Evaluating Arguments One Step at a Time

We're studying factored cognition: under what conditions can a group of people accomplish complex cognitive tasks if each person only has minimal context?

In a recent experiment, we focused on dividing up the task of evaluating arguments. We created short, structured arguments for claims about movie reviews. We then tried to distinguish valid from invalid arguments by showing each participant only one step of the argument, not the review or the other steps.

In this experiment, we found that:

1. Factored evaluation of arguments can distinguish some valid from invalid arguments by identifying implausible steps in arguments for false claims.
2. However, experiment participants disagreed a lot about whether steps were valid or invalid. This method is therefore brittle in its current form, even for arguments which only have 1–5 steps.
3. More diverse argument and evidence types (besides direct quotes from the text), larger trees, and different participant guidelines should improve results.

In this technical progress update, we describe these findings in depth.

Table of Contents

Methods

Representing arguments as claim trees

In each trial of our experiment, we first sample a random Roger Ebert movie review. An *expert* is instructed to read the entire review and then generate a *root claim* about the review that is either at least 90% likely to be true or at least 90% likely to be false. For example:

> *The film takes a progressive stance on gender relations.*

The expert then builds a claim tree of evidence that contains:

- *Subclaims* that support the root claim.
- *Quotes* from the text supporting each subclaim.

A different expert also reads the text and contributes rebuttals, quotes from the text that are intended to undermine each step of the tree. These rebuttal quotes might show that:

- The root claim is not supported in the text.
- The subclaims are not supported in the text.
- The quotes supporting the subclaim are taken out of context.

Each claim tree has a *depth* (see Figure 1):

- A depth-0 claim tree is just the root claim, nothing else.
- A depth-1 claim tree has the root claim, quotes supporting the root claim, and rebuttal quotes refuting the root claim.
- A depth-2 claim tree has the root claim, subclaims, quotes supporting each of the subclaims, and rebuttal quotes.
- In principle, we can generate arbitrarily large claim trees using nested subclaims and rebutting claims. In this experiment, we use at most depth 2.

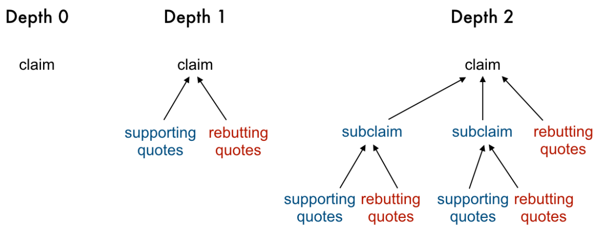


Figure 1: Structure of the claim trees used in this experiment.

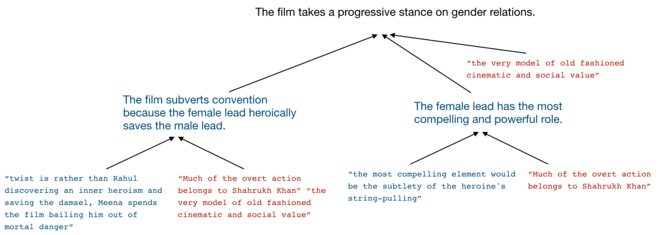

The film takes a progressive stance on gender relations.
"the very model of old fashioned cinematic and social value"
The film subverts convention because the female lead heroically saves the male lead.
The female lead has the most compelling and powerful role.
"twist is rather than Rahul discovering an inner heroism and saving the damsel, Meena spends the film bailing him out of mortal danger"
"Much of the overt action belongs to Shahrukh Khan" "the very model of old fashioned cinematic and social value"
"the most compelling element would be the subtlety of the heroine's string-pulling"
"Much of the overt action belongs to Shahrukh Khan"

Figure 2: Claim tree example. More here.

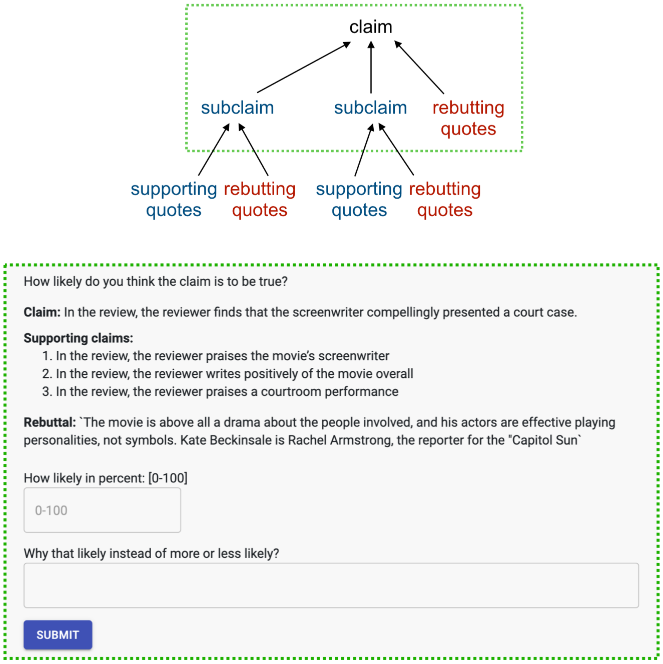
Factored evaluation of arguments

To test whether we can correctly evaluate if a claim about a text is true or false by aggregating the work of context-limited *judges*, we show each step of the claim tree to an experiment participant without showing them the other steps. We then ask how likely the step is to be *valid* by asking how likely the claim (or subclaim) is to be true given the subclaims (or quotes).

This distributed, stepwise evaluation is the main difference compared to Debate as described by Irving et al (2018). In Debate, the judges usually look at the entire debate before they declare a winner.

Evaluating root claims given subclaims

Some participants saw the root claim, the subclaims, and the rebuttals. They had to answer "Assuming all of the subclaims are true, how likely is the root claim to be true?" with a credence from 0% to 100% likely to be true.



How likely do you think the claim is to be true?

Claim: In the review, the reviewer finds that the screenwriter compellingly presented a court case.

Supporting claims:

1. In the review, the reviewer praises the movie's screenwriter
2. In the review, the reviewer writes positively of the movie overall
3. In the review, the reviewer praises a courtroom performance

Rebuttal: `The movie is above all a drama about the people involved, and his actors are effective playing personalities, not symbols. Kate Beckinsale is Rachel Armstrong, the reporter for the "Capitol Sun`

How likely in percent: [0-100]

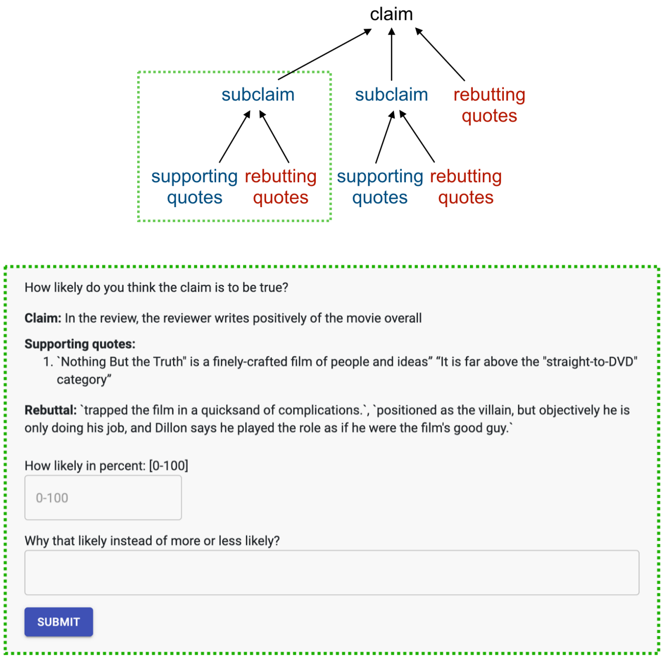
0-100

Why that likely instead of more or less likely?

Figure 3: A judge evaluates the credibility of a root claim given subclaims. Position of argument step (top) and simplified screenshot (bottom).

Evaluating subclaims given quotes

Other participants saw one subclaim, the supporting quotes, and the rebuttals. They had to answer "Given these quotes, how likely is the subclaim to be true?" with a credence from 0% to 100% likely to be true.



How likely do you think the claim is to be true?

Claim: In the review, the reviewer writes positively of the movie overall

Supporting quotes:

1. `Nothing But the Truth" is a finely-crafted film of people and ideas" "It is far above the "straight-to-DVD" category"

Rebuttal: `trapped the film in a quicksand of complications.`, `positioned as the villain, but objectively he is only doing his job, and Dillon says he played the role as if he were the film's good guy.`

How likely in percent: [0-100]

0-100

Why that likely instead of more or less likely?

Figure 4: A judge evaluates the credibility of a subclaim given quotes. Position of argument step (top) and simplified screenshot (bottom).

Claim tree evaluation as binary classification

We want to measure how well factored evaluation (FE) can distinguish true and false claims. The ground truth is provided by the experts who read the whole text. We use the following definitions:

1. Our ground truth is that a claim is *true* if an expert assigns it a >=90% probability of being true, and false if they assign it a >=90% probability of being false
2. FE *classifies a claim as true* if judges evaluate all steps in its claim tree as likely valid, and false otherwise.

Note that (2) specifies a "weakest-link" semantics. If judges think a single step in the claim tree is likely invalid, then FE classifies the claim as false.

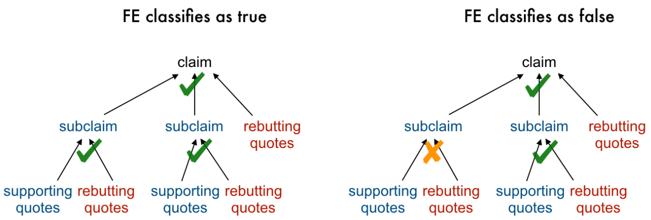


Figure 5: Examples of claim trees classified as true and false by FE. The green check mark indicates that judges evaluated the step as likely valid, and the orange cross as likely invalid.

Having defined how FE of arguments works as a binary classifier, we can use standard confusion matrix metrics to understand its performance. We highlight the two kinds of errors that are possible:

- False positives: A false positive occurs when the expert thinks a root claim is false but FE classifies it as true.
- False negatives: A false negative occurs when the expert judges a root claim to be true but FE classifies it as false.

The confusion matrix allows us to calculate:

- Recall: What fraction of true claims are classified as true?
- Precision: What fraction of claims classified as true are in fact true?
- Accuracy: What fraction of all classifications (both true and false) were correct?

We care most about high precision, and only maximize recall subject to that; reducing false positives matters most to us.

Ensembling step judgments

For each step, participants estimated the likelihood that the (sub)claim was true given the evidence. At least 4 different participants evaluated each step.

To decide whether a claim tree is valid, we need to specify two parameters:

1. Ensembling percentile: take the n-th percentile judgment. For example, the 25th percentile of judgments [10%, 20%, 30%, 40%] is 17.5% with linear interpolation.
2. Threshold: if the ensembled judgment is above this threshold, call the step valid. A claim tree is valid if all steps are valid.

Data

We generated a dataset of 51 claim trees about 51 Roger Ebert movie reviews. Each claim tree had a depth-0, depth-1, and a depth-2 version. For 49 trees we have 4+ judgments for each step in the tree. We restrict the analysis below to these 49 trees.

We collected 2,722 judgments, split as follows:

	Depth 0	**Depth 1**	**Depth 2**
Trees	49	49	49
- Trees with true root claims	24	24	24
- Trees with false root claims	25	25	25
Steps	49	49	207
Mean steps per tree	1	1	4.2
Judgments	826	344	1552
Mean judgments per step	18.86	7.02	7.50

Table 1: Descriptive stats of the claim tree dataset.

Each judgment comes with a natural language explanation by the judge as shown above in Figures 3 and 4.

You can access the full dataset (in json) with all trees, evaluations, and participant explanations here. Trees in human-readable format are here.

Analysis

We present exploratory analysis to synthesize what we learned from this experiment.

These results have methodological limitations that we discuss in the Appendix, but we hope that the analysis provides a framework for informing and evaluating future experiments.

Summary

1. Factored evaluation of arguments can distinguish between some valid and invalid arguments.
    a. The depth-0 baseline was at chance in distinguishing true and false root claims.
    b. Both depth-1 and depth-2 evaluations exceed this baseline across a range of parameters.
    c. For false root claims, factored evaluation of depth-2 claim trees is more likely to identify at least one step in the tree that is unlikely to be valid.
2. However, high variance in judgments across participants leads to brittle performance. Performance is sensitive to the ensembling parameters.
3. By analyzing false positives and false negatives, we've identified specific problems that we can address.
    a. To reduce false positives, we can let the rebuttals include explanations in addition to quotes and instruct claim tree creators to reduce the complexity of individual steps.
    b. To reduce false negatives, we can improve quality control for root claims to ensure that they are indeed clearly true or false and increase tree depth to support arguments in cases where evidence is less direct.

Factored evaluation can distinguish some valid from invalid arguments

The depth-0 baseline is at chance

We confirmed that participants did not have strong prior beliefs about our root claims that could have influenced how they evaluated the tree. At depth 0, the median credence for both false and true claims was 50%. In a binary forced-choice task, participants guessed the truth of the claim correctly for 20 of the 51 trees.

Depth-1 and depth-2 evaluations exceed the depth-0 baseline

Compared to depth 0, depths 1 and 2 result in more accurate evaluations across a range of parameter settings, as shown by the fact that there are parameter settings with above-chance accuracy (light blue pixels below).

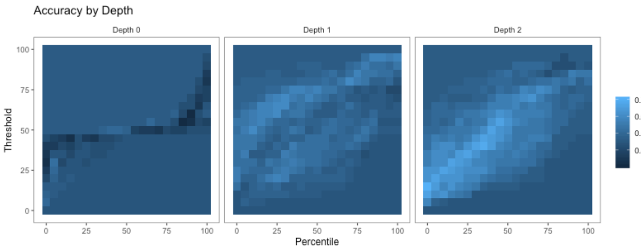


Figure 6: Accuracy as a function of the ensembling parameters. Depth 0 is at chance. For depth 1 and 2, there are parameter settings with moderate accuracy.

Previously, we said that we're most interested in high precision (few false positives). If we optimize the threshold and percentile parameters independently for each depth to maximize recall subject to >80% precision, we get the following confusion matrices:

	Depth 0	**Depth 1**	**Depth 2**
Percentile	*72*	*30*	*1*
Threshold	*59*	*69*	*23*

Confusion matrix	`2 TP 23 FN`
	`0 FP 26 TN`
		`9 TP 16 FN`
		`2 FP 24 TN`
			`10 TP 14 FN`
			`_1 FP 24 TN`

Precision	1.0	0.82	0.91
Recall	0.08	0.36	0.42
Accuracy	0.55	0.65	0.69

Table 2: Classification performance for each tree depth with posthoc fitting of ensembling parameters.

In our analysis below, we'll focus on depth 2. Our results don't show whether depths 1 and 2 differ significantly. Evaluating where depth-1 evaluations differ from evaluations at higher depths is among the most interesting pieces of follow-up work.

Factored evaluation identifies invalid steps in some claim trees for false claims

The figure below shows the ensembled judgments for each step of each depth-2 tree, separating trees with false root claims (left) from trees with true root claims (right).

Each row represents a claim tree. Each dot across that row represents the 25th percentile of all judgments for one step in that tree; each dot represents a different evaluation step. The leftmost dot is therefore the “weakest step” of this tree. We see that the judgments for false root claims skew more left than those for true claims. There is only one tree with an ensembled judgment below 25% percent for true root claims, whereas there are 9 such trees for false root claims.

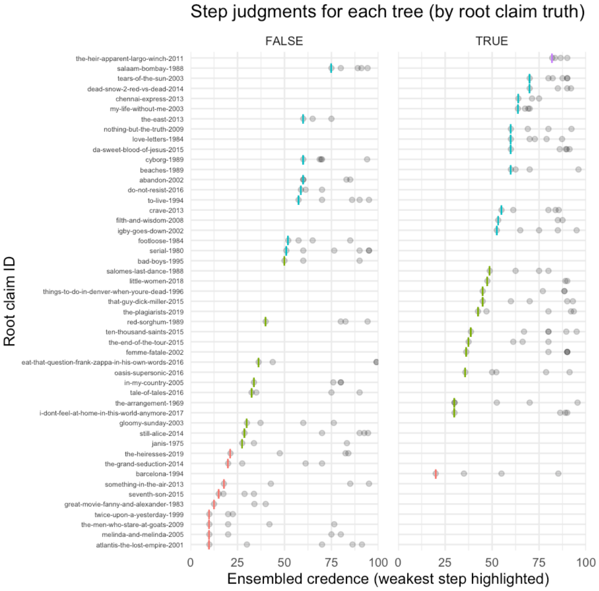


Figure 7: Ensembled step judgments for each tree. There are more steps that judges consider very weak in trees with false root claims, but the difference isn't large.

High variance in judgments leads to brittle performance

Step judgments have high variance

We expected that ensembling would be crucial for smoothing out occasional errors, but learned that evaluations across participants vary widely. Below we show the participant-submitted credences for all 207 steps in the 49 depth-2 claim trees:

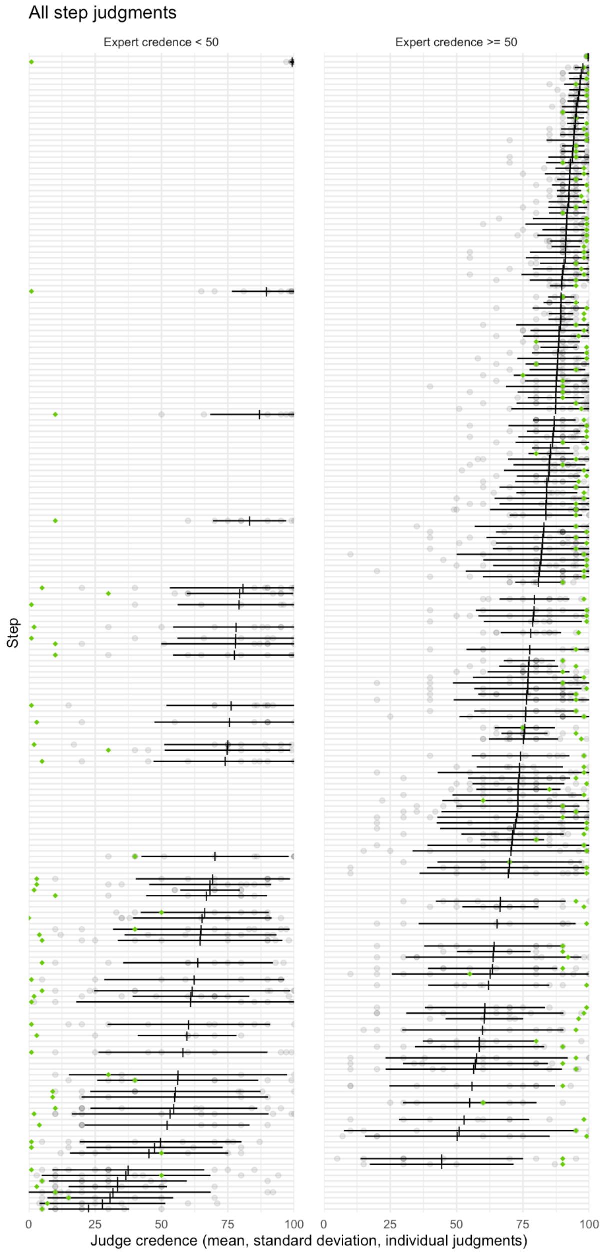


Figure 8: All individual judgments (light gray) for all steps of all depth-2 trees, both quotes to subclaim and subclaim to root claim. The green dots show the credence the expert assigned to the corresponding (sub-)claim.

Each row represents a step from our 49 depth-2 claim trees. The two charts are split out by true and false claims: on the left are the steps evaluating claims (both subclaims and root claims) that experts estimated to be less than 50% likely to be true. On the right are steps evaluating claims that experts estimated to be at least 50% likely to be true. The green dots indicate the credence that the expert assigned the claim (all the green dots in the left chart are below 50%, all the green dots on the right are above 50%).

This plot suggests a few things:

- Judges' credences vary widely for each step; different judges disagree a lot on whether the same claim is likely to be true or false given the evidence.
- Ensembling matters, as we discuss below. We'd have a higher false positive rate if we used the mean of the judgments (vertical black bar) as the overall evaluation for a step.
- As we'd hope, participants give higher credences for claims that are true than for claims that are false.
- Participants generally seem to think that claims are true. The means of most judgments exceed 50%, even for claims that experts generated as false.
- Experts (claim generators) think that their false claims are more obviously false than the judges do (green dots on the left chart skew further left than the

judgments).

- Some true claims seem obviously true to everyone. Some true claims have very high mean judgments with low variance. We do not see the same for false claims.

Performance is sensitive to ensembling parameters

A limitation of our work is that the accuracy metrics and confusion matrix are highly sensitive to the ensembling parameters (judgment percentile, threshold for ensembled credence). For example, here are three settings and the corresponding metrics:

	Parameter setting 1	**Parameter setting 2**	**Parameter setting 3**
Percentile	*1st*	*25th*	*50th*
Threshold	*29%*	*50%*	*75%*
Precision	0.90	0.60	0.60
Recall	0.38	0.50	0.25
Accuracy	0.67	0.59	0.55

Table 3: Different ensembling parameters lead to different metrics.

If we visualize the space of all parameter settings, we see that high values of precision and accuracy (light blue pixels) are sparse:

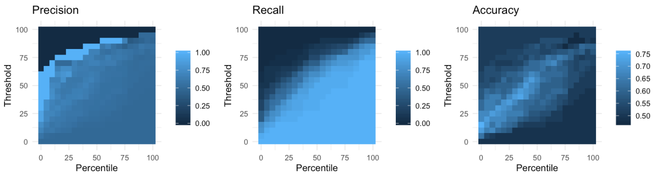


Figure 9: Ensembling parameters that lead to high precision and accuracy are sparse.

Qualitative analysis of false positives and false negatives

For the sensitivity reasons presented above, summary statistics present an imperfect picture of the experiment. In this section, we dive deeper into the trees to qualitatively understand why incorrect evaluations occur and how we can reduce the presence of false positives and negatives in future experiment iterations. For the sake of the qualitative analysis below, we choose to call a tree valid if the 25th percentile judgment exceeds threshold 50.

We find that:

1. False positives are primarily caused by
    a. missing evidence that was difficult to highlight in a rebuttal quote
    b. individual steps with high complexity, leading to judge mistakes
2. False negatives are primarily caused by
    a. mistakes the experts made in choosing root claims
    b. indirect evidence that is difficult to distill into small claim tree

We propose ways to mitigate each of these causes, but haven't implemented the mitigation strategies yet so can't be confident that they would work.

False positives

9 of the 49 claim trees we evaluated were false positives under the threshold set above. Factored evaluation returned that the root claim was true when in fact it was not. The table below summarizes the most common reasons for this incorrect evaluation, followed by discussion of the top two reasons.

Reason for false positive	**# of trees**	**Trees**	**Ways to mitigate**
Judges overlooked the absence of key evidence. Judges failed to notice that an aspect of the (sub)claim is not supported by evidence	8	26, 27, 28, 29, 30, 31, 32, 33	Allow rebuttal via claims in addition to quotes. Better judge instructions and training
Individual steps with high complexity	8	27, 28, 29, 30, 31, 32, 33, 34	More practice and feedback, reduce complexity and difficulty of individual steps (requiring larger trees)
Evidence for claim is ambiguous or uses figurative language, so it’s hard to rebut and judge	3	29, 30, 34	Allow rebuttal via claims in addition to quotes. Train participants on dealing with figurative language

Reason for false positive	# of trees	Trees	Ways to mitigate
Root claim not necessarily false	3	27, 29, 30	Quality control for root claim generation (multiple people review the root claim before building out a tree.)
Rebuttals poorly chosen	2	29, 34	Quality control for rebuttals

Table 4: Reasons for false positives. A tree is listed for a reason if the reason seemed to make a substantial contribution to the tree being a false positive. Trees may be listed for more than one reason.

Claims with ambiguous or figurative quotes are difficult to rebut and judge

Many claims had the following structure: The review text t contains a short quote q that provides strong evidence for the false claim S absent further context:

$$P(S \mid q) > 0.6$$

However, given access to t, the evidence from q is cancelled or explained away:

$$P(S \mid q, t) < 0.1$$

Yet it's not easy to extract a 200-character quote from t that would do the same cancelling. In some cases, there is a slightly longer quote (say 300 characters) that would be sufficient. In other cases, the relevant context from t is distributed throughout the review, which might total 5000 characters. The ambiguity in q sometimes resulted from figurative language or irony, where the literal interpretation was more plausible without context.

For example, consider the claim (simplified from source):

> *"the film features a group of high-school activists"*

The supporting quotes are:

1. "activists are rich kids angry at their parents"
2. "becomes a movie about slumming brats"

The words "kids" and "brats" could either refer to high-school students or to 20-something adults who are immature (which is what the reviewer intends). Various statements in the review suggest the group are not high-school age (e.g. one is a college graduate) but it's hard to condense these statements into a short quote.

Individual steps are complex. Spotting the absence of key evidence is difficult

Factored cognition aims to break down cognitive work into small pieces, making each step of work easier to check and automate. Unfortunately, many steps in this experiment still required judges to do complex things like evaluating subtleties of the relationship between the evidence and the rebuttal, or weighing fine points of the phrasing or implications of a claim.

One source of complexity that confused judges on multiple trees was subclaims that seemed to justify a root claim but that actually contain an unjustified logical leap.

For example, consider the claim (simplified from source):

> *“The reviewer expresses that they are glad that the movie’s political statement was distributed globally despite the movie being banned in China”*

Subclaims:

1. “The reviewer emphasizes the movie’s political statements about the interpretation of Chinese history”
2. “The reviewer emphasizes the movie’s criticism of Chinese Communism”
3. “The reviewer describes the movie’s politically charged reception”
4. “The reviewer expresses excitement that the movie is spreading despite being banned in China”

The subclaims appear to support the claim, but actually they include no evidence that the reviewer was specifically glad to see that the movie’s political statement was distributed globally, as opposed to other aspects of the film. And in fact, the aspect of the movie that the reviewer was excited to see disseminated was its artistic take on history, not its political statements. Judges overlooked the missing evidence, and it was impossible for the rebuttal to directly point out the flaw in the evidence because the rebutter could not express the flaw in direct quotes from the text.

False negatives

12 of the 49 claim trees we evaluated were false negatives. Factored evaluation returned

that the root claim was false when in fact it was true. The table summarizes the most common reasons:

Reason for false negative	# of trees	Trees	Ways to mitigate
Evidence for root claim is ambiguous or indirect, so it’s hard to make a convincing small tree	6	13, 14, 19, 20, 21, 23	Increase tree depth
Truth of root claim is ambiguous	5	15, 20, 21, 22, 23	Quality control for root claim generation through more ensembling or review
Claim tree didn’t provide clear enough evidence for claim (claim tree creator mistake)	3	17, 18, 23	Quality control for tree construction through more ensembling or review
Ambiguous or poorly defined claim (claim tree creator mistake)	3	19, 20, 21	Quality control for tree construction through more ensembling or review
Overly specific intermediate claims (claim tree creator mistake)	3	13, 16, 19	Quality control for tree construction through more ensembling or review

Table 5: Reasons for false negatives. A tree is listed for a reason if the reason seemed to make a substantial contribution to the tree being a false negative. Trees may be listed for more than one reason.

Claim tree creators make mistakes

As described above, we evaluated root claims by checking the probability judges assigned to the weakest step in the claim tree. If any one of the steps evaluating whether quotes support a subclaim, or whether subclaims support the root claim, is invalid, then the root claim evaluates to false. For 8 of the 12 false negatives, mistakes made by the experts constructing claims, subclaims and quotes contributed substantially to the result. These could be mitigated by quality control measures such as allowing experts to get more feedback during the generation process.

Making convincing trees is difficult if evidence for the root claim is indirect

Some false negatives seem to result from fundamental limitations of factored evaluation with small trees. Suppose that a movie is artistically innovative or avant-garde. Instead of stating this explicitly, the review might spend two paragraphs describing the scenes that make it avant-garde. It might be hard to convey the overall effect of those two paragraphs in a few short quotes. Similarly, the reviewer might suggest to readers that a movie is good without stating it. Here's a paragraph from a review (tree data):

> *This is a film about the utter indifference and outright hostility that people encounter every day, and how essentially decent people like Ruth suffer and suffer through it, almost always silently, until they finally snap. The break-in is the culmination of a series of unfortunate encounters: she has to deal with an old racist at the nursing home where she works. She gets stuck in traffic and spies a jerk in a pickup truck at the head of the lane whose tailpipe spews inky smoke as he revs his engine. In a scene that will break the hearts of many regulars who read reviews, Ruth enjoys a drink at a neighborhood bar while reading a new book, only to have a plot twist casually spoiled by another customer that she initially mistakes for a nice guy.*

No single sentence in this paragraph conveys much about the quality of the movie, but the paragraph as a whole is positive. We anticipate that increasing tree size will allow for discussions about such nuances and improve overall performance.

Conclusion

We ran an exploratory experiment in which a distributed group of participants evaluated tree-structured arguments that make claims about movie reviews. We started with shallow arguments that have 1–5 steps and measured success using common classification metrics (precision, recall, accuracy). We found that:

1. Factored evaluation of arguments can distinguish some valid from invalid arguments by identifying implausible steps in arguments for false claims.
2. Experiment participants disagreed a lot about whether claims were true or false. This method is therefore brittle in its current form, even for arguments which only have a few steps.
3. More diverse argument and evidence types (besides direct quotes from the text), larger trees, and different participant guidelines should improve results.

Over time, we'd like to show that accuracy improves as we increase the depth of claim trees, and that we can apply methods like this to much longer texts. A depth-5 tree should reliably discern the truth of a larger set of claims than a depth-2 tree, and we should be able to evaluate claims about entire collections of books, not just single-page reviews. Eventually, we want naive judges to spot-check complex arguments from domain experts even when the judges are entirely unfamiliar with the domain.

We're excited that this experiment established foundations such as operationalizing success for experiments in factored evaluation and creating benchmarks for us and others to improve upon in future work.

Appendix

Acknowledgments

We'd like to thank many different people who contributed to the experiments and their presentation in this blog post.

1. The research was done by William Saunders, Ben Rachbach, Owain Evans, Jungwon Byun, and Andreas Stuhlmüller. Zachary Miller and Andrew Schreiber built the infrastructure.
2. Our experiment participants provided important data and feedback on the experiments. William K, Erol C A, Karin N, Henry A, Julian D, Vojtech B, Henrique D B, Liam D, and Eric H in particular contributed many hours to the experiments.
3. Feedback from Beth Barnes, Vishal Maini, and Milan Griffes helped make the blog post clearer.

This work was supported by many donors, including the Future of Life Institute (RFP2-178).

Citation

Please cite this blog post as:

```
Saunders et al. (2020). Evaluating Arguments One Step at a Time.
```

BibTex citation:

```
@misc{ought2020arguments,
  author = {Saunders, William and Rachbach, Ben and Evans, Owain and Miller, Zachar
  title = {Evaluating Arguments One Step at a Time},
  year = {2020},
  howpublished = {\url{https://ought.org/updates/2020-01-11-arguments}},
  note = {Accessed 11-January-2020}
}
```

Methodological flaws and room for improvement

We're excited about these initial results and about having a more concrete framework for running factored evaluation experiments, but we also recognize that our work is far from perfect. We want to improve upon the following next time and hope readers will cautiously interpret our results in light of these limitations.

Sample root claims independent of the claim tree generation process

We don't believe that our results apply to a broad set of claims because:

- The same expert generated both the claim and the corresponding claim tree. The expert was told to generate claims that are best supported by depth-2 claim trees.
- This generation was done by Ought employees who understood the goals of the experiment and may have been biased in a particular way.
- The inferential gap between the text and our claims was small (by necessity due to small tree size). Our results may not provide much information about claims that require more complex inferences about a text.

The fact that performance on these claims was ambiguous suggests that we didn't stumble upon a narrow set of convenient claims, but we want to control for this more carefully in the future by, e.g., generating claims independent of claim trees.

Check that claim tree generation doesn't have systematic biases

Future experimenters may want to check that the process used to generate claim trees doesn't distort the results. For example, untrained experts could be worse at supporting false root claims than true root claims, or bad at rebuttals for particular types of claims.

Control context for depth-1 and depth-2 judges more carefully

The amount of text that a judge can read to evaluate their step should be the same at all steps and across all depths so that we can isolate the impact of adding more steps at increasing depths. However, some depth-2 judges had more context than depth-1 judges. Judges who evaluated whether or not a root claim was true in light of subclaims saw up to 400 characters of subclaims + 200 characters of rebuttal quotes. All depth-1 judges only saw 200 characters of quotes + 200 characters of rebuttal quotes. Some of the extra characters at the subclaim-to-root-claim level were template characters that provided no new information, which means that the actual difference was smaller.

Even with this additional advantage for depth 2, we don't see much differentiation between depth 2 and depth 1. For experiments that do establish a difference between depth 1 and 2, controlling context size will be important.

Pre-register the experiment

A future iteration of this experiment should have more features of the experiment defined upfront. In this iteration:

- We chose ensembling percentiles and thresholds after seeing the data. We did set a threshold beforehand informed by past work, but the setup differed enough that comparing to our ex-ante thresholds wasn't helpful.
- We didn't control the total number of judgments per step. We limited our analysis to trees with a minimum of 4 judgments for all steps but some of those steps had more than 4 judgments, while others had exactly 4 judgments. We had to balance the distribution of judgments collected per step with considerations like information contamination or providing a reliable stream of work for participants and chose to err on the side of collecting more data when possible.
- Instructions to judges changed slightly throughout data collection as we received feedback from participants. These changes did not seem like they would change results meaningfully to us e.g. they provided more specific instructions for dealing with information contamination.

Clarify the task to reduce variance across judgments

The high variance in judgments we discussed in the analysis section suggests that our task is insufficiently clear to participants. It may also be worth starting with an even simpler task (such as judging arguments about arithmetic).

Minimize information contamination

Given the pool of participants we had access to, many participants evaluated multiple steps from the same tree. In the worst case, this could lead to "information contamination", where a participant's judgment for a step is different from the judgment they would have made if they had no context.

We took steps to mitigate this. We avoided scheduling people to the same tree when possible, we asked participants if they were contaminated and excluded their judgments if so, and each participant only saw the depth-1 or depth-2 tree, not both. A larger pool of participants will minimize the likelihood of contamination further.

Test rebuttals as claims, not just quotes

Instead of being a quote, each rebuttal could be a claim, with supporting quotes and a rebuttal of its own. This would make rebuttals easier to interpret.

Clearly show that depth 2 outperforms depth 1

We want performance to improve with greater depth—everything we do at depth 2 shouldn't be done just as easily at depth 1. This is more of an improvement opportunity than a methodological limitation of this experiment. It's also possible that depths 1 and 2 are too close and that we need to compare a larger depth to depth 1 or 2 to see a difference.

This post was published on January 11, 2020 by Ought.