Bayesian models have played a central role in cognitive science for over three decades and provided a widely used framework for studying perception, learning, and reasoning. They are sometimes presented as capturing the universal structure of cognition (Griffiths et al., 2011) or as enabling reverse engineering of the mind (Griffiths, Chater, & Tenenbaum, 2024). Despite their success, debate persists about what Bayesian models explain about human mind/brain. Some argue that Bayesian models are purely computational and carry no ontological commitments (e.g., Griffiths et al., 2012), while critics note that they are often treated as implying that the mind literally performs Bayesian inference (e.g., Bowers & Davis, 2012). Others claim that the evidence is insufficient for probabilistic mental representations and favor an instrumental interpretation (e.g., Block, 2018), while proponents argue that the success of Bayesian models supports their psychological reality (e.g., Rescorla, 2025).

Much of the debate is a dispute about how Bayesian models are used in practice, with each side attributing a stance to "Bayesian researchers" and the other side rejecting that attribution. This is illustrated by the commentaries on Jones and Love's (2011) *Behavioral and Brain Sciences* article. Although leading scholars disagree on many points, yet converge on a single conclusion: there is a widespread confusion about what explanatory commitments Bayesian models are meant to express. While prior work offers important theoretical analyses, there remains no standardized and scalable method for empirically characterizing explanatory stance in the literature at scale. Our goal in this work is to analyze the language in academic articles to uncover the implicit assumptions authors make when using Bayesian concepts.

Here we present a large-scale empirical analysis of explanatory stances in Bayesian cognitive science using a theory-driven annotation framework. We introduce a codebook for classifying "assumption-bearing quotes": sentences that make claims about Bayesian modeling/inference as explanations of behavioral, cognitive, computational, or neural processes (Table 1). The codebook maps these claims onto two continuous and mirroring scales: "realism" and "instrumentalism". Realism treats Bayesian models as describing the actual mechanism of the human mind, while instrumentalism considers them as useful tools without ontological commitments. We adopt this framework because it reflects the current debates in Bayesian cognitive science, subsumes previous theoretical distinctions in the literature, and aligns with the broader realism-instrumentalism tradition in the philosophy of science.

To validate this framework prior to the main analysis, 250 quotes from 15 articles were extracted and hand-annotated by an expert. The data were split into training (80%) and test (20%) sets, and three large language models (LLMs) were prompted with the codebook and training examples. LLM–human agreement on ordinal categorical assignments was moderate-to-good for Gemini (Krippendorff's α=.65-realism; α=.70-instrumentalism), OpenAI (α=.58-realism; α=.58-instrumentalism), and Claude (α=.63-realism; α=.63-instrumentalism). For continuous scores, agreement with human ratings was high for Claude (*ICC*(2,1)=.90-realism & instrumentalism; *ICC(*2,k)=.95) and moderate-to-high for OpenAI (*ICC(*2,1)=.70-realism & instrumentalism;

ICC(2,k)=.82). Overall, these results indicate that the models could apply the codebook reliably to support the analyses.

Using this framework, we analyzed 6,941 assumption-bearing quotes from 211 peer-reviewed cognitive science articles. We built the corpus by seeding 50 APA database results from the relevant Bayesian modeling keyword search and screening abstracts for substantive engagement with Bayesian modeling (empirical applications, computational modeling, or theoretical discussion; not mere statistical use). We then expanded the set using ResearchRabbit's citation-network map with additional abstract screening. Across the three models, annotation agreement was moderate-to-good for ordinal categorical subcategories (α=.73-for realism; α=.72-for instrumentalism) and high for continuous scores (ICC(2,1)=.75-for realism; ICC(2,1)=.74-for instrumentalism; ICC(2,k)≈.90-across both scales). We averaged the three models' continuous scores for downstream analyses; a later prompt-optimization pipeline reproduced the same reliability and domain patterns.

These analyses presented three important results. First, intraclass correlation analyses revealed substantial within-paper heterogeneity: between-article variance (*SD* = 14.7) was smaller than within-article variance (*SD* = 20.8), with only 33.5% of variance attributable to between-paper differences (*ICC* = .34) (Figure 1). This pattern indicates that explanatory commitments vary substantially within articles: individual papers often contain conflicting realist and instrumentalist assumptions in their language.

Second, explanatory stance differed systematically by domain. Here, domain refers to the level of process a paper primarily targets: lower-level perceptual and motor processes versus higher-level cognitive processes. We fit a linear mixed-effects model predicting quote-level realism from domain level (high vs. low), article type (computational, experimental, and theoretical), and publication year. The model included random intercepts for articles to account for multiple quotes per paper. Articles focused on lower-level were more realist than those focused on higher-level ($\beta = 7.60$, $SE = 2.03$, $t = 3.75$). Including domain level improved model fit ($\chi^2(1) = 13.63$, $p < .001$) (Figure 2). This effect remained while controlling for article type and publication year ($\beta = 9.57$, $SE = 2.05$, $t = 4.68$). Article type did not improve the model fit ($\chi^2(2) = 5.73$, $p = .057$), as publication year did not reliably predict paper's stance (realism: $\chi^2(1)=3.56$, $p=.059$; instrumentalism: $\chi^2(1)=3.44$, $p=.064$).

Third, averaging quote scores within articles and comparing realism/instrumentalism scores enabled us to classify 160 (75.8%) articles as instrumentalist-leaning, 39 (18.5%) as realist-leaning, and 12 (5.7%) as mixed/ambivalent (weighted by quote volume: 5,279 (76%) instrumentalist-leaning quotes, 1,332 (19.2%) realist-leaning quotes, and 330 (4.8%) mixed/ambivalent quotes). This result suggests that cognitive science literature shows an overall instrumentalist tendency, both at the article level and at the level of individual quoted claims.

By turning a long-standing theoretical dispute into a measurable empirical target, this work provides a quantitative map of how Bayesian explanations are framed across cognitive science

and offers a scalable method for discourse/stance analysis of explanatory commitments using LLM-assisted annotation. We have achieved three main results through our work: (1) individual research articles frequently mix realist and instrumentalist language, (2) explanatory stance shifts between high- and low-level cognition fields, and (3) there is an instrumentalist tendency in the literature overall. These results provide empirical support for the previous theoretical discussions regarding the "confusion" about the role of Bayesian models in explaining human mind and also the tension between explicit and implicit assumption within research articles using Bayesian modeling.