A Appendix A. On explainability evaluation.

Quantitatively validating latent attention as explanation: As previously noted, evaluating language model explanations is not yet standardized. Despite the effort to make human evaluation fair and reliable, such qualitative measurements are still prone to bias and subjectivity. To validate that latent attention can be used as an explanation, we conduct a stand-alone experiment on the *BeerAdvocate* dataset used by McAuley et al. (2012) and adapted by Lei et al. (2016). This is a dataset that has ground-truth annotations of sentences relevant to prediction results. Although the dataset is not crafted for the purpose of rationale evaluation, we use it as a proxy to examine the quality of our attention scores.

served from the bottle into a becker glass a-poured a semi-hazy orange color with a one finger soapy head s- very sweet floral , slightly citrusy aroma … reminds me more of a pale ale…

Blue background: attended tokens in annotation
Red background: attended tokens not in annotation
Underscore: annotation

Figure 5: Test case example of BeerAdvocate dataset.

The full *BeerAdvocate* dataset contains 1.5 million beer reviews describing four aspects (i.e., *appearance, smell, palate, and taste*), each corresponding to a rating on a scale of 0 to 5. Lei et al. (2016) published a subset of 90k reviews selected to minimize correlation between *appearance* and other aspects. In our experiment, we use these 90k reviews for training, and 994 annotated reviews for testing. The training set only has rating labels, whereas the testing set has both rating labels and human annotations of sentence-level relevancy. Since all aspects have the exact same setups, it suffices to use the *appearance* rating prediction as a proof-of-concept.

We build a model with only two components, described in Section 3.1, namely BERT (pretrained base-case model) and latent attention. We feed static token embeddings from BERT to a latent attention layer, which output sequence representations to be used for regression through a linear layer with a sigmoid activation. We train the model for 20 epochs and select the best performing one for testing.

In contrast to our clinical model, this model only attends to individual tokens and only generates word-level explanations. For words separated by the WordPiece tokenizer, we merge the tokens and average the attention weights. For each sentence, we sort the words based on their attention weights and take the top n words as the prediction rationale, where n equals the total length of the human-annotated sentences. We only use attention mechanisms without additional constraints, such as selection continuity, which makes the testing task even more challenging, as the annotations are ranges of words.

The model is evaluated according to mean squared error (MSE) and rationale precision

$$P_{rationale} = \frac{\sum_{i=1}^{N} |S_i \cup A_i|}{\sum_{i=1}^{N} |S_i|},$$

where N is the number of test cases, y is the ground truth rating of appearance, $\hat{y}$ is the predicted rating, A_i is the set of word indices in the annotated covers, S is the set of word indices selected as model explanations, and $|S| = |A|$.

Our model reaches a rationale precision of 76.39%, which indicates that our most attended words are mostly consistent with the annotations. Figure 5 shows an example of *appearance* test results. The experiment demonstrates the usability of latent attention as an explanation mechanism.

Definition of explanation utility in the rating task: For mortality, each sentence is evaluated individually based on how the described situation would contribute to a patient's survival rate. Sentences describing highly life-threatening complications (such as multiple organ failures) support a positive prediction, whereas sentences indicating improving conditions (such as stable lab measurements) support a negative prediction. In both cases, these sentences are considered helpful. Sentences that are irrelevant (i.e., that support neither a positive nor negative prediction) are considered unhelpful in both populations.

Many of the conditions that present themselves with sepsis onset (such as hypotension) can have numerous etiologies. Diagnostic criteria specify that bacteremia (i.e., bacteria in the bloodstream) must be present in order to predict the development of sepsis. Yet the administration of antibiotics is also not considered as a direct indication of bacteremia without other indications of potential sepsis. Therefore, sentences describing sepsis-related symptoms are not rated as helpful in understanding a positive sepsis prediction until the indication of infection (for example, compromised skin integrity)

Sepsis - Positive Case

… initially admitted on after he was spotted to have partial complex seizure which lasted minutes … vitals show he was intermittently **febrile** and **hypotensive** [1]… patient did not feel lightheaded but did appear sleepy … Pt last had chemo several days prior to admission and is undergoing XRT. Pt currently started on **Zosyn, Flagyl** for broad coverage pending speciation[2]… hypotension relieved with IVF. **bld cultures** check A, urine cultures place central line for access[3]… review with primary team, pt noted to have bloody bowel movement. GI have already been consulted with the concern being ischaemic colitis given the level of **hypotension** in the ED[4]… Review of systems is unchanged from admission except as noted below … doctor aware of **bloodstream infection** in patient with portacath as he placed portacath bld cultures GPC check[5], possible pressors less likely now place line if hypotension persists or pressors are initiated …

Sepsis - Negative Case

… patient lungs clear … Later BP became hypotensive and HR decreased **after receiving neuromuscular reversal agents**[6]. Later BP hypertensive with increased pain … Patient followed commands , afebrile … Temporary pacer set at backup rate and wires sense and pace… ventilation settings changed to CPAP and resulting ABG wnl, patient was **extubated without complications**[7]. Temporary atrial sensitivity threshold…No acute distress, **regular respiratory**, chest expansion symmetric[8] … **Incision clean dry intact**[9]… Encourage wakefulness. Increase activity as tolerated…

Mortality - Positive Case

… admit from OR coronary artery bypass graft … OR course significant for RV failure … Neuro Arrived sedated … Resp Poor oxygenation … Remains on CMV … Glucose gtt started & titrated. Social updated daughters in on pt condition … high dose pressors gtts noted … family decision to withdraw pressors discussed doctors decision made to allow … bp cont decreasing despite titration … vent support weaned significantly… pt became hypotensive with junctional rhythm … CVP levo and vasopressin weaned significantly today …UO boarderlne this am treated with mg iv lasix w/o response. PT started on CVVHD in pm … dialysate running at cc hr endo pt con at insulin gtt BS s.

Mortality - Negative Case

… Poss extub … extubated this morning … wean propofol off. rise to voice, follow commands … diet advanced to clears. swallows no difficulty. mushroom cath intact w watery melena … brother visit this eve … stable for floor transfer … sat room air sat lungs clear. enc deep breathing and coughing. mobilizing clear secretions. Pt still stable for transfer to floor … concern re: hypotension but art line shows stable at systolic … With propofol on board SBP have consistently been … Lung clear on right, coarse and diminished at left base … Obese Abdo with active bowel sounds … Surgery involved. SKIN Intact … LINES and in place and functioning well. Right radial arterial line also present …

---

[1] Description of sepsis-related symptoms which may indicate potential sepsis
[2] Use of antibiotics indicating possible infections
[3] Mentioning blood culture test
[4] Potential complications of hypotension that may happen in septic patients
[5] Description of bacteraemia which means high risk of sepsis
[6] Description non-sepsis-related cause of hypotension
[7] General good sign of improved patient condition
[8] No respiratory distress (infection and inflammatory response can cause irregular respiratory rate)
[9] Presence of incision that is clean and dry, suggesting no infection

Figure 6: Example explanations. Highlighted sentences are rationales picked by our model. Elaboration on the meanings of sentences is written in footnotes. These examples have been edited for increased privacy.

also appears, and vice versa. For negative cases, sentences that are either irrelevant to sepsis or explain other origins of sepsis-related symptoms are rated as helpful. Given this definition, the existence of any helpful sentences means the explanation is valid for a positive case. Similarly, the existence of any unhelpful sentences invalidates a negative case.

Examples of sepsis and mortality explanations are shown in Figure 6. We truncate and edit these texts to avoid data disclosure.