Appendix

A. Supplementary results on errors of LLM inference and prior constraints

In this section, we present the count of incorrect causal statements inferred by GPT-4 along with the erroneous prior constraints across various backbone algorithms and distinct observed data sizes of eight datasets. Figure 4 delineates the results pertinent to the hard constraining approaches, while Figure 5 elucidates those relevant to the soft constraining approaches.

B. Trends in Learned DAG Quality Over Iterations

This section outlines the iterative trends of scaled SHD (aiming for a decrease, denoted as SHD↓) and True Positive Rate (aiming for an increase, denoted as TPR↑) for various backbone algorithms across eight datasets, as depicted in Figure 6.

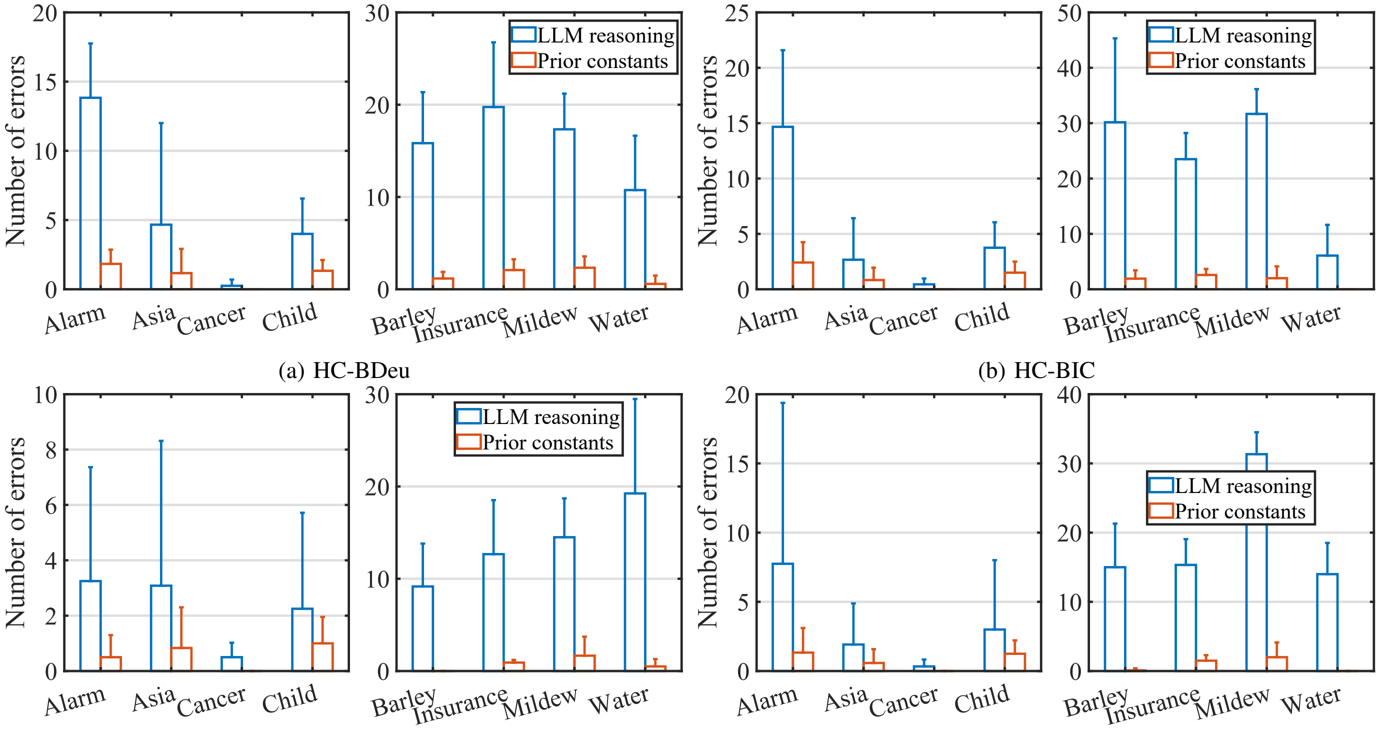
C. Trend of Constraints Derived from LLM Over Iterations

This section discusses the trend in the number of total and erroneous prior constraints derived from various backbone algorithms on eight datasets, as illustrated in Figure 7.

The following key observations emerge from the analysis:

- **Increasing Total Prior Constraints with Few Errors:** As the iterations progress, the number of total prior constraints sees a rise, while the increase in erroneous constraints is considerably smaller. This trend highlights the robust capability of ILS-CSL in generating high-quality, reliable constraints, enhancing the overall efficiency and reliability of the causal discovery process.
- **Occasional Decrease in Constraints:** Despite a general increase, some iterations exhibit a decrease in the number of prior constraints. This phenomenon is attributed to the same statistical artifact discussed in Appendix B. Some cases conclude in earlier iterations, leading to a varied set of statistical points across consecutive iterations, thereby affecting the total count of constraints.

These observations further affirm the effectiveness of ILS-CSL in consistently generating high-quality constraints throughout the iterations.



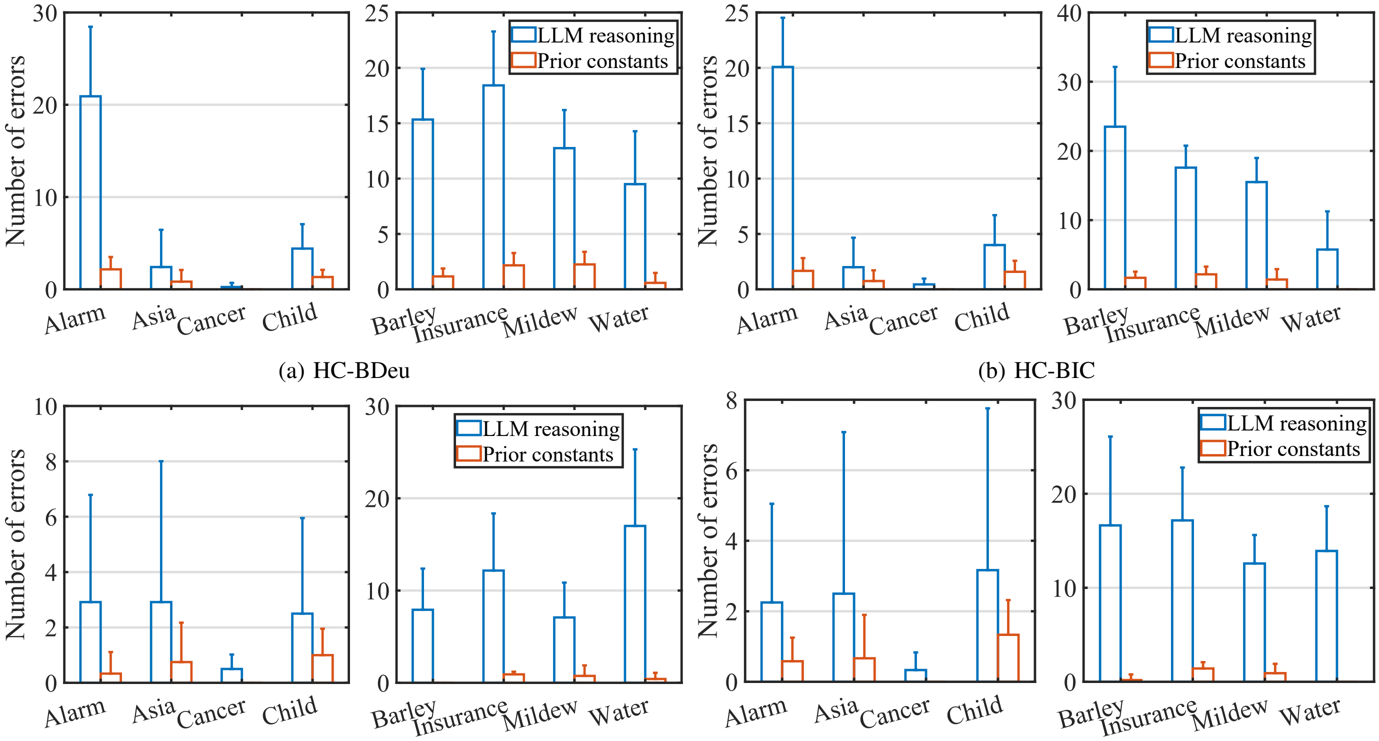
(a) HC-BDeu

(b) HC-BIC

(c) MIONBSx-BDeu

(d) MIONBSx-BIC

Fig. 4: Number of erroneous LLM inference and prior constraints during ILS-CSL related to hard constraining approaches on various algorithms and datasets.



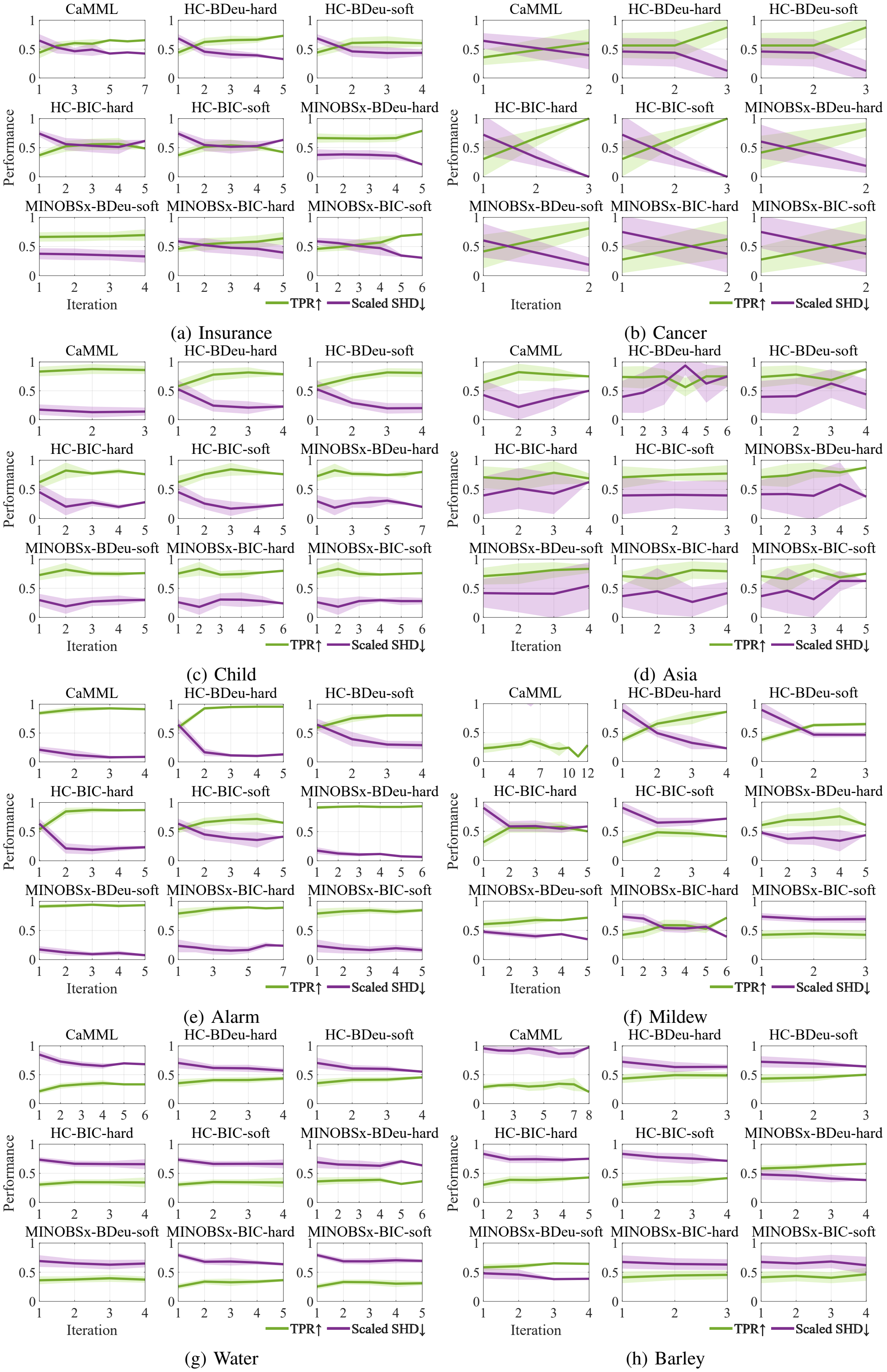
(a) HC-BDeu

(b) HC-BIC

(c) MIONBSx-BDeu

(d) MIONBSx-BIC

Fig. 5: Number of erroneous LLM inference and prior constraints during ILS-CSL related to soft constraining approaches on various algorithms and datasets.



(a) Insurance

(b) Cancer

(c) Child

(d) Asia

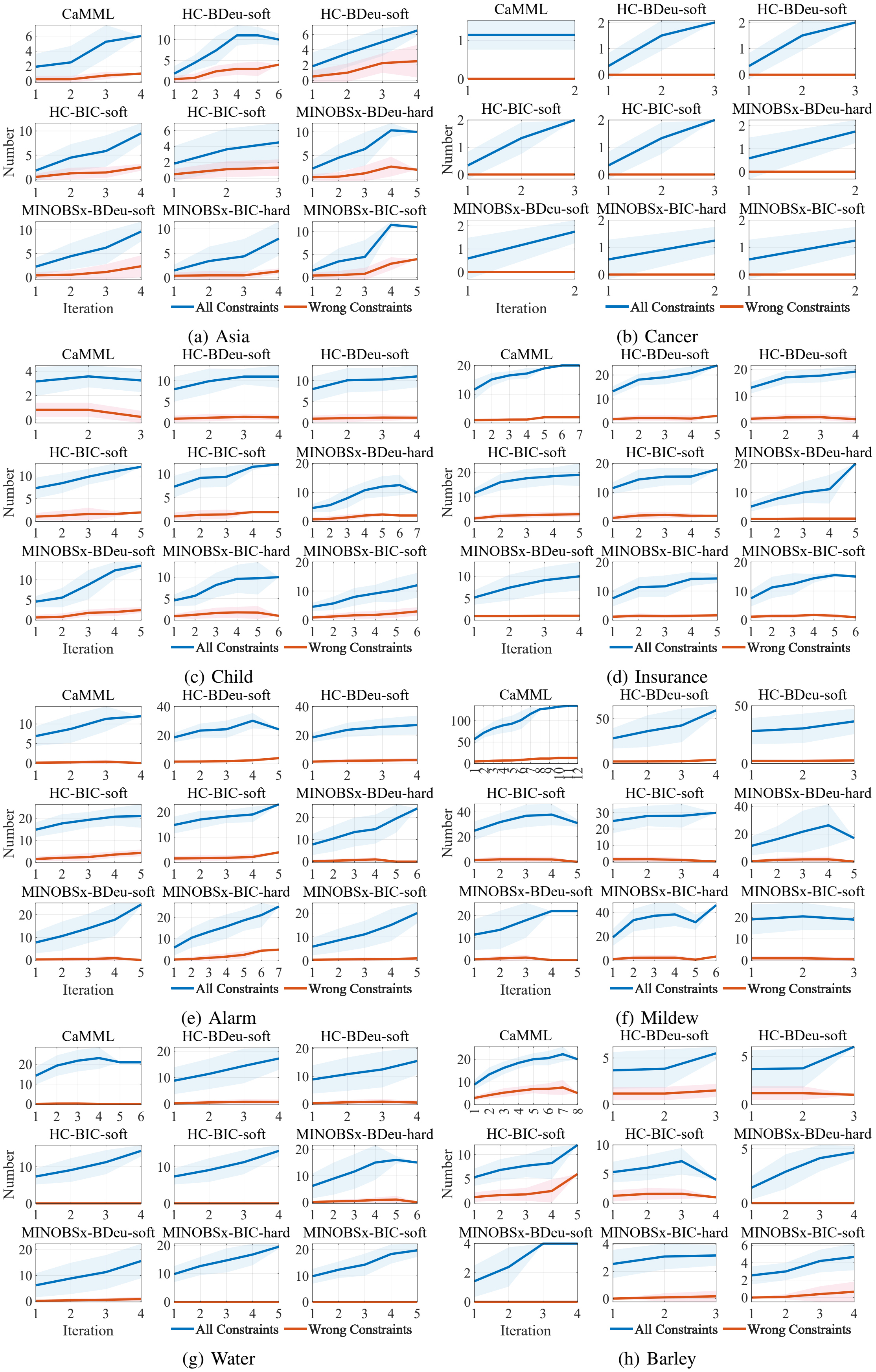
(e) Alarm

(f) Mildew

(g) Water

(h) Barley

Fig. 6: TPR↑ (green line) and scaled SHD↓ (purple line) alongwith derivations (colored area) in ILS-CSL with various algorithms on various datasets.



(a) Asia

(b) Cancer

(c) Child

(d) Insurance

(e) Alarm

(f) Mildew

(g) Water

(h) Barley

Fig. 7: Number of total (blue line above) and erroneous (red line below) prior constraints along with derivations (colored area) in ILS-CSL with various algorithms on various datasets.