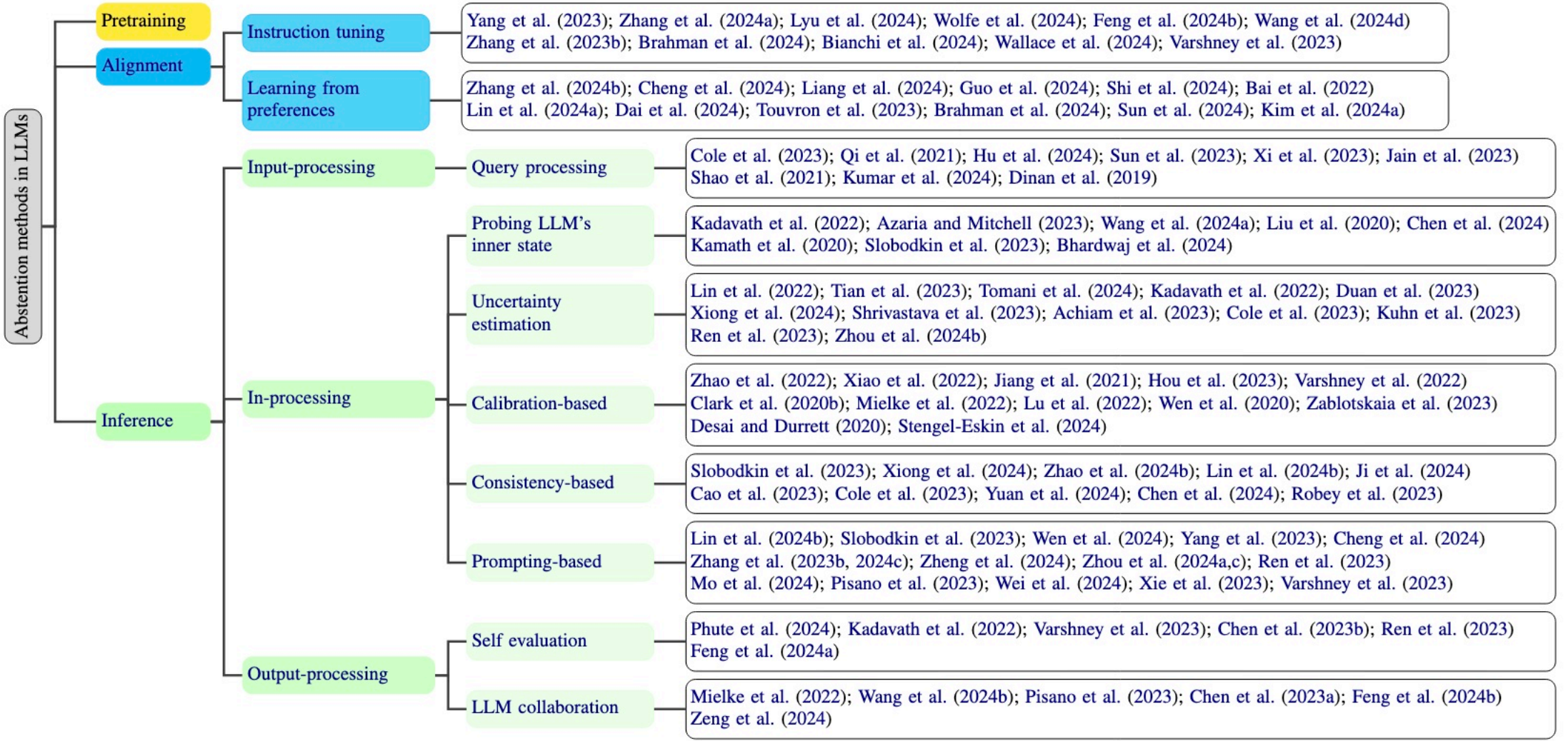
Abstention methods in LLMs
Pretraining
Alignment
Instruction tuning
Yang et al. (2023); Zhang et al. (2024a); Lyu et al. (2024); Wolfe et al. (2024); Feng et al. (2024b); Wang et al. (2024d)
Zhang et al. (2023b); Brahman et al. (2024); Bianchi et al. (2024); Wallace et al. (2024); Varshney et al. (2023)
Learning from preferences
Zhang et al. (2024b); Cheng et al. (2024); Liang et al. (2024); Guo et al. (2024); Shi et al. (2024); Bai et al. (2022)
Lin et al. (2024a); Dai et al. (2024); Touvron et al. (2023); Brahman et al. (2024); Sun et al. (2024); Kim et al. (2024a)
Inference
Input-processing
Query processing
Cole et al. (2023); Qi et al. (2021); Hu et al. (2024); Sun et al. (2023); Xi et al. (2023); Jain et al. (2023)
Shao et al. (2021); Kumar et al. (2024); Dinan et al. (2019)
In-processing
Probing LLM's inner state
Kadavath et al. (2022); Azaria and Mitchell (2023); Wang et al. (2024a); Liu et al. (2020); Chen et al. (2024)
Kamath et al. (2020); Slobodkin et al. (2023); Bhardwaj et al. (2024)
Uncertainty estimation
Lin et al. (2022); Tian et al. (2023); Tomani et al. (2024); Kadavath et al. (2022); Duan et al. (2023)
Xiong et al. (2024); Shrivastava et al. (2023); Achiam et al. (2023); Cole et al. (2023); Kuhn et al. (2023)
Ren et al. (2023); Zhou et al. (2024b)
Calibration-based
Zhao et al. (2022); Xiao et al. (2022); Jiang et al. (2021); Hou et al. (2023); Varshney et al. (2022)
Clark et al. (2020b); Mielke et al. (2022); Lu et al. (2022); Wen et al. (2020); Zablotskaia et al. (2023)
Desai and Durrett (2020); Stengel-Eskin et al. (2024)
Consistency-based
Slobodkin et al. (2023); Xiong et al. (2024); Zhao et al. (2024b); Lin et al. (2024b); Ji et al. (2024)
Cao et al. (2023); Cole et al. (2023); Yuan et al. (2024); Chen et al. (2024); Robey et al. (2023)
Prompting-based
Lin et al. (2024b); Slobodkin et al. (2023); Wen et al. (2024); Yang et al. (2023); Cheng et al. (2024)
Zhang et al. (2023b, 2024c); Zheng et al. (2024); Zhou et al. (2024a,c); Ren et al. (2023)
Mo et al. (2024); Pisano et al. (2023); Wei et al. (2024); Xie et al. (2023); Varshney et al. (2023)
Output-processing
Self evaluation
Phute et al. (2024); Kadavath et al. (2022); Varshney et al. (2023); Chen et al. (2023b); Ren et al. (2023)
Feng et al. (2024a)
LLM collaboration
Mielke et al. (2022); Wang et al. (2024b); Pisano et al. (2023); Chen et al. (2023a); Feng et al. (2024b)
Zeng et al. (2024)