# An exploratory study of hybrid explanation of Bayesian network derived likelihood ratios for assessing the value of evidence

Jeroen Keppens[1,*,†]

[1]King's College London, Department of Informatics, 30 Aldwych, London WC2B 4BG, UK

**Abstract**

Within Bayesian approaches to evidential reasoning, the likelihood ratio is the standard metric to assess the value of evidence in differentiating support for competing hypotheses. Such valuations of support can be used to determine expert testimony in court and to direct resources in investigations. Therefore, it is important that likelihood ratios computed by decision support systems are explainable. State-of-the-art foundation models have the potential to explain complex concepts, such as the provenance of a likelihood ratio. Unlike most traditional methods for explaining Bayesian inference, they are capable of engaging interactively with users answering specific questions in cohesive and compelling natural language. This paper presents a study to explore the potential of hybrid systems in producing faithful, stable, and interpretable explanations of likelihood ratios computed by means of a Bayesian network. Its contributions are as follows: (i) the paper defines an explanation problem for likelihood ratios computed from Bayesian networks, (ii) a number of symbolic reasoning techniques are adapted to produce supporting information a foundation model/large language model can base its answers on, and (iii) a qualitative study examines, by means of a number of carefully selected case studies, the performance of a foundation model with and without various forms of symbolically derived supporting information.

**Keywords**

Explainable AI, Evidential reasoning, Bayesian network, Likelihood ratio.

## 1. Introduction

A *Bayesian network (BN)* is commonly used to model evidential reasoning problems in law and forensics. A BN consists of a directed acyclic graph (DAG) over a set of variables and, for each variable $X$, a set of probability distributions, one for each combination of values of the parents of $X$. BNs provide a natural representation, rooted in probability theory, of the way hypothetical events and situations produce evidence. With a given BN, posterior probability distributions of variables in the problem, given observed or hypothetical information, can be calculated. In Bayesian evidential reasoning, such conditional probability distributions are used to calculate a *likelihood ratio (LR)*: the ratio of a likelihood representing the plausibility of evidence occurring under a null hypothesis, to the likelihood representing the plausibility of the evidence occurring under an alternative hypothesis. The LR is a means to quantify the value of specific piece or collection of evidence, in isolation of prior beliefs. This makes the LR a useful tool to prepare expert testimony on forensic evidence and directing investigative casework [1]. Within this field, a wide range of template BNs have been proposed in peer reviewed literature for analysis of different evidence types and combinations of evidence (e.g. the cross transfer problem, and the two trace problem).

A significant obstacle to the use of LRs derived from BNs is that BN models and the likelihood ratios derived from them can be difficult to understand. A substantial range of explanation techniques exist to explain BN models, evidence within BNs, and reasoning/inference within BNs [2]. Separate from developments in BN explanation, there is a body of work on interpreting LR values and validation of domain-specific approaches to compute LRs [3]. To the best of our knowledge, there is no work explicitly dedicated to producing verbal explanations of how a LR is derived from inference in a BN, as opposed to interpreting the LR value itself or explaining probabilistic influence more generally.

*RELAF'26: Reasoning with Evidence in Law Enforcement and Forensics, Workshop at ICAIL 2026, Singapore.*

jeroen.keppens@kcl.ac.uk (J. Keppens)

This paper explores how hybrid approaches, combining symbolic calculations with a state-of-the-art foundation model to produce cohesive and compelling explanations for a likelihood ratio using natural language. The paper introduces a novel LR explanation problem. It adapts a number of symbolic techniques to produce information to support a foundation model in explaining LRs. GPT-5.4, a state-of-the-art foundation model at the time of writing, in combination with different types of supporting symbolic information, is tasked to produce LR for several carefully selected case studies, and the results are critically evaluated.

## 2. Background

### 2.1. Bayesian networks for evidential reasoning

A BN is a type of probabilistic graphical model, representing the joint probability distribution over a (potentially large) number of variables. The graphical model of a BN is a DAG, where each node corresponds to a variable. The edges of the DAG define the dependencies between variables in such a way that the probability distribution of each variable only needs to be defined conditioned on the values of its parents in the DAG[1]. This simplifies the specification of complex joint probability distributions. With a suitable model structure (or DAG), the conditional probability distributions reflect knowledge the model's creator can derive from first principles, learn from data, or elicit expert opinion on.

A BN can be defined as a tuple $\langle \mathbf{X}, \mathbf{E}, \mathbf{P} \rangle$, consisting of the following components. A set of variables $\mathbf{X}$. Each variable has a domain of values, discrete or continuous. A set of directed edges $\mathbf{E}$ between pairs of variables in $\mathbf{X}$, such that $G = \langle \mathbf{X}, \mathbf{E} \rangle$ is a DAG. A set of conditional probability tables (CPTs) $\mathbf{P}$, one for each variable $X \in \mathbf{X}$. The CPT associated with a variable $X$ is a function:

$$P_X : \text{dom}(X) \times \prod_{Y \in \text{pa}(X)} \text{dom}(Y) \mapsto [0, 1]$$

such that for any fixed assignment $\vec{y} \in \prod_{Y \in \text{pa}(X)} \text{dom}(Y)$, the CPT defines a probability distribution for $X$:

$$\sum_{x \in \text{dom}(X)} P_X(x, \vec{y}) = 1 \tag{1}$$

Following the Markov property, a BN defined in this way represents the joint probability distribution over all variables in $\mathbf{X}$. Let $\mathbf{X} = \{X_1, \ldots, X_n\}$. Then,

$$\mathbf{P} : \prod_{X \in \mathbf{X}} X \mapsto [0, 1] : \mathbf{P}(X_1, \ldots, X_n) = \prod_{X \in \mathbf{X}} P_X(X | \text{pa}(X)) \tag{2}$$

Figure 1 shows an example of a simple evidential reasoning BN modelling a one-trace transfer problem proposed by Garbolino and Taroni [4]. It models a scenario of a violent crime, where a trace may have transferred from the perpetrator to the victim, and a suspect has been identified. The BN consists of four variables, all propositions. The possible values of each variable $X$ are $x$ (true) and $\overline{x}$ (false).

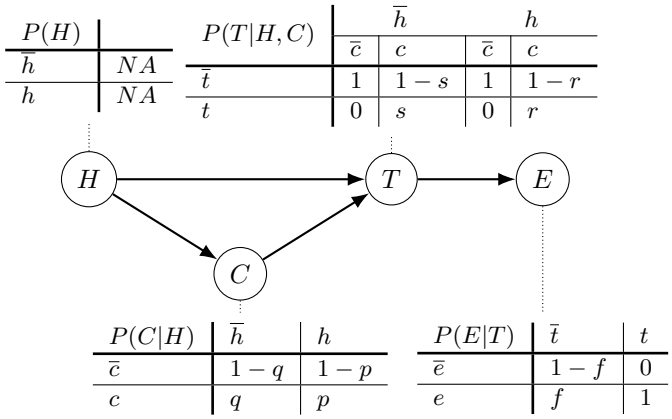
Each variable comes with a CPT that consists of a probability distribution for that variable, for each set of value assignments of its parents[2]. As is common in evidential reasoning BNs, these CPTs contain parameters to be instantiated with the numbers that apply to the case. With this almost fully specified BN, conditional probability distributions, such as the probability of acquiring evidence $e$ given that hypothesis $h$ is true ($P(e|h)$), can be calculated more easily because the joint probability distribution over all variables $P(H, C, T, E)$ simplifies to $P(E|T) \times P(T|H, C) \times P(C|H) \times P(H)$. Thus,

$$P(e|h) = \frac{\sum_T \sum_C P(e|T) P(T|hC) P(C|h) \cancel{P(h)}}{\cancel{P(h)}} = pr + fp(1 - r) + f(1 - p)$$

Similarly, $P(e|\overline{h}) = qs + fq(1 - s) + f(1 - q)$.

[1] For conciseness, the concept of *d-separation*, which defines the notion of conditional dependencies between variables not covered here, but the paper can be understood without it.

[2] No probabilities are provided for $P(H)$ because, as we shall see, these are not needed.




| Variable | Meaning |
|---|---|
| $H$ | *Hypothesis*: The suspect is the perpetrator of the crime |
| $C$ | There was contact between suspect and victim |
| $T$ | Traces were transferred from suspect to victim |
| $E$ | *Evidence*: A trace recovered from the victim matches the suspect |

**Figure 1:** Simple one-trace transfer BN (from [4], Section 5). The CPTs depend on parameters $p$, $q$, $s$, $r$, and $f$. Prior probabilities for $P(H)$ are not provided.

## 2.2. The likelihood ratio

In a Bayesian evidential reasoning approach, the value of evidence is assessed in the form of a likelihood ratio (LR) [5]. Let $e$ be a piece of evidence, and $h_0$ and $h_a$ (herein named the null hypothesis and the alternative hypothesis respectively) be an exhaustive pair of mutually exclusive hypotheses (i.e. $P(h_0 \cup h_a) = 1$ and $P(h_0 \cap h_a) = 0$). The hypotheses could, for example, be the prosecution hypothesis and the defence hypothesis respectively. The likelihood ratio of $e$ with respect to $h_0$ and $h_a$ equals:

$$LR(h_0 : h_a|e) = \frac{P(e|h_0)}{P(e|h_a)} \tag{3}$$

Intuitively, the LR appears to quantify how much better (or worse) an explanation the evidence is for one hypotheses as opposed to the other. According to Bayes' rule (in odds form):

$$\frac{P(h_0|e)}{P(h_a|e)} = \frac{P(e|h_0)}{P(e|h_a)} \times \frac{P(h_0)}{P(h_a)}$$

This formula states that the posterior odds of $h_0$ and $h_a$ (after observing evidence $e$) equal the LR multiplied by the prior odds of the hypotheses (before observing evidence $e$). Thus, the LR represents how the model requires our beliefs, expressed as odds, to change by observing the evidence. If the LR is close to 1, then the evidence does not have a meaningful impact on the odds. In other words, in that situation, the evidence has little value. If the evidence is greater than (smaller than) 1, then the evidence provides support for $h_0$ ($h_a$) as opposed to $h_a$ ($h_0$). The greater $LR(h_0 : h_a|e)$ (or conversely $LR(h_0 : h_a|e)^{-1}$), the stronger the level of support for $h_0$ ($h_a$).

Observe that the LR makes no judgement about the prior or posterior odds of the hypotheses within the model, only about the effect of the evidence. This makes the LR a valuable decision support aid in computing the rational implications of the observations, data, and knowledge available to them, in isolation of aspects of the case that do not concern their expertise.

For the purposes of working out numerical examples, let $p = 0.9$, $q = 0.01$, $s = 0.2$, $r = 0.9$, and $f = 0.02$. The BN of Figure 1 can be used to calculate the LR for the evidence that a trace recovered from the victim is matched to the suspect, with respect to the hypothesis that the suspect is the perpetrator as opposed to the hypothesis that the suspect is not the perpetrator: In this case $LR \sim 37$, which would indicate that, under these beliefs as specified by the parameters, the evidence provides moderate support for the hypothesis $h$.

## 2.3. A belief propagation model of inference

Let $\langle \mathbf{X}, \mathbf{F}, \mathbf{E} \rangle$ denote a factor graph, where $\mathbf{X}$ is a set of variables, $\mathbf{F}$ is a set of factors, and $\mathbf{E} \subseteq \mathbf{X} \times \mathbf{F}$ is the set of edges. A factor graph is a bipartite graph consisting of two disjoint sets of nodes: variables

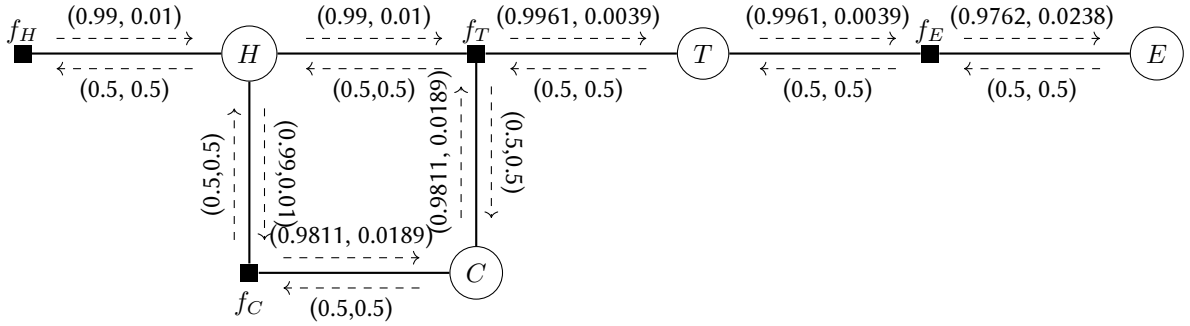



**Figure 2:** A solution to (5) as calculated via loopy belief propagation

($\mathbf{X} = X_1, \ldots, X_n$) and factors ($\mathbf{F} = f_1, \ldots, f_m$), where each edge $(X, f) \in \mathbf{E}$ connects a variable $X \in \mathbf{X}$ to a factor $f \in \mathbf{F}$. A direct factor graph representation of a BN is obtained by (i) representing each BN variable as a factor graph variable, (ii) representing each prior or conditional probability table as well as each observation as a factor, and (iii) connecting each factor to the variables in its scope. Each factor $f$ in the factor graph represents a function:

$$f : \prod_{X \in \mathcal{N}(f)} \text{dom}(X) \mapsto [0, 1] \tag{4}$$

where $\mathcal{N}(f)$ represents the neighbours of $f$ in the the factor graph. This function is defined by the probability distributions (i.e. the prior distribution, conditional distributions, or the observation) represented by the factor. Each factor graph edge $(X, f)$ is associated with two messages: a variable-to-factor message $m_{X \to f}$ and a factor-to-variable message $m_{f \to X}$, each of which is a function $\text{dom}(()X) \to \mathbb{R}$.

Let $p^\star_X$ represent exact marginal probability distributions for each $X \in \mathbf{X}$. The closest approximation representable by belief propagation (BP) is obtained by solving:

$$\min_m \quad \sum_{X \in \mathbf{X}} w_X D(b_X, p^\star_X)$$

subject to:

$$\begin{aligned}
b_X(x) &= \frac{1}{Z_X} \prod_{f \in \mathcal{N}(X)} m_{f \to X}(x) \qquad \forall X \in \mathbf{X}, \\
m_{X \to f}(x) &= \frac{1}{Z_{X \to f}} \prod_{g \in \mathcal{N}(X) \setminus \{f\}} m_{g \to X}(x) \qquad \forall (X, f) \in \mathbf{E}, \\
m_{f \to X}(x) &= \frac{1}{Z_{f \to X}} \sum_{x_{\mathcal{N}(f) \setminus \{X\}}} f\big(x, x_{\mathcal{N}(f) \setminus \{X\}}\big) \prod_{Y \in \mathcal{N}(f) \setminus \{X\}} m_{Y \to f}(x_Y) \qquad \forall (X, f) \in \mathbf{E}.
\end{aligned} \tag{5}$$

where $D$ is a divergence or distance metric between pairs of probability distributions, $b_X$ represents the approximate marginal probability distributions calculated by BP, and $Z_X$, $Z_{X \to f}$, and $Z_{f \to X}$ are normalisation factors ensuring that $\sum_{x \in \text{dom}(X)} b_X(x) = \sum_{x \in \text{dom}(X)} m_{X \to f}(x) = \sum_{x \in \text{dom}(X)} m_{f \to X}(x) = 1$. In practice, loopy belief propagation provides a heuristic for approximating solutions to this optimisation problem. While it is not guaranteed to converge, nor to attain a global optimum, solution are often close enough to exact solution for explanation purposes.

Figure 2 shows a solution in the form of BP messages for the optimisation problem specified in (5) in the form of a factor graph for the BN of Figure 1, assuming no observations/evidence. In this figure, factors are represented by squares and variables by circles. Approximate results can be calculated from these messages. For example, $b_T = m_{f_T \to T} \times m_{f_E \to T} = (0.9961, 0.0039)$, whereas $p^\star_T = (0.9899, 0.0101)$. As such, the BP messages largely explain the probability distribution with an error $< 0.01$.

Given a BP model of inference, the marginal distribution of a variable $X$ can be approximated by the product of the incoming BP messages $m_{f \to X}$:

$$P(x) = \frac{1}{Z_X} \prod_{f \in \mathcal{N}(X)} m_{f \to X}(x)\, \eta_X(x) \tag{6}$$

where $\eta_X$ is a multiplicative error term. If $\eta_X(x) \approx 1$, the BP model (almost) fully accounts for the inference. If $\eta_X(x) > 1$ ($\eta_X(x) > 1$), the BP model underestimates (overestimates) the exact probability.

For the purpose of explanation, we only retain factor-to-variable messages. A factor-to-variable message from a factor $f$ to a variable $X$ expresses the support offered by variables $\mathcal{N}(f) \setminus \{X\}$ for each of the values in the domain of $X$. This characterisation of information flows in the model is helpful in identifying the sources for differences in support under alternative hypotheses, especially where competing effects lead to counterintuitive results.

# 3. Problem

While there is a substantial body of literature on explaining BNs and the inference within them, there are to the best of our knowledge no approaches for explaining a LR computed by means of a BN. Nonetheless, prior to proposing extensions on existing work, it is important to understand its limitations. This section specifies the explanation problem this paper aims to address. Then, the section briefly identifies relevant existing work and the issues to be addressed with regards to the present explanation problem.

## 3.1. Problem specification

This paper presents a method to generate explanations for assessments of the value of evidence as quantified by a LR, based on likelihoods computed by means of a BN. Specifically, given a piece of evidence, an exhaustive pair of mutually exclusive hypothesis, and (optionally) a set of observations, *why does the piece of evidence support or not support one hypothesis over an another (under the known observations)?*

LRs can be computed by means of BNs in a variety of ways. To simplify the explanation problem somewhat, this paper assumes that the two hypotheses in the LR, as defined in (3), are the only two values of a single hypothesis variable in the BN. In other words, the BN contains a hypothesis node $H$ with $\text{dom}(H) = \{h_a, h_0\}$. In what follows, $h_0$ is referred to as the null hypothesis and $h_a$ as the alternative hypothesis. This ensures that the hypotheses considered for computing the LR constitute an exhaustive pair of mutually exclusive variables, as required by (3). Without loss of generality, the notation abstracts away any prior observations that may have occurred in the model.

## 3.2. Related symbolic explanation approaches

While an exhaustive review of all symbolic BN inference explanation methods is outside the scope of this paper, this subsection outlines the most relevant categories of approaches with a few (but not all) relevant systems.

Qualitative probabilistic networks abstract the conditional probability distributions specified in BNs into monotonic qualitative relations (e.g. observing X=x increases belief in Y=y), with efficient sign-propagation algorithms enabling purely qualitative belief propagation [6]. Although originally motivated by scalability concerns, such qualitative abstractions have also been shown to support verbal explanations of how observations affect posterior beliefs [7]. A fundamental limitation of sign-propagation is its coarseness: qualitative models capture only the direction, not the magnitude/significance, of belief change. While later developments work have refined these approaches significantly, explaining LRs requires quantitative reasoning with probabilities. Nonetheless, these qualitative approaches inform the present work.

Beyond qualitative belief propagation, several approaches provide *structural* explanations of Bayesian inference. These explanations focus on which variables, dependencies, and paths are relevant to an

inference rather than on the numerical strength of their influence. Approaches include (but are not limited to INSITE [8], EBI [9], and Bayes-ball [10]. Although these approaches are not designed as explanation frameworks per se, they provide principled criteria for determining which evidence can influence a variable of interest and could be adapted to structure explanations of likelihood-based reasoning.

A different line of work adopts an argumentative perspective. BANTER constructs explanations by framing posterior probabilities under evidence sets as arguments for or against hypotheses, organising evidential support and counter-support into an explicit argument structure [11]. Building on this idea, Timmer et al. generate argumentation models from Bayesian networks via an intermediate support-graph representation that captures all admissible inference paths [12]. Keppens combines this framework with qualitative belief propagation and exact inference to explain counter-intuitive posterior beliefs [13]. However, the reliance on qualitative abstractions limits its applicability to explaining likelihood ratios.

### 3.3. Foundation models

Conventional symbolic approaches to explanation typically rely either on custom representations, which require additional parsing by users, or on template-based text generation, which often results in repetitive and inflexible descriptions. In contrast, a large language model (LLM) can generate explanations that are cohesive, dynamically emphasise relevant factors, and adapt to user requirements and interaction context. Bayesian inference and LRs are conceptually complex and require the processing of substantial amounts of structured information. Larger LLMs exhibit capabilities that emerge with scale, enabling more effective handling of such tasks than smaller models. In particular, they are better able to integrate multiple dependencies, maintain coherence across longer reasoning chains, and generate more accurate and informative explanations.

Indeed, state-of-the-art foundation models can address complex reasoning tasks, including the probabilistic computations required to derive and explain LRs. This motivates examining whether symbolic explanation methods remain necessary. Nevertheless, despite often performing well on reasoning tasks, large language models exhibit well-documented limitations in formal reasoning, including fundamental and robustness-related errors that persist even in simple settings [14]. In mathematical domains, models demonstrate arithmetic inaccuracies, instability across equivalent problem formulations, and degraded performance with increasing numerical and compositional complexity [15]. These issues are particularly problematic for probabilistic inference tasks such as marginals and LRs by means of BN and likelihood ratio computation, as these operations require precise multi-step reasoning and strict adherence to probabilistic constraints.

## 4. Proposed approach

This paper proposed a hybrid, two-stage approach to explanation generation. First, key characteristics of the inference within the BN for the calculation of the LR are extracted using symbolic approaches. Second, the resulting symbolic explanation model is fed to a LLM with a prompt to produce a cohesive natural-language explanation.

A comparison of nuanced differences between alternative LLMs or foundation models is outside the scope of this study. Instead, a single advanced foundation model, GPT-5.4, is used as a representative instance of this class of systems. The objective is not to benchmark models against one another, but to investigate the extent to which a hybrid approach can improve the quality and reliability of generated explanations. Focussing on a single model allows for a more controlled analysis of how symbolic support influences explanation generation, without introducing confounding variation.

Large language models generate explanations by producing probabilistic continuations of text rather than executing formal inference procedures. As a result, their outputs are fluent and adaptable but may be unfaithful, numerically inaccurate, and sensitive to prompting. In particular, LLMs are prone to hallucination, generating plausible but incorrect statements, and provide limited guarantees of

correctness or calibrated uncertainty. Furthermore, generated explanations should be understood as post hoc textual artefacts rather than faithful traces of internal reasoning. These characteristics make LLMs well suited for rendering explanations in natural language, but unsuitable as standalone mechanisms for computing or validating probabilistic inference.

To assess the capabilities of a hybrid explanation system, this study considers three modalities, each defined by a distinct prompt design. In all cases, the prompt includes a description of the explanation task (including the hypothesis variable and evidence assignment), instructions on how to interpret the provided information, and a constraint that responses must be based exclusively on the information contained in the prompt or logically inferred from it. The modalities are as follows:

1. *Foundation-only* (not hybrid): The model is provided with a complete specification of the BN and any observations. Under this modality, the foundation model is required to perform all necessary computations and generate the explanation without additional support.

2. *Foundation and marginals* (hybrid): In addition to the information provided in the foundation modality, the model is given the marginals under both hypotheses, and the LR. This approach outsources calculations to a symbolic approach, ideally preventing errors and reducing the computational demands on the foundation model.

3. *Foundation, marginals, and partial BP model* (hybrid): In addition to the information provided in the previous modality, the model is given a partial BP model, providing structural and message-level information to support explanation generation. This information may be useful in identifying information flows within a BN.

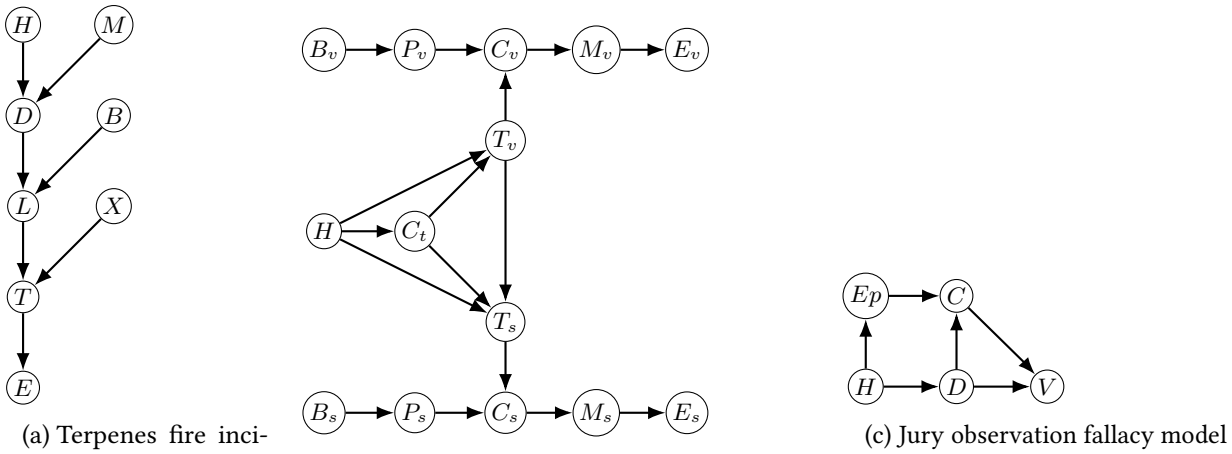
### 4.1. Posterior probability distributions

Let $X$ be a variable on a path between hypothesis node $H$ and evidence node $E$ in the BN. Comparing the posterior distributions $P(X|h_0)$ and $P(X|h_a)$ enables identification of the role of intermediate variables in explaining the likelihood ratio. The distance between posterior probability distributions $P(X|H = h_0)$ and $P(X|H = h_a)$ reflects the capacity of $X$ to support one hypothesis over another. It follows from (3) that, if the posterior distributions under the alternative hypotheses are close, the value of observing $X$ will be low. In other words, distances enable identification of what variables are important to an explanation of the likelihood ratio (including those without ordered domain).

### 4.2. Belief propagation model of inference

For variables $X$ with significant distance between $P(X|h_0)$ and $P(X|h_a)$ and connections to multiple other variables, it may not be straightforward (and therefore informative) to identify how other variables (particularly neighbouring ones) affect it. The posterior probability distribution of each variable in the BN depends on the posterior distributions of other variables in the model. The relationship between any pair of variables is often complex, as it is mediated by other variables in the BN and by observation-induced updates to their posterior distributions. This dynamic interdependence is not well suited to human-interpretable explanations of inference in a BN. Consequently, any effective explanation method must abstract away some of the nuances of BN inference, focusing on those aspects that are most salient.

Belief propagation (BP) provides a model for such an abstraction. Intuitively, when applied to BNs, BP represents inference via bidirectional message passing along the edges of the graph. These messages are not independent: each message is computed from incoming messages from neighbouring nodes, together with the prior and conditional probability distributions specified by the BN and any observed evidence. As a result, BP captures the influence that neighbouring variables exert on one another while enforcing local consistency. In tree-structured BNs, this is sufficient to yield exact posterior distributions. In the presence of loops, however, BP no longer guarantees global consistency, as messages may reinforce or contradict one another around cycles. Nevertheless, BP often produces accurate approximations




(a) Terpenes fire incident model [17].

(b) Cross-transfer trace model [18].

(c) Jury observation fallacy model [19].

**Figure 3:** Additional BNs used for evaluation

in such settings, as much of the inference is governed by local interactions that are captured by the message-passing scheme (particularly in graphs with weak loops or limited long-range dependencies) [16].

## 5. Evaluation

This section presents a structured empirical evaluation of the LR explanation techniques introduced in 4. The objective of this evaluation is to assess the extent to which a foundation model, both independently and in combination with the symbolic techniques described in 3.2, can produce suitable explanations, as defined by the evaluation criteria introduced later in this section. Natural -anguage explanations are inherently high-dimensional and, therefore, do not lend themselves to one-dimensional scoring. Accordingly, this study adopts a qualitative evaluation protocol based on a small number of carefully designed case studies, rather than a large-scale quantitative analysis.

### 5.1. Setup

#### 5.1.1. Evaluation criteria

The suitability of natural language explanations is determined by multiple criteria. Ideally, an explanation is faithful, complete, concise, robust, and interpretable. *Faithful* explanations accurately reflect the underlying dependencies and probabilistic computations of the model. *Complete* explanations capture all relevant dependencies and influences in the network. *Concise* explanations convey substantial information with minimal redundancy. *Robust* explanations remain consistent under repeated generation and under small perturbations of model parameters or observations. *Interpretable* explanations are consistent with domain knowledge and support correct reasoning by human users. This exploratory study focusses exclusively on the generation of faithful explanations, with a more comprehensive assessment left for future work.

#### 5.1.2. Case studies

For the purposes of evaluation, four BNs models on evidential reasoning will be used. Each of these has a number of interesting features. The first model is the simple example introduced in Figure 1 and used as an example of key concepts in this paper. Because this model is so small, it is easy to understand and verbalise the reasoning occurring within it.

To illustrate the nature of the other BNs, their structure is presented in Figure 3[3]. The second model proposed by Biedermann et. al. analyses fire incidents potentially caused by arson by means of terpenes ($H$) from trace evidence ($E$) [17]. This 8 variable model does not contain any loops. Instead, it consists of a longer chain of cause-and-effect relationships with an alternative explanation at each level. Under its default parameterisation, it results in an LR close to 1. In other words, the evidence provides no meaningful value in differentiating between null/prosecution and alternative/defence hypotheses.

The third model is a BN proposed by Aitken et. al. to analyse the cross-transfer of trace material, from victim to suspect and from suspect to victim [18]. It consists of 14 variables and a number of interesting features. The hypothesis ($H$) can lead to two items of evidence: a match of trace material retrieved from suspect with the victim ($E_s$), and a match of trace material retrieved from the victim with the suspect ($E_v$). Prior observation of one of these affects the value of the other. The model also represents background contamination as a possible cause for a failure to match trace material.

The fourth model is the Jury observation fallacy model proposed by Fenton and Neil [19]. This 5 variable BN models a scenario where a jury member discovers evidence that a defendant committed prior offences ($E_p$), after a trial that returned a "not guilty" verdict (prior observation $V = \text{not\_guilty}$), and it is used to analyse the value of this evidence in assessing the guilt hypothesis ($H$). The model is tightly connected and contains two conflicting effects: a direct one where prior offences promote the guilt hypothesis ($H \rightarrow E_p$), and a less obvious (but potentially stronger) "explaining-away" effect (via $C \rightarrow V \leftarrow D$). The explaining away effect overrides the direct relation between $H$ and $E_p$, resulting in a counterintuitive LR.

## 5.2. Experiments and results

### 5.2.1. Foundation-only

In our experiments, we have observed that the foundation-only modality is capable of producing compelling explanations for an LR computed from a BN. This shall be revisited below, where the two hybrid modalities are considered. The main limitation of using a foundation model only is that the symbolic reasoning it attempts to perform, while often correct, is unreliable and fails without indication of doing so. As the foundation-only modality fails on faithfulness, its evaluation focusses on this aspect exclusively.

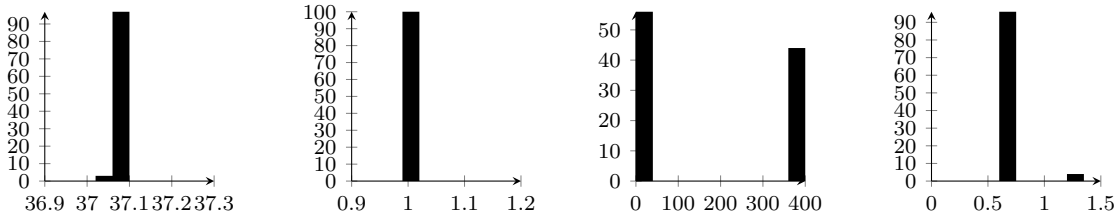
As noted in 3.3, prior studies suggest that foundation models cannot be relied on to perform complex reasoning tasks reliably. To verify whether these findings extend to LR analysis by means of BNs, GPT-5.4 was tasked to perform 100 repetitions of a specific LR calculation. To replicate normal explanation conditions, the full "foundation model only" prompt was provided and the model allowed to verbalise its reasoning process (as not doing so can impact results). The final result was then stripped from the output. The results of the calculations are presented in Figure 4.

For each run, there are small variations in the LRs that are produced. However, the variations tend to be too small to notice in the distributions. Apart from small rounding errors, a majority of the produced LRs is inherently correct, apart from minor rounding errors. Nevertheless, as LLMs produce these LRs by means of probabilistic continuations of text rather than executing formal inference procedures, these errors are likely to grow as model sizes increase.

More significantly, subfigures 4c and 4d exhibit bimodal distributions of outcomes, with the smaller peak representing a collection of incorrect results. Both scenarios involve a prior observation, though the nature of the underlying error is different in each scenario.

In the cross transfer model, nearly half of the LRs are in the 390–391 range, when the correct value is $LR(E_v, H|E_s = \text{true}) \cong 18.07$. The larger values correspond closely to the LR obtained without conditioning on the prior observation, i.e. $LR(E_v, H) \cong 390.4$. In these cases, the foundation model has effectively ignored the observation $E_s = \text{true}$ when performing the calculation. A typical explanation produced in this scenario includes the statement:

[3]Symbols used have been changed from some of the original publications to ensure the hypotheses variables are named $H$ and the evidence variables $E$.




(a) Example model results for $LR(E, H)$. Correct answer$\cong 37.06$.

(b) Terpenes evidence in fire incidents model results $LR(E, H)$. Correct answer$\cong 1.006$.

(c) Cross transfer model results for $LR(E_v, H|E_s = \text{true})$. Correct answer$\cong 18.07$.

(d) Jury observation fallacy model results for $LR(E_p, H|V = \text{not\_guilty})$. Correct answer$= 0.6620$.

**Figure 4:** Distribution of the results of 100 LR calculations by means of GPT-5.4 under "foundation model only" conditions (temperature=0.2).

> *This means that, within this Bayesian network model, finding the evidence ($E_v$=true) is about 390 times more probable if the suspect is the perpetrator than if the suspect is not the perpetrator, given that we have already observed ($E_s$=true). [...]*

In other words, the model computes a value consistent with $LR(E_v, H)$ but presents it as though it were $LR(E_v, H|E_s = \text{true})$. The explanation subsequently includes the following "Important note on $E_s$":

> *In this particular model, once we already know whether ($H$) is true or false, the observation ($E_s$=true) does not further change the probability of ($E_v$=true). That is because, according to the probabilities supplied, ($E_s$=true) turns out to be independent of ($T_v$) conditional on ($H$). [...]*

At this stage, the model appears to recognise that the computed value coincides with $LR(E_v, H)$, and subsequently generates an incorrect justification for why the observation $E_s$ has no effect on the likelihood ratio. This is an instance of the well-documented phenomenon whereby an LLM produces post hoc rationalisations of erroneous outputs rather than revising them.

In the Jury observation fallacy model, 4 out of 100 runs produce a likelihood ratio of $\approx 1.21$, rather than the correct value of $\approx 0.66$. In this smaller model, the prior observation V=not_guilty was incorporated in all runs, indicating that the error does not arise from omission of the observation. Attempts to reproduce the behaviour by changing the model specification, including the removal of individual dependencies from the BN, were unsuccessful. This suggests that the error arises from a failure to correctly propagate inference through the network, rather than from a simple structural simplification. One possible explanation is that the tight connectivity of the model introduces interactions that are not handled consistently. A precise characterisation of this failure could be obtained through symbolic inference under controlled fault models. This analysis is beyond the scope of this paper.

These findings show that, while a foundation model is capable of producing compelling explanations of a LR, these explanations must be constrained by symbolically derived calculation results. While the errors observed in these experiments might be prevented by more effective prompting strategies, the post hoc rationalisation of errors makes them hard to detect and, therefore, avoid.

### 5.2.2. Foundation + marginals

Under this modality, the prompt is augmented with exact marginal distributions under both the null (prosecution) and alternative (defence) hypotheses, together with the exact likelihood ratio. To prevent the model from recomputing these quantities and thereby introducing avoidable error, the prompt explicitly instructs it not to recalculate the provided marginals or likelihood ratio. As expected, this eliminates the inaccuracies observed in the foundation-only modality. This is confirmed by repeating the accuracy experiment in 5.2.1, which yields 100% correct results across all cases.

An additional advantage of this approach is the reduction in computational demand placed on the foundation model. As a proxy for this computational load, we examine aggregate statistics (mean and

**Table 1**
Token count and output word count over 100 repetitions. Results reported as: average $\pm$ std.

| Model | Token count | | Output word count | |
|---|---|---|---|---|
| | foundation-only | Foundation + marginals | foundation-only | Foundation + marginals |
| Terpenes fire incident | $2400 \pm 82$ | $1773 \pm 106$ | $651 \pm 50$ | $478 \pm 65$ |
| Cross-transfer | $4023 \pm 453$ | $3261 \pm 176$ | $925 \pm 137$ | $646 \pm 75$ |
| Jury observation fallacy | $2973 \pm 403$ | $1597 \pm 113$ | $818 \pm 117$ | $387 \pm 65$ |

standard deviation) of the total number of processed tokens (i.e., token count) across 100 repeated requests under each modality. The results, summarised in Table 1, demonstrate a substantial difference between the modalities across all case studies, with mean differences far exceeding the associated variability.

As a starting point, we expect an explanation to report (i) what the likelihood ratio is, (ii) what the reported likelihood ratio means, and (iii) an explanation b. If left unconstrained, the output of the foundation model under all modalities tend to be verbose (i.e. not concise at all), with typical output word counts running up to many hundreds of words (and too long to report here). Output word counts over 100 test runs are summarised in Table 1. To address this, we have extended the prompt with a request for an executive summary containing "What is the value of the likelihood ratio and what is the value of the evidence in differentiating between the prosecution hypothesis and defence hypothesis? What is/are the main reason/reasons why the value is as high/low/close to 1 as it is?".

A sample explanation produced for the terpenes fire incident model is:

- *Likelihood ratio: 1.0059*
- *Value of the evidence: Detecting terpenes has virtually no value in distinguishing between the prosecution hypothesis and the defence hypothesis. It is only very slightly more probable under the prosecution hypothesis.*
- *Main reasons the value is so close to 1:*
  1. *The prosecution hypothesis increases the chance of a terpene-containing spill only a little: ($P(D = \text{true} \mid H = \text{true}) = 0.019$).*
  2. *Terpenes are already likely to be present for other reasons, especially because the floor contains terpenes with probability (0.75).*
  3. *Therefore, the probability of detecting terpenes is very similar under both hypotheses: (0.759575) versus (0.755113).*

  *In short, the evidence of terpene detection is essentially neutral in this model.*

This summary explanation correctly concludes that the evidence in this model is without value. It also summarises the key reason for this succinctly: i.e. that under the beliefs implied by the model the likely presence of terpenes is high irrespective of the hypotheses and that the activities of an arsonist do not alter these significantly.

A sample explanation produced for the cross-transfer model is:

- *Likelihood ratio: 18.07*
- *Value of the evidence: The finding ($E_v = \text{true}$) supports the prosecution hypothesis over the defence hypothesis. It has moderate probative value in differentiating between them.*
- *Main reasons for this value:*
  1. *Under the prosecution hypothesis, the model makes contact certain and transfer from suspect to victim likely, so ($E_v = \text{true}$) is quite probable ((0.7997)).*
  2. *Under the defence hypothesis, ($E_v = \text{true}$) is much less probable, but not extremely rare ((0.0443)), especially because the already observed ($E_s = \text{true}$) increases the chance of contact and transfer even under the defence case.*
  3. *Background unrelated trace material is very rare in the model, so the result is driven mainly by the transfer mechanism rather than background sources.*

  *So, in short: the evidence is about 18 times more likely if the suspect is the perpetrator than if not, and therefore it favors the prosecution, but not overwhelmingly because the prior observation ($E_s = \text{true}$) also makes ($E_v = \text{true}$) somewhat plausible under the defence hypothesis.*

This summary explanation correctly identifies the key contributing factors to the likelihood ratio. (i) The evidence supports the prosecution. (ii) The prior evidence victim to suspect trace transfer makes contact, and therefore suspect to victim transfer more likely under both hypothesis. This suppresses the likelihood ratio. (iii) Background contamination is unlikely and has, therefore, little impact.

One of the strengths of employing a foundation model for explanation is that it allows continued interaction with the system, based on the previously provided information and responses. In other words, the user can query the explanation and seek further answers. For example, we can ask the explanation system:

> *You: Does the prior observation that traces matching the victim were retrieved from the suspect have any impact on the value of this new evidence?*

A typical response includes:

> *Assistant: [...]*
> *Yes, the prior observation ($E_s$ =true) affects the value of the new evidence.Its main effect is to reduce the additional evidential value of ($E_v$ =true), because the earlier observation already makes contact and transfer more plausible, especially under the defence hypothesis. Therefore the new evidence still favors the prosecution, but with less force than it would if assessed in isolation.*

In our interactions with the explanation system, we have attempted to probe it with a view to commit to fallacies.

> *You: In other words, given $E_s$ =true, the new evidence tells us is 18 times more likely now that the suspect is guilty.*

However, the foundation error tends to be resilient to common errors:

> *Assistant: No. A more correct statement is: Given ($E_s$ =true), the new evidence ($E_v$ =true) is 18.07 times more probable if the suspect is the perpetrator than if the suspect is not the perpetrator. [...]*
>
> - *Incorrect: "The suspect is 18 times more likely to be guilty."*
> - *Correct: "The evidence is 18 times more likely if the suspect is guilty than if the suspect is not guilty."*
>
> *That distinction is very important.*

### 5.2.3. Foundation, marginals, and partial BP

While the Jury observation fallacy model model is small, its outcome under default values (and many variations thereof), is counterintuitive. The LR$\cong 0.66$ indicates that after the "not_guilty" verdict, evidence of prior offences support the defence position that the defendant/suspect is not_guilty. A typical explanation generated by the foundation + marginals models explains why as follows:

> - *Likelihood ratio: 0.6621*
> - *Value of the evidence: The finding Priors=true provides some support for the defence hypothesis rather than the prosecution hypothesis, because it is more probable under $H$ =false than under $H$ =true once $V$ =not_guilty is taken into account.*
> - *Why the value is like this: The main reason is that, in this Bayesian network, the observed verdict not_guilty changes the probability pattern so that Priors=true becomes less likely under the prosecution hypothesis (0.1103) than under the defence hypothesis (0.1665). Because these probabilities are fairly close, the likelihood ratio is below 1 but not far below 1, so the evidence has limited discriminatory value and only mildly favours the defence.*

This description is valid, but it is not particularly insightful as it focusses exclusively on the marginals under the alternative hypothesis. In this particular case, the LR arises from the information flows in the model. On the one hand, the guilt hypothesis promotes the probability of prior offences. On the other hand, once we know the defendant was charged and found not guilty, prior offences explain away the evidence needed for the defendant to be charged. Thus, an innocent suspect with prior offences is more likely to charged than an innocent suspect without prior offences. It is difficult to describe this explaining away effect relying solely on marginals.

If we ask the foundation + marginals system: "*This is counterintuitive. Guilty people tend to be more likely to have committed priors. How is it possible that priors are less likely when we assume the defendant is not guilty?*" it answers something along the lines of:

*Yes, in the model without the verdict, priors are much more likely if the suspect is the perpetrator. But after conditioning on $V$ =not_guilty, the picture changes. The not_guilty verdict gives information about charge and hard evidence. Under $H$ =true, the case is mostly explained by hard evidence, so priors add relatively little. Under $H$ =false, hard evidence is mostly absent, so priors become relatively more important to explain why there was a charge. That is why $E_p$=true ends up more likely under the defence hypothesis than under the prosecution hypothesis in this conditioned setting, giving a likelihood ratio of 0.6621.*

A prompt that includes the partial BP messages often includes a description of the "explaining-away" effect. For example, a typical response under the foundation + marginals + and partial BP model modality is:

*Without conditioning on the verdict: guilt makes priors more likely. After conditioning on $V$ = not_guilty:*

- *the model knows there was a charge*
- *under guilt, the charge is mainly explained by hard evidence*
- *under innocence, hard evidence is largely unavailable*
- *so under innocence, priors become a stronger alternative explanation for the charge*

*That is why priors end up less likely under ($H$ =true) than under ($H$ =false) in this particular conditional comparison.*

This suggests that a specification of partial BP messages prompts the foundation model to consider the influences on the distinct variables, detecting the conflicting effects that give rise to the counterintuitive results.

## 6. Conclusions and future work

This paper explores how the combination of symbolic reasoning techniques with a foundation model can be employed to produce natural language explanations for a likelihood ratio (LR) calculated by means of a Bayesian network (BN). The study shows that a state-of-the-art foundation model is capable of generating compelling natural language explanations for a LR derived from a BN with only high-level guidance. This simplifies the construction of explainable decision support systems in this field considerably. Consistent with prior studies, even a state-of-the-art foundation model cannot be relied to perform the Bayesian reasoning consistently required in this analysis without guidance. While its results can be accurate, this is not reliably the case. To yield reliable explanations, foundation models need to be augmented with symbolic reasoning. This also reduces the computational load on the foundation model substantially.

As a limited examination of four relatively small models, this exploratory study has obvious significant limitations. The key priorities for future work are as follows. Firstly, a critical issue in using foundation models is that their output is variable. This study has not examined the consistency of the characteristics of the explanatory output: repeat experiments were reserved test easily quantifiable elements of the exercise. Assessing consistency would involve running repeated explanation generation runs, and scoring each output on a carefully designed rubric. Consistency also requires fine tuning prompts to ensure the system meets certain conventions and standards consistently. Extending prompts with a detailed step-by-step plan may be a fruitful approach. Secondly, scaling up explanations to substantially larger models (i.e. increasing the variable count and/or connection density) will introduce additional challenges. In particular, increasing model size will cause tension between the completeness, conciseness, and interpretability criteria of generated explanations. Last but not least, foundational model generated explanations are to be relied on by human decision makers, the safety and trustworthiness of the approach needs to be carefully tested.

## Declaration on Generative AI

During the preparation of this work, the author has used ChatGPT to assist in the creation of a prototype Python implementation of the ideas expressed in this paper. All code has been manually tested and

refactored. No generative AI tools were used in authoring the paper.

## References


[1] R. Cook, I. Evett, G. Jackson, P. Jones, J. Lambert, A model for case assessment and interpretation, Science and Justice 38 (1998) 151–156.

[2] C. Lacave, F. Díez, A review of explanation methods for Bayesian networks, Knowledge Engineering Review 17 (2002) 107–127.

[3] Meuwly, D., Ramos, D., Haraksim, R., A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Science International 276 (2017) 142–153.

[4] P. Garbolino, F. Taroni, Evaluation of scientific evidence using Bayesian networks, Forensic Science International 125 (2002) 149–155.

[5] C. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd edition ed., Wiley, 2004.

[6] M. Druzdzel, M. Henrion, Efficient reasoning in qualitative probabilistic networks, in: Proceedings of the 11th Conference on Artificial Intelligence, 1993, pp. 548–553.

[7] M. Druzdzel, Qualitative verbal explanations in bayesian belief networks, Artificial Intelligence and Simulation of Behaviour Quarterly, Special issue on Bayesian belief networks 94 (1996) 43–54.

[8] H. Suermondt, Explanation in Bayesian belief networks, Ph.D. thesis, Stanford University, Department of Computer Science, 1992.

[9] Yap, G-E., Tan, A-H., Pang, H-H., Explaining inferences in bayesian networks, Applied Intelligence 29 (2008) 263–278.

[10] R. Shachter, Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams), in: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998, pp. 480–487.

[11] Haddawy, P., Jacobson, J., Kahn, C., BANTER: A Bayesian network tutoring shell, Artificial Intelligence in Medicine 10 (1997) 177–200.

[12] S. Timmer, J.-J. Meyer, H. Prakken, S. Renooij, B. Verheij, A two-phase method for extracting explanatory arguments from Bayesian networks, International Journal of Approximate Reasoning 80 (2017) 475–494.

[13] J. Keppens, Explaining bayesian belief revision for legal applications, in: Proceedings of the 29th International Conference on Legal Knowledge and Information Systems, 2016, pp. 63–72.

[14] Song, P., Han, P., Goodman, N., Large language model reasoning failures, Transactions on Machine Learning Research (2026). doi:10.48550/arXiv.2602.06176.

[15] Boye, J., Mo ell, B., Large language models and mathematical reasoning failures, 2025. URL: https://arxiv.org/abs/2502.11574. arXiv:2502.11574.

[16] K. Murphy, Y. Weiss, M. Jordan, Loopy belief propagation for approximate inference: an empirical study, in: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, 1999, pp. 467–475.

[17] A. Biedermann, F. Taroni, O. Delemont, C. Semadeni, A. Davison, The evaluation of evidence in the forensic investigation of fire incidents. part ii. practical examples of the use of bayesian networks, Forensic Science International 147 (2005) 59–69.

[18] C. Aitken, F. Taroni, P. Garbolino, A graphical model for the evaluation of cross-transfer evidence in DNA profiles, Theoretical Population Biology 63 (2003) 179–190.

[19] N. Fenton, M. Neil, Risk Assessment and Decision Analysis with Bayesian Networks, CRC Press, 2013.