity between groups of sentences, we employ the centroid distance among sentence groups: high values mean the groups are clearly distinct in LMs' representations. The centroid distance between classes with different syntactical structures (declaratives vs interrogatives) measures the relevance of the syntactic dimension; the distance between classes with different pragmatic intents (assertive vs requests) measures the relevance of the pragmatic dimension.

4. Results

The results for Qwen2.5-7B are reported in Figure 1. The 2D PCA plot for the base model (on the left) shows two clusters driven by the syntactic dimensions of the sentences, meaning the model may lack the ability to differentiate sentences according to their communicative intent (i.e., pragmatic understanding). For the instruction-tuned model (PCA on the right) the distinction follows the communicative intent of the sequences: instruction-tuning teaches the model to prioritise the communicative intent over syntactical distinctions. It is worth noting that the groups declarative_request vs. interrogative_request, the most separated groups in the base model, become the most similar after instruction-tuning (65% absolute reduction in terms of centroid distance). This result explains how the instruction-tuned model correctly interprets indirect speech acts: declaratives like *I'd appreciate corrections for 'seperate'* are interpreted as pragmatically equivalent interrogatives (*Can you correct 'seperate'?*), triggering a helpful response despite the syntactical difference.

In terms of absolute centroid distance, the syntactical clusters in the base model are linearly separable (mean centroid distance 196.4), while the groups in the instruction-tuned models are closer in the 2D space (lower mean of 121.1). This could lead the instruct model to misinterpret assertives as requests. Answering rhetorical questions is a phenomenon that could, indeed, be observed even in large proprietary models.

Figure 2 reports the centroid distance according to the syntactic and pragmatic dimensions for Qwen2.5-7B. In the base model, the syntactic dimension accounts for more variance than the pragmatic dimension, whereas this relationship reverses in the instruction-tuned model (pragmatic > syntax). The same (normalised) relative centroid distance is used to track the transition to pragmatic dominance over the syntactic dimension in instruct models. Fig. 3 shows the crossover from the syntactical to the pragmatic dimension happening around layer 16 (Qwen2.5-7B-Instruct). This is in line with both linguistic theory, where syntax precedes pragmatic interpretation, and the multi-task abilities of instructed-LM by Zhao et al. (2024).

The observations reported hold for the other models in this analysis. The core insight—that instruction-tuning teaches models to prioritise communicative intent over syntactic form—is a consistent phenomenon across different Qwen2.5 model sizes (1.5B, 3B and 7B), as illustrated by Figure 4, and different models' families, as in Figure 5.

5. Conclusion

This work focuses on the high-level properties emerging from instruction-tuning, inspected through the lens of models' internal representations. We observe a clear pattern on how instruction-tuning modify inputs processing: before instruction-tuning, models prioritise the syntactical form of the input sequences; after instruction-tuning, LMs focus on the pragmatical intent under this surface form. This finding is confirmed across different models' sizes and families. Furthermore, we localise this change as occurring in middle layers of the LMs, in line with the hierarchial organisation of human language processing, where syntax procede the pragmatic understanding.

6. Bibliographical References

Leonardo Bertolazzi, Sandro Pezzelle, and Raffaella Bernardi. 2025. How language models conflate logical validity with plausibility: A representational analysis of content effects.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Reto Gubelmann. 2024. Large language models, agency, and why speech acts are beyond them (for now) ? a kantian-cum-pragmatist case. *Philosophy and Technology*, 37(1):1–24.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schuetze, Nima Mesgarani, and Jonathan Brennan. 2025. Large language models as neurolin-

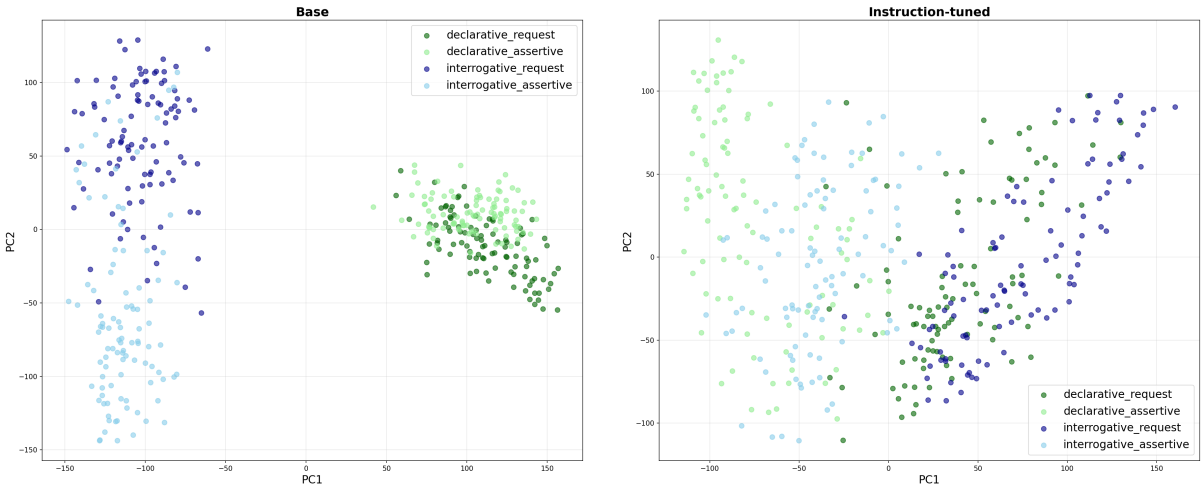

Figure 1: **PCA on Qwen2.5-7B** on the last layer for Base (left) vs Instruction-tuned (right) model on our dataset. Base model: interrogatives on the left and declaratives on the right. IT model: on the left assertive statements, on the right requests.

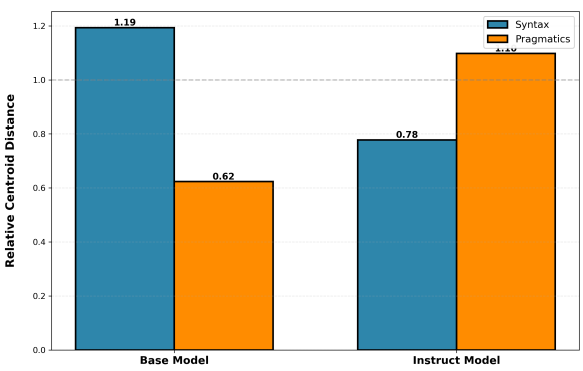

Figure 2: **Relative Centroid Distance** for Syntax and Pragmatics for the last layer of Qwen2.5-7B. The 'Syntax' bar reports the distance for syntactically separated groups (declaratives vs interrogatives), the 'Pragmatics' bar is between pragmatically separated ones (requests vs assertives). The values are normalised by the mean of the centroid distance among all groups.

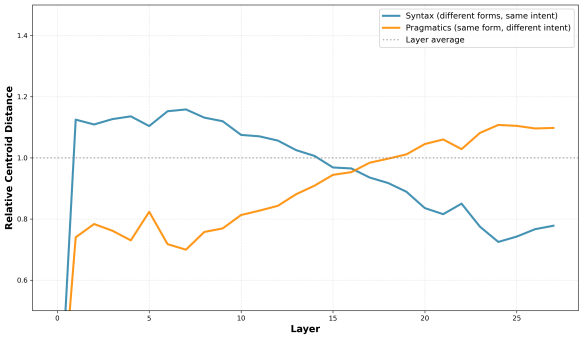

Figure 3: **Layerwise Relative Centroid Distance** for Qwen2.5-7B-Instruct. In early layers (1-15), syntax dominates over pragmatics, with sentences with different forms well-separated. After layer 16, the pragmatic dimension dominates (the sentences with different intents become more separated), while syntactic distinctions diminish.

guistic subjects: Discrepancy between performance and competence. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19284–19302, Vienna, Austria. Association for Computational Linguistics.

Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley You Ren, Andrew Miller, Udhyakumar Nallasamy, and Jaya Narain. 2025. Do LLMs "know" internally when they follow instructions? In *The Thirteenth International Conference on Learning Representations*.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex

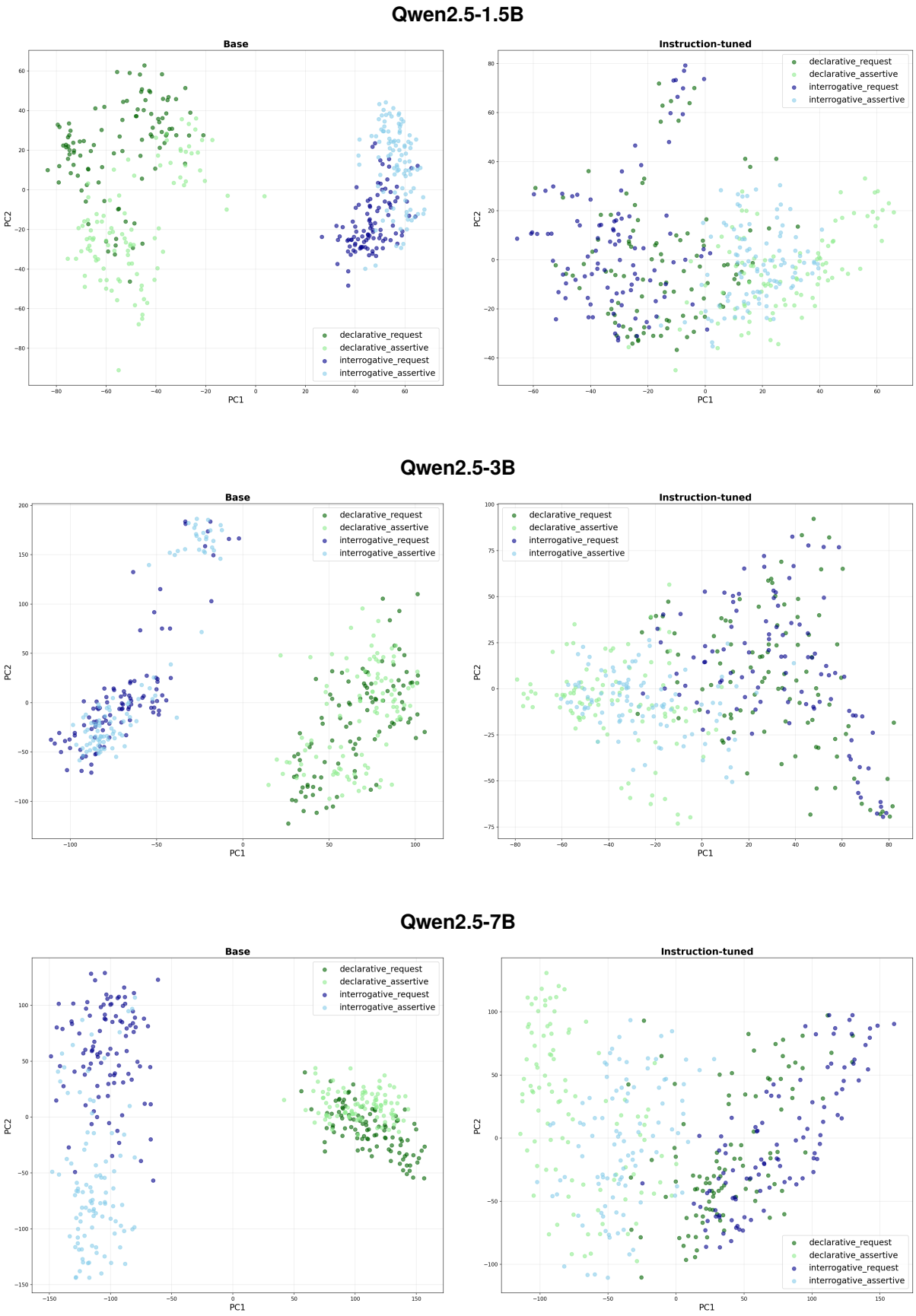

Figure 4: **PCA on Qwen2.5 with different parameters**: the switch from prioritising the syntactical dimension to the pragmatic dimension after instruction-tuning is confirmed across different Qwen2.5 1.5B, 3B and 7B.

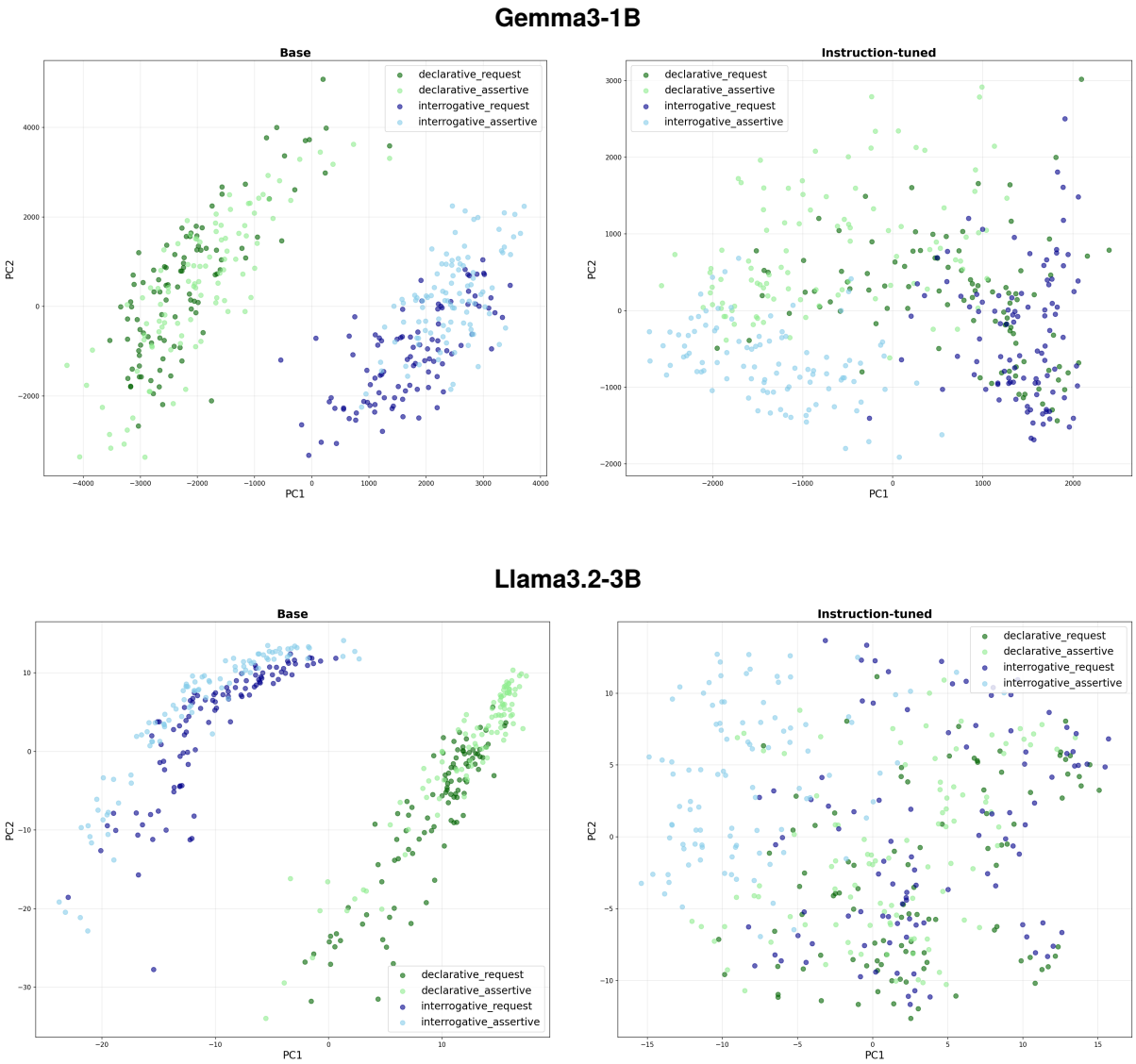

Figure 5: **PCA on different models' families**. For Gemma3-1B and Llama3.2-3B, it is possible to observe the syntactic dimension dominating the base model's representations, and the pragmatic one the instruction-tuned version.

Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2024. From language modeling to instruction following: Understanding the behavior shift in LLMs after instruction tuning. In *Proceed-*

ings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2341–2369, Mexico City, Mexico. Association for Computational Linguistics.

Zheng Zhao, Yftah Ziser, and Shay B Cohen. 2024. Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15195–15214, Miami, Florida, USA. Association for Computational Linguistics.