nature research

Peer Review File

Manuscript Title: A fully autonomous debating system

Editorial Notes: none

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

KEY RESULTS: The key contribution of the work is the construction of the Project Debater system. As made clear in the abstract (and in the paper itself), Project Debater "the first automatic system that can engage in a competitive debate with humans". This represents a huge achievement for the field of Natural Language Processing as well as
for Artificial Intelligence, more generally.

VALIDITY: Does the manuscript have flaws which should prohibit its publication? If so, please provide details.

No, there are no flaws prohibiting publication.

ORIGINALITY AND SIGNIFICANCE: If the conclusions are not original, please provide relevant references. On a more subjective note, do you feel that the results presented are of immediate interest to many people in your own discipline, and/or to people from several disciplines?

The conclusions are original: the IBM team has been developing Project Debater over the course of many years, and publishing results on its various components at top conferences and the relevant workshops in Natural Language Processing along the way.

Given the very broad interest in other AI Grand Challenge-type projects like the AlphaStar Go system and IBM's Watson question answering system, the results of Project Debater should have a similarly broad audience. It is truly an incredible achievement.

DATA & METHODOLOGY: Please comment on the validity of the approach, quality of the data and quality of presentation. Please note that we expect our reviewers to review all data, including any extended data and supplementary information. Is the reporting of data and methodology sufficiently detailed and transparent to enable reproducing the results?

I've reviewed the main paper as well as the supplementary material. The methodology and overall approach are completely sound, and have been documented in multiple publications at the top international conferences in NLP over the past years. The results have been subject to rigorous double-blind review prior to their publication; all of the relevant supporting papers have been referenced in the main publication.

Because of the scope of the project and its many subcomponents, it would be the undertaking of many years to try to duplicate the system-level results reported in the main paper. This is completely expected for a project of this scope and should not be taken as a criticism of the work. In particular, the component-level results (which were not included in the paper and *should not* be included in the paper or supplemental material) would, I believe, be able to be duplicated after reading the relevant referenced paper. But the component-level results are not the focus of this paper.

The paper and supplement are generally very clearly and transparently written. A couple of places where this could be improved are:

(1) It seemed throughout the paper that Project Debater is only able to argue "for" the motion (and so is always the first debater to argue). If this is the case, it is important for readers to be aware of this simplification, even if it is only via the supplementary material.

(2) In the same vein, the supplementary material includes details regarding the simplification of the debating task that I think should be at least referred to in the main text (and the reader pointed to the relevant section of the supplement). In particular, motions presented for debate cannot be too specific (which is partially why, I think, the motion needs to conform to a specific syntactic/semantic template in which the "topic" must be the title of a Wikipedia article (or one of its redirects) and the "action" and/or "view" must match a pre-defined list of terms).

(3) When first describing the high-level architecture (around lines 78-79), it would have been useful to provide a bit of information on the range of computing paradigms that the system components comprise: traditional knowledge-based AI (including the use of manually constructed components), Information Retrieval, of course, machine-learning and NLP techniques.

(4) In the ARGUMENT MINING COMPONENT:

- The description doesn't emphasize what makes this component particularly ground-breaking. The focus seems to be on the fairly standard IR-style sentence-level retrieval from a corpus. But it is the series of neural models that follow and that tailor the component to the argument mining task that are what makes this component impressive I think.

- Lines 98-99, "some 400 million articles^6": It is important to provide some idea of what kind of articles are part of the corpus...news, law, etc.? What was the motivation for selecting this corpus for the debating task? I.e., how are the articles relevant to your argumentation context? Are all articles arguments of some sort?

- Lines 101-106: We get no sense of how arbitrary sentences retrieved using what sounds like a token and entity-based indexing scheme will be guaranteed to have argumentative content. Is it because of the "tailored queries" (line 105) (because the steps that follow only appear to filter out and identify the stance of the retrieved sentences)?

- Line 101, pre-defined lexicon words: need a bit more info on this.

- Line 102 ...when the system is presented with a novel motion, it...:

"Novel" with respect to what, and in what sense? At this point in the system description, there has been no discussion of training vs. test time processing. Isn't it the case that the same

process would be followed regardless of whether the motion had (in some sense) been "seen" during system development? The question of what constitutes a "novel motion" is still an issue after reading the entire paper because the system is so complex and many components are trained on datasets that the paper does not (and cannot) describe.

- Line 106, tailored queries: more info needed on these.

(5) Regarding the KNOWLEDGE BASE:

- Lines 124-125 "We aim to model in advance principled arguments and counter-arguments and common-place examples which may be relevant for a wide range of topics."
Here, it would be great to include an example and to describe very briefly what is meant by "principled".

- For the knowledge base, how many "key sentiment terms" are there? Where did they come from? Same questions regarding the "pattern-based authored responses".

(6) Regarding the REBUTTAL component:

- Lines 143-144 "...and (3) arguments extracted from iDebate in the relatively rare cases..."

How many arguments are extracted from iDebate? How was 'rarity' measured/determined? Why were iDebate arguments employed in the rebuttal system but not used in the argument mining component discussed above?

(7) For DEBATE CONSTRUCTION:

- Line 156, "Finally, the Debate construction is a rule-based Natural Language Generation system."
What is the size? scope?

- Line 159-160 "... the system identifies a main theme, a Wikipedia concept
that is statistically enriched in the cluster's arguments ..."

Please provide an example of a "theme" and also a more concrete description of what is meant by "stasticially enriched"? As currently stated, this step is too vague.

(8) For the EVALUATION section:

Given the difficulty of fairly and consistently determining the quality of a debater, the evaluation of the system over time is well done. Some questions still arise though.

- First, why was the opening system-generated speech (referred to as (i) in line 189) always in favor of the motion instead of half in favor and half in opposition? Is there evidence that arguing in favor is easier/harder than arguing in opposition? Can Project Debater argue the opposition side?

- In addtion, how were the motions used in the evaluations of Figure 2 *and* the additional evaluation of the Feb 2019 system selected? This is important to know, especially for the 36 motions in the second evaluation, as is some determination of the "novelty" of the motions.

- Lines 194-197: It is not clear from the text how/whether the second speech (referred to as (iii)

in line 195) corresponds to the "second speech" sections of Figure 2. The in-line text implies that the second speech includes the rebuttal; does it also include the rebuttal in Figure 2? Or do the second speech sections of Figure 2 include only the non-rebuttal portion?

- Line 199 "...performance along various dimensions on..."
How many dimensions? Mention at least some of them.

- Shouldn't lines 201-202 be rewritten as follows?n
Original 201-202 "... Considering the top 25% of the motions in each evaluation, the average overall score in our last estimation was 3.6 ..."

Rewrite-->? The average overall score of the top 25% of the motions in the final evaluation (i.e., in 2019) was 3.6.

If this is not a correct rewrite, then probably the description of the evaluation needs to be clarified. In addition, if I understood the evaluation correctly, 3.6 should correspond roughly to the average of the three purple bars (under Top 25%) corresponding to the opening, second speech, and rebuttal, respectively. But it doesn't seem like it could be --- the rebuttal score is too low to allow an average of 3.6. Maybe I'm missing something.

- Line 200 ...the results were averaged
Please be more precise. The scores on all dimensions are averaged across all annotators into a single score?

- Line 201 "Considering the top 25% of the motions in each evaluation..."

The paper summarizes the best performance of the system; it should probably also similarly summarize the overall characterization of the remaining 75% of the motions. In particular, we never learn (even in the supplementary material) what each of the scores (1-5) corresponds to in words.

APPROPRIATE USE OF STATISTICS AND TREATMENT OF UNCERTAINTIES: All error bars should be defined in the corresponding figure legends; please comment if that's not the case. Please include in your report a specific comment on the appropriateness of any statistical tests, and the accuracy of the description of any error bars and probability values.

Figure 3 contains error bars but does not indicate what they represent. Standard deviations? No statistical significance tests were reported.

CONCLUSIONS: Do you find that the conclusions and data interpretation are robust, valid and reliable?

In general, I found that the conclusions and data interpretation were sound. In a large system that combines machine learning-based and manually constructed portions (including a knowledge base), there is always a concern that the manually constructed portions are brittle and/or lack scalability and generalization. To combat this concern, it is important that we get a bit more detail on each such component as they are introduced, including (roughly) its size and how critical it is for system performance. The components that fall into this category are the following:

Argument mining: pre-defined lexicon words used for sentence-level indexing; tailored queries;

topic expansion module

Knowledge base: key sentiment terms; pattern-based authored responses

Debate construction: rule-based NLG system; connecting phrase library; pre-defined generation template (is there just one?).

SUGGESTED IMPROVEMENTS: Please list additional experiments or data that could help strengthening the work in a revision.

The transcripts are fascinating! From them, it seems that it is possible to automatically identify the source/component of each sentences/phrase of any generated debate. It would be interesting to see counts of each of the 25 types of content across all of the debates. This would help to determine which parts of the system are employed more than others and could also be used to determine the distribution of content types in debate subsets of varying quality.

REFERENCES: Does this manuscript reference previous literature appropriately? If not, what references should be included or excluded?

References to previous literature are appropriately cited.

CLARITY AND CONTEXT: Is the abstract clear, accessible? Are abstract, introduction and conclusions appropriate?

The abstract, introduction and conclusions are clear and accessible and appropriate with one possible exception. Lines 244-245 state: "We describe a fully autonomous debate system THAT WAS SHOWN TO MEANINGFULLY PARTICIPATE IN A LIVE DEBATE with expert human debaters." I think that this main claim in the conclusion/discussion needs to be qualified. (The wording of the matching claim in the abstract is fine, however.) In particular, the evaluation showed that Project Debater can meaningfully participate in a live debate with expert human debaters, on average, about 25% of the time.

==========

SMALL THINGS

The references in 46-47 (see below) concern only summarization while the sentence seems to apply to all of the tasks mentioned in the preceding sentence.

45 ...there is growing interest in developing
46 widely acceptable benchmark data [14, 15] along
with gaining a deeper understanding of the
47 limitations of current evaluation metrics [16].

Footnote 3: I think it is not quite right to refer to a "motion" as a Boolean statement. The latter is a statement that is either true or false, but a motion is neither true nor false. In addition, footnote 3 might be a good place to make it clear to readers that motions submitted to Project Debater must conform to a particular syntactico-semantic form.

72 "in the supplementary material": Say specifically WHERE in the supplementary material so that readers don't start reading the supplementary material from its beginning (as I did) looking for the transcript. (At this point in the reading of main paper, most of what's in the supplement won't make sense. Minimally, we first need to learn about the system architecture.) At line 72 of

the main paper, I think pointing the reader directly to 5.2 of the supplement (i.e. where the transcripts are) would work.

Figure 2: the white text in the boxes is very difficult to see.

79 tasks received --> tasks had received

79 attention --> prior attention

94 "matches leads coming from the first two components": what does this mean? what is a "lead" in this context? This hasn't been defined yet.

98 At an offline stage --> In an offline stage

106 in them indeed being --> in their indeed being

121 humans conduct--> [remove]

138 arguments who focus --> arguments that focus

210 suggesting that if any exists -->
suggesting that if any overfitting exists

p. 2 of the supplement:
"...We ensured that our motion sets do not include duplicate topics"
-->
"...We ensured that the motion sets employed in our full-system evaluations do not include duplicate topics"

Referee #2 (Remarks to the Author):

This report describes an ongoing project to create a system that can engage in debate competitions with humans. I admire the ambition of this project, and agree with the authors that it presents a variety of interesting and important challenges for AI and human-AI communication. However, I don't feel that this submission has the goals or inferential force of a scientific paper, especially one appropriate for Nature.

The purpose of a scientific paper is to contribute sound and generalizable knowledge to the scientific community. Thus, among other aspects, it is necessary that the key scientific claims are clearly stated, that the methods are described in enough detail to allow replication, and that the evaluations are sufficient to support the claims. The current submission fails to rise to these standards. (Although I should emphasize that it succeeds in other ways: the writing is clear, the philosophical stance is intriguing, and the engineering effort is rather epic.)

Key contributions: It isn't clear what the key ideas/hypotheses are meant to be. The system as described is a complex combination of modules for subtasks. The subtasks are interesting, but they and the modules used are previously reported elsewhere (and are not described in nearly enough detail to evaluate). Perhaps that contribution then is the combination of these pieces? But the exact method of combination is not well described and is not compared to any kind of alternatives. Instead it seems as though the main contribution is that this system participated in

a debate with a human. That is a very cool event, but it is not a scientific contribution in itself.

Methods: Throughout the paper, the methods are barely described and never in the detail that would permit (even conceptual) replication. Explanation of the system are given with phrases like “neural methods”, “machine-learning techniques”, and “rule-based .. generation”. This is like a paper on agriculture saying that “farming techniques” were used — fine for popular press but entirely insufficient for technical insight or replicability.

Evaluation: Perhaps the greatest flaw of this report is the lack of serious evaluation and comparison. The extent of evaluation is the ratings of 5 (non-blinded) annotators on a single likert scale (“can decently take part in a live debate”). There is no attempt to argue or establish the validity of this measure. Why are the judgements of these raters to this question a good metric? The instructions/training of the raters aren’t described, and the raters are not blind to the system or goals of the project. (And more prosaically inter-rater reliability is not reported.) Critically there are no baselines or comparison to alternatives. The only comparison reported is to earlier versions of the system, but this is meaningless because we aren’t told what was changed between versions. Minimally I would have expected comparison to human performance and comparison to meaningfully ablated versions of the system (and comparison to other systems if there were any). The fact of the system participating in a live debate is appropriately not presented as evaluation, yet it is indicated as evidence in the conclusion. I’m ok reading past this, but do want to emphasize that this event alone does not establish any scientific contribution.

Referee #3 (Remarks to the Author):

A. Summary of the key results

This is a very large AI project, combining many different components into a slick final product. The utility of a grand challenge for AI is undeniable, and Project Debater is ambitious in scale, scope and profile. Its argument mining components are amongst the best extant algorithms, and as a cross-domain, if not cross-genre project, the NLP techniques are extraordinarily robust. The paper summarises the architecture of the system, describes how debate contributions are assembled and lays out preliminary evaluation.

B. Originality and significance

The work described here is a novel assembly of many parts, some of which are the state of the art in themselves. To be able to engage in debate is indeed a major milestone in AI evolution — but it is not clear either that the system here does in fact produce all of its own contributions, and more concerningly, the methodology of the evaluation does not convince that the evaluation would distinguish success from failure.

C. Data & methodology

The conception of argument structure and organisation that underlies the system is weak, and the results can be seen in the stilted and halting presentation of material that is created. Evidence cannot be argued for; relevance cannot be supported; counterarguments cannot undercut. These problems could have been headed off at the pass if the foundational work had connected to the literature (or even textbook accounts) from argumentation theory [2]. The problems become manifest in the footnote at 104, in which the definitions are worryingly

incoherent. A Claim is defined as having a clear stance towards the motion. Evidence is defined as something that clearly supports or contests the motion yet is not a Claim. A Claim must furthermore be 'concise', whilst Evidence must be 'a single sentence'. Assuming that single sentences are more or less concise, and that having a clear stance is more or less the same as clearly supporting or contesting, these two definitions are *identical*. All that remains is the final sentence: '[Evidence] provides an indication for whether a … Claim is true' — suggesting that the only way to tell Evidence from Claim is to see whether or not it supports something else.

(On a minor note, there is also a comment that gives pause for thought - what is sentence-level argument mining? [103] Arguments can surely be as long as a book or as short as a word?)

More significantly, the SM reveals at least some of the genesis of the output. Presumably, the black text is templated. It seems that material in classes 4 thru 8 and 21 thru 23 is automatically processed; in classes 9 thru 20 it is extracted from a database of manually created arguments. The problem is twofold: first, the proportion of manually created argument that is just matched and dropped in is very high (almost exactly 50% by word count on the preschool first speech; a little over 50% on the preschool second speech). But much more significant is that almost all of the organisational material lies amongst this manually prepared material that is being selected on the basis of simple topic matching. That is to say, the parts of the speeches that are 'clever' from a linguistic or rhetorical point of view are all canned text.

There is not enough detail in either paper or SM to be sure, but it seems as though the structure of the speech is predetermined, with (admittedly variable number or amount) of mined data being inserted into fairly narrow and tightly constrained slots. There is a lot of work in argumentation theory, rhetoric and cognitive science on the organisation of debates (see, e.g, [1], [2] and rhetoric handbooks from Cicero onwards), which seems to have been missed entirely.

D. Appropriate use of statistics

The most significant concern surrounds evaluation. One of the advantages of previous IBM challenges, Deep Blue and Jeopardy, was that they had crystal-clear success conditions. Whilst, as the authors argue, this is partly as a result of lying inside AI's 'comfort zone' (which may be at least in part easier to say with the benefit of twenty or thirty years of downstream research). Here, there are two aspects to evaluation: polling and annotation. The polling reported in SM§6 is extremely superficial. There is very little statistical power in the results, and it is not even clear that it is measuring the right thing. The evaluation should be telling the reader the extent to which there is an effect. If there was a machine, for example, that performed a google search on, say, "telemedicine facts" and reads out the web page of the top hit, might that have at least as much success on the "knowledge enrichment" poll, and quite possibly the "stance change"? We need to see some kind of baseline or comparator — and preferably a series of benchmarks — in order to formulate and interpret evaluation. The year-on-year improvement described in the paper (186-203) is similarly thin: what is the questionnaire? What is the annotator agreement? What are the grounds for thinking that the subjective assessment of five people is reliable? Are the objective measures of quality that can be used? Why a five point scale? What is meant by 'decently taking part in a live debate'? Do the annotators remain the same? How are they insulated from bias? Finally, in presenting the results, why do we see the top quartile performance? This seems to be saying nothing other than that best performance is better than the average performance. Ultimately, neither the main paper nor SM deliver evaluations with either sufficient detail or sufficient rigour to count as such.

(One example of this superficiality is a comment in SM§4: 'The Unseen Set [was] … by and large

hidden from developers'. This is a worrying phrase: was it hidden or not?)

E. Conclusions

It is certainly the case that something like debating lies beyond typical, older AI systems, but a call to arms to tackle generalized AI or composite AI is not new. (See for example hybrid AI systems from Minsky onwards; the HAIS workshop series; the current debate in the popular press about AGI; and proponents of avoiding narrow focus from McCarthy onwards). And whilst it is certainly reasonable to argue that debates often 'have no clear winner' this seems to be equated with not needing quantitative evaluation which detracts from the significance of the paper significantly.

F. Suggested improvements

For the paper, it would be important to have much more detailed SM, allowing the reader a more detailed understanding of how the system is put together, and where its limitations lie. It would also be important to develop a much more rigorous evaluation rubric. For the work itself, it would be good to see better connection to the literature on rhetoric and on persuasion (at least by linguistic means) to try to address the large proportion of the delivered arguments that are currently simply canned text.

G. References

Referencing in general is very good, though there is something of a lack of referencing outside the authoring team. The system architecture necessarily connects the pieces of the authors' work reported elsewhere, but makes little reference to alternative ways of accomplishing the same tasks. (The reference list as a whole has good balance but generally only in setting up context, not in the detail of, for example, argument mining or rebuttal generation. I know that space is limited, but there needs to be explication of how the work presented here goes beyond the state of the art — and where it doesn't).

H. Clarity & Context

The paper is clear and well written, with an excellent abstract that is crystal clear and exciting. The conclusions are much more fuzzy and hand-wavy and don't leave up to the promise of the abstract.

References

[1] Cialdini, R. Influence: The Psychology of Persuasion, Harper Business, 2006.
[2] van Eemeren, F.H. et al. (eds) Handbook of Argumentation Theory, 2014.

Author Rebuttals to Initial Comments (please note that the authors have quoted the reviewers in italic font and responded in plain font):

1. Response to Referee-1

The key contribution of the work is the construction of the Project Debater system. As made clear in the abstract (and in the paper itself), Project Debater "the first automatic system that can

engage in a competitive debate with humans". This represents a huge achievement for the field of Natural Language Processing as well as for Artificial Intelligence, more generally.

(There) are no flaws prohibiting publication. The conclusions are original: the IBM team has been developing Project Debater over the course of many years, and publishing results on its various components at top conferences and the relevant workshops in Natural Language Processing along the way. Given the very broad interest in other AI Grand Challenge-type projects like the AlphaStar Go system and IBM's Watson question answering system, the results of Project Debater should have a similarly broad audience. It is truly an incredible achievement.

I've reviewed the main paper as well as the supplementary material. The methodology and overall approach are completely sound, and have been documented in multiple publications at the top international conferences in NLP over the past years. The results have been subject to rigorous double-blind review prior to their publication; all of the relevant supporting papers have been referenced in the main publication.

*Because of the scope of the project and its many subcomponents, it would be the undertaking of many years to try to duplicate the system-level results reported in the main paper. This is completely expected for a project of this scope and should not be taken as a criticism of the work. In particular, the component-level results (which were not included in the paper and *should not**

be included in the paper or supplemental material) would, I believe, be able to be duplicated after reading the relevant referenced paper. But the component-level results are not the focus of this paper.

The paper and supplement are generally very clearly and transparently written. A couple of places where this could be improved are:

1.1 It seemed throughout the paper that Project Debater is only able to argue "for" the motion (and so is always the first debater to argue). If this is the case, it is important for readers to be aware of this simplification, even if it is only via the supplementary material.

Response: In order to simplify the development of the system, we focused on the scenario where Project Debater argues for the motion. Note, that the motion can be phrased with the opposite stance, e.g., "We should *not* subsidize preschool", resulting with Project Debater expected to oppose the notion of subsidizing preschool. We clarified this point in Section 1 in the Supplementary.

1.2 In the same vein, the supplementary material includes details regarding the simplification of the debating task that I think should be at least referred to in the main text (and the reader pointed to the relevant section of the supplement). In particular, motions presented for debate cannot be too specific (which is partially why, I think, the motion needs to conform to a specific syntactic/semantic template in which the "topic" must be the title of a Wikipedia article (or one of its redirects) and the "action" and/or "view" must match a pre-defined list of terms).

Response: Per the referee suggestion in comment 1.14 below, we corrected Footnote-3 to indicate that we focused on constraint motion phrasings as described in the Supplementary.

1.3 When first describing the high-level architecture (around lines 78-79), it would have been useful to provide a bit of information on the range of computing paradigms that the system components comprise: traditional knowledge-based AI (including the use of manually constructed components), Information Retrieval, of course, machine-learning and NLP techniques.

Response: We added a corresponding sentence in Line 76.

1.4 In the ARGUMENT MINING COMPONENT

1.4.1 The description doesn't emphasize what makes this component particularly ground-breaking. The focus seems to be on the fairly standard IR-style sentence-level retrieval from a corpus. But it is the series of neural models that follow and that tailor the component to the argument mining task that are what makes this component impressive, I think.

Response: We agree. As indicated in our response to comment 1.4.3 below, we have added more details in Lines 107-110 to convey the essential value of the neural models in the Argument Mining component.

1.4.2 - Lines 98-99, "some 400 million articles^6": It is important to provide some idea of what kind of articles are part of the corpus...news, law, etc.? What was the motivation for selecting this corpus for the debating task? I.e., how are the articles relevant to your argumentation context? Are all articles arguments of some sort?

Response: We used the LexisNexis corpus of English newspaper articles, covering a wide range of media outlets. The focus on newspapers is now explicitly stated in Line 101. Our motivation was to demonstrate argument-mining capabilities over a massive general corpus, with billions of sentences, as opposed to focusing on smaller corpora that are known to be argumentative in nature.

1.4.3 - Lines 101-106: We get no sense of how arbitrary sentences retrieved using what sounds like a token and entity-based indexing scheme will be guaranteed to have argumentative content. Is it because of the "tailored queries" (line 105) (because the steps that follow only appear to filter out and identify the stance of the retrieved sentences)?

Response: The sentence level retrieval, even when relying on the designed queries, is far from enough. The percentage of relevant arguments within sentences that satisfy the designed queries is around 30%, as indicated in Footnote-4 in Ein-Dor et al., AAAI-20. The neural models boost this precision, when averaging across different motions, to around 95% in the top 40 candidates. These details are now explicitly stated in Lines 107-110.

1.4.4 - Line 101, pre-defined lexicon words: need a bit more info on this.

Response: The pre-defined lexicons we used are now described in full in the Supplementary, in Section 5.1.

1.4.5.1 - Line 102 ...when the system is presented with a novel motion, it...: "Novel" with respect to what, and in what sense? At this point in the system description, there has been no discussion of training vs. test time processing. Isn't it the case that the same process would be followed regardless of whether the motion had (in some sense) been "seen" during system development?

Response: We agree, and correspondingly removed the word ‘novel’ in this sentence.

1.4.5.2 - The question of what constitutes a "novel motion" is still an issue after reading the entire paper because the system is so complex and many components are trained on datasets that the paper does not (and cannot) describe.

Response: Indeed, different components of the system were often trained over different sets of motions. Nonetheless, to assess the system in its entirety, we introduced two sets of motions, referred to as Pipeline-set-1 and Pipeline-set-2, which were constructed to estimate the performance when a new, unknown motion will be presented. These sets are described in detail in Section 4 in the Supplementary. The results reported in the paper, for tracking progress over time, comparison to baseline systems, and final evaluation of the system, are all with respect to motions taken from these two sets. In addition, we have set aside a separate collection of motions, termed the Unseen Set, from which motions were selected for live demonstrations of the system. See Section 4 in the Supplementary for more details, and also the reply to comment 1.8.2, which is related to this issue.

1.4.6 - Line 106, tailored queries: more info needed on these.

Response: The queries we developed and used are now described in full in Sections 5.1.1 and 5.1.2 in the Supplementary.

1.5 Regarding the KNOWLEDGE BASE:

1.5.1 - Lines 124-125 "We aim to model in advance principled arguments and counter-arguments and common-place examples which may be relevant for a wide range of topics." Here, it would be great to include an example and to describe very briefly what is meant by "principled".

Response: We have added a detailed description of the types of texts that exist in the Argument Knowledge-Base, alongside examples, in Section 5.2 in the Supplementary.

1.5.2 - For the knowledge base, how many "key sentiment terms" are there? Where did they come from? Same questions regarding the "pattern-based authored responses".

Response: For the key-sentiment terms component, we constructed a list of 6 positive sentiment terms (e.g., “valuable”), and a list of 9 negative sentiment terms (e.g., “harmful”), based on the analysis of human debaters’ speeches. This gave rise to simple sentiment-leads of the form [topic is/are sentiment-term] (e.g., “preschool is harmful”). We then searched for sub-sentences in the opponent speech with high semantic similarity to a sentiment-lead with the correct stance. If found, we generated a short response using one out of five simple potential templates. This component is now described in detail in Section 5.3.3 in the Supplementary.

1.6 Regarding the REBUTTAL component: - Lines 143-144 "...and (3) arguments extracted from iDebate in the relatively rare cases..." --

1.6.1 How many arguments are extracted from iDebate?

Response: The rebuttal sub-component that relies on iDebate is now described in detail in Section 5.3.5 in the Supplementary. We limited this component to generate at most one rebuttal argument per debate.

1.6.2 How was 'rarity' measured/determined?

Response: In our collection of motions used for training and/or development of the various system components, we identified a matching iDebate motion for around 10% of the motions.

1.6.3 Why were iDebate arguments employed in the rebuttal system but not used in the argument mining component discussed above?

Response: We preferred not to rely too much on manually curated data, especially as these data naturally have limited coverage.

1.7 For DEBATE CONSTRUCTION:

1.7.1 - Line 156, "Finally, the Debate construction is a rule-based Natural Language Generation system." - What is the size? scope?

Response: The Debate Construction component is now described in more detail in Section 5.4 in the Supplementary.

1.7.2 - Line 159-160 "... the system identifies a main theme, a Wikipedia concept that is statistically enriched in the cluster’s arguments ..." - Please provide an example of a "theme"

and also a more concrete description of what is meant by "statistically enriched"? As currently stated, this step is too vague.

Response: We use the p-value of the hyper-geometric distribution to characterize the enrichment of a Wikipedia concept in a cluster of arguments. Specifically, this is derived from the (1) total number of arguments; (2) number of arguments in the cluster; (3) number of arguments that mention the concept; and (4) overlap – namely, the number of arguments that are in the cluster and mention the concept. This is now described in detail in Section 5.4.4 in the Supplementary. In addition, an example was added to the main paper in Line 162.

1.8 For the EVALUATION section: Given the difficulty of fairly and consistently determining the quality of a debater, the evaluation of the system over time is well done. Some questions still arise though.

1.8.1 - First, why was the opening system-generated speech (referred to as (i) in line 189) always in favor of the motion instead of half in favor and half in opposition? Is there evidence that arguing in favor is easier/harder than arguing in opposition? Can Project Debater argue the opposition side?

Response: As mentioned in the response to Comment 1.1 above, for simplicity we focused on the scenario where Project Debater argues for the motion. In principle. rephrasing the motion to the opposite stance – e.g., "We should *not* subsidize preschool", should result with Project Debater arguing against the notion of subsidizing preschool. In addition, the common wisdom is that arguing *for* the motion is more challenging for various reasons. E.g., the opposition can use the time of the government opening speech to further prepare, and in addition speaks last which usually leaves a stronger impression with the audience.

*1.8.2 - In addition, how were the motions used in the evaluations of Figure 2 *and* the additional evaluation of the Feb 2019 system selected? This is important to know, especially for the 36 motions in the second evaluation, as is some determination of the "novelty" of the motions.*

Response: Section 3 and Section 4 in the Supplementary now describe in more detail the process by which the motion sets were selected in each of the reported evaluations. In addition, as now stated in Lines 194-196, Pipeline-set-1, that eventually consisted of 78 motions (and not 80, as mistakenly stated in our original submission), was constructed to estimate the performance of the system in its entirety, when a new, unknown motion will be presented.

1.8.3 - Lines 194-197: It is not clear from the text how/whether the second speech (referred to as (iii) in line 195) corresponds to the "second speech" sections of Figure 2. The in-line text implies that the second speech includes the rebuttal; does it also include the rebuttal in Figure 2? Or do the second speech sections of Figure 2 include only the non-rebuttal portion?

Response: The 'Second Speech' throughout the paper, denoted as S3, refers to the entire second speech by Project Debater. This includes the rebuttal to the opponent opening speech, denoted S2, but also new content brought to the debate by the system. This is now explained in detail in in Section 7 and in Figure 3 in the Supplementary, where we use the phrasing 'Rebuttal Aspect of Second Speech' as the label.

1.8.4 - Line 199 "...performance along various dimensions on..." - How many dimensions? Mention at least some of them.

Response: By dimensions we mean the quality of the opening speech, the second speech, etc. We made this more explicit in Line 241. In addition, in Section 7.3.1 in the Supplementary we now provide the complete annotation instructions for this evaluation, that elaborate the meaning of each of the dimensions we considered.

1.8.5.1 - Shouldn't lines 201-202 be rewritten as follows? Original 201-202 "... Considering the top 25% of the motions in each evaluation, the average overall score in our last estimation was 3.6 ..."; Rewrite-->? "The average overall score of the top 25% of the motions in the final evaluation (i.e., in 2019) was 3.6". If this is not a correct rewrite, then probably the description of the evaluation needs to be clarified.

Response: The *Evaluation and Results* section was rewritten, according to the comments by the referees. As part of that, and to address comment 3.4.2.9 by the third referee, we removed the report of the top quartile results.

1.8.5.2 - In addition, if I understood the evaluation correctly, 3.6 should correspond roughly to the average of the three purple bars (under Top 25%) corresponding to the opening, second speech, and rebuttal, respectively. But it doesn't seem like it could be --- the rebuttal score is too low to allow an average of 3.6. Maybe I'm missing something.

Response: The *Progress Over Time* evaluation included a short questionnaire, asking the annotators to reflect their impression of the overall system performance, as well as its performance on more specific dimensions, or aspects – e.g., the quality of the opening speech, the quality of the rebuttal, etc. These instructions are now described in full in Section 7.3.1 in the Supplementary. Correspondingly, the reported 'overall score' is not necessarily a simple average of the scores reported for the other dimensions.

1.8.6 - Line 200 ...the results were averaged - Please be more precise. The scores on all dimensions are averaged across all annotators into a single score?

Response: In the *Progress Over Time* evaluation questionnaire, the votes by all annotators were averaged independently for each question. This is now explicitly stated in Line 242.

1.8.7.1 - Line 201 "Considering the top 25% of the motions in each evaluation..." - The paper summarizes the best performance of the system; it should probably also similarly summarize the overall characterization of the remaining 75% of the motions.

Response: As mentioned, we removed the report of the top quartile results. Still, in Section *In-Depth Analysis* (previously entitled *Error Analysis*) we added analysis to characterize the features of low versus high system performance.

1.8.7.2 - In particular, we never learn (even in the supplementary material) what each of the scores (1-5) corresponds to in words.

Response: The instructions for the *Progress Over Time* evaluation, including the meaning of each of the 1-5 scores, are now described in full in Section 7.3.1 in the Supplementary

1.9 - Figure 3 contains error bars but does not indicate what they represent. Standard deviations? No statistical significance tests were reported.

Response: In all figures in the revised version the error bars represent standard error of the mean. This is now indicated in the figure captions, e.g., the caption of Figure 3. In addition, we added statistical significance tests, as described in Section 7 in the Supplementary.

1.10 - In general, I found that the conclusions and data interpretation were sound. In a large system that combines machine learning-based and manually constructed portions (including a knowledge base), there is always a concern that the manually constructed portions are brittle and/or lack scalability and generalization. To combat this concern, it is important that we get a bit more detail on each such component as they are introduced, including (roughly) its size and how critical it is for system performance. The components that fall into this category are the following: Argument mining: pre-defined lexicon words used for sentence-level indexing; tailored queries; topic expansion module. Knowledge base: key sentiment terms; pattern-based authored responses. Debate construction: rule-based NLG system; connecting phrase library; pre-defined generation template (is there just one?).

Response: The need for more detail in the description of the system components was also raised by the other two referees. Correspondingly, we added Section 5 in the Supplementary that includes a detailed description of each of the main components of the system.

1.11 - The transcripts are fascinating! From them, it seems that it is possible to automatically identify the source/component of each sentences/phrase of any generated debate. It would be interesting to see counts of each of the 25 types of content across all of the debates. This would help to determine which parts of the system are employed more than others and could also be used to determine the distribution of content types in debate subsets of varying quality.

Response: We thank the referee for this suggestion. We added a corresponding analysis, that we believe indeed provides additional insights regarding the system's operation. To avoid an overly fine-grained resolution, we performed a more high-level analysis, distinguishing between mined arguments, content coming from the Argument Knowledge Base, rebuttal content, rebuttal leads, and canned text. This new analysis is reported in Lines 303-311 and in Figure 4, as well as in more detail in Section 10 in the Supplementary.

1.12 - The abstract, introduction and conclusions are clear and accessible and appropriate with one possible exception. Lines 244-245 state: "We describe a fully autonomous debate system THAT WAS SHOWN TO MEANINGFULLY PARTICIPATE IN A LIVE DEBATE with expert human debaters." I think that this main claim in the conclusion/discussion needs to be qualified. (The wording of the matching claim in the abstract is fine, however.) In particular, the evaluation showed that Project Debater can meaningfully participate in a live debate with expert human debaters, on average, about 25% of the time.

Response: We agree, and we rewrote the opening paragraph of the *Discussion* to better reflect the additional analyses reported in the revised version.

1.13 – Minor - The references in 46-47 (see below) concern only summarization while the sentence seems to apply to all of the tasks mentioned in the preceding sentence.

"...there is growing interest in developing widely acceptable benchmark data [14, 15] along with gaining a deeper understanding of the limitations of current evaluation metrics [16]."

Response: We added a reference in Line 43, referring to similar efforts by the community developing dialog systems.

1.14 – minor - Footnote 3: I think it is not quite right to refer to a "motion" as a Boolean statement. The latter is a statement that is either true or false, but a motion is neither true nor false. In addition, footnote 3 might be a good place to make it clear to readers that motions submitted to Project Debater must conform to a particular syntactic-semantic form.

Response: We agree, and we corrected the footnote accordingly.

1.15 – minor - 72 "in the supplementary material": Say specifically WHERE in the supplementary material so that readers don't start reading the supplementary material from its beginning (as I did) looking for the transcript. (At this point in the reading of main paper, most of what's in the supplement won't make sense. Minimally, we first need to learn about the system architecture.) At line 72 of the main paper, I think pointing the reader directly to 5.2 of the supplement (i.e. where the transcripts are) would work.

Response: We agree, and we now refer from the main paper to specific sections in the Supplementary, e.g., in Lines 85-86.

1.16 – minor - Figure 2: the white text in the boxes is very difficult to see.

Response: We updated the figure and hope the text is now easier to read.

1.17 – Grammar etc. –

- 79 tasks received --> tasks had received

- 79 attention --> prior attention

- 98 At an offline stage --> In an offline stage

- 106 in them indeed being --> in their indeed being

- 121 humans conduct--> [remove]

- 138 arguments who focus --> arguments that focus

- 210 suggesting that if any exists --> suggesting that if any overfitting exists

- p. 2 of the supplement: "...We ensured that our motion sets do not include duplicate topics" --> "...We ensured that the motion sets employed in our full-system evaluations do not include duplicate topics"

Response: All the above suggestions were integrated in the main paper and in the Supplementary.

1.18 - Grammar etc. -

- 94 "matches leads coming from the first two components": what does this mean? what is a "lead" in this context? This hasn't been defined yet.

Response: We rephrased Line 96 according to this comment.

2. Response to Referee-2

2.1 - This report describes an ongoing project to create a system that can engage in debate competitions with humans. I admire the ambition of this project and agree with the authors that it presents a variety of interesting and important challenges for AI and human-AI communication. However, I don't feel that this submission has the goals or inferential force of a scientific paper, especially one appropriate for Nature. The purpose of a scientific paper is to contribute sound and generalizable knowledge to the scientific community.

Response: We worked to improve the manuscript in line with this important comment, as described next.

2.1.1 - Thus, among other aspects, it is necessary that the key scientific claims are clearly stated...

Response: We agree. We now state our key claim explicitly in Lines 52-53. In short, we argue that Project Debater, as described in this paper, can decently perform in a debate with a human expert debater. See also our response to Comment 2.1.3.

2.1.2 - … that the methods are described in enough detail to allow replication...

Response: We agree that the methods description was lacking, as also stated by the two other referees. Correspondingly, we added Section 5 in the Supplementary, that describes the main components of the system in more detail. We can add further details as needed.

2.1.3 - … and that the evaluations are sufficient to support the claims.

Response: In response to this comment we conducted new and thorough evaluations to substantiate our main claim. The results are described in the rewritten *Evaluation and Results* Section, Lines 216-231 and in Figure 3b. In short, in 64% of the 78 motions in the evaluation set, the assessment by human crowd annotators supported the claim that Project Debater was exemplifying decent performance in the debate.

2.1.4 - The current submission fails to rise to these standards. (Although I should emphasize that it succeeds in other ways: the writing is clear, the philosophical stance is intriguing, and the engineering effort is rather epic.)

Response: As suggested by our responses above, we hope that the revised version satisfies the standards correctly pointed out by the referee.

2.2 - Key contributions: It isn't clear what the key ideas/hypotheses are meant to be. The system as described is a complex combination of modules for subtasks. The subtasks are interesting, but they and the modules used are previously reported elsewhere (and are not described in nearly enough detail to evaluate). Perhaps that contribution then is the combination of these pieces? But the exact method of combination is not well described and is not compared to any kind of alternatives. Instead it seems as though the main contribution is that this system participated in a debate with a human. That is a very cool event, but it is not a scientific contribution in itself.

Response: We agree that this issue should have been stated more explicitly in our original submission. We revised the manuscript accordingly. The key contribution of this paper is the description of the Project Debater system, along with newly added results in the *Evaluation and Results* Section, that qualify and support the claim that Project Debater can decently perform in a debate. In addition, the system itself is now described in more detail via Section 5 in the Supplementary, and comparison to baseline systems was added in Lines 190-215 and in Figure 3a.

2.3 - Methods: Throughout the paper, the methods are barely described and never in the detail that would permit (even conceptual) replication. Explanation of the system are given with phrases like "neural methods", "machine-learning techniques", and "rule-based .. generation". This is like a paper on agriculture saying that "farming techniques" were used — fine for popular press but entirely insufficient for technical insight or replicability.

Response: As mentioned, we agree to this comment, and correspondingly added Section 5 in the Supplementary, that describes the main components of the system in more detail.

2.4 - Evaluation: Perhaps the greatest flaw of this report is the lack of serious evaluation and comparison. The extent of evaluation is the ratings of 5 (non-blinded) annotators on a single likert scale ("can decently take part in a live debate"). There is no attempt to argue or establish the validity of this measure.

Response: Following this comment, and related comments by the third referee, we have performed additional thorough evaluations and rewrote the *Evaluation and Results* Section to include the corresponding new results. This section is now divided into three parts, covering three evaluation schemes. First, *Comparison to Baseline Systems* was added. To the best of our knowledge, Project Debater is the only automatic system capable of holding a full debate. However, generating a debate opening speech is presumably more feasible, hence we report the comparison of the performance of Project Debater to baseline systems - as well as to expert human debaters - in this more limited task. Second, we added *Evaluation of the Final System.* Here, we used a classical Likert scale in a scenario-based task. Our goal was to specifically quantify, over the evaluation set of 78 motions, to what extent blind crowd workers agree to the statement that Project Debater, as described in the paper, can decently participate in a debate. Finally, the *Progress Over Time* part, covers the evaluation included in our original submission, along with clarifying the rationale behind the design of this evaluation, that was mainly intended to track our progress over the years, in preparation for the Grand Challenge debate.

2.4.1 - Why are the judgements of these raters to this question a good metric?

Response: As mentioned, the evaluation was significantly expanded. In the newly added evaluation of the final system, we examined whether crowd workers that examine the first three speeches in a full debate, agree that the system is exemplifying a decent performance in the debate. For each motion, we collected the votes of 20 crowd workers that were blind to the origin of the speeches, and further added two different controls to ensure annotation quality. The formulation of the question was intended to examine our key claim, as described in our response to Comment 2.1.1 above.

2.4.2 - The instructions/training of the raters aren't described…

Response: In Section 7 in the Supplementary, we now describe the full instructions given to the annotators, in each of the three evaluation schemes. No additional training was performed.

2.4.3 - … and the raters are not blind to the system or goals of the project.

Response: The two new evaluation schemes - of *Comparison to Baseline Systems* and *Evaluation of the Final System* – were performed in the crowd, by raters that were blind to the system and goals of the project. Further, they were not told that they are evaluating the performance of an automatic system. The *Progress Over Time* evaluation, as implied by its name was aimed to assess progress and reflect how prepared the system is for a Grand Challenge debate. Hence, we wanted the annotators to be aware of the fact that this is an automatic system built for the purpose of debating in front of an audience, as this would be known to the audience at the public debates. This is now explicitly stated in Lines 236-238.

2.4.4. - (And more prosaically inter-rater reliability is not reported.)

Response: In Section 7 in the Supplementary we now report inter-annotator agreement for all three evaluations reported in the revised version. In addition, in the same section we report the results of multiple analyses, that all support the reliability of the reported annotation results, and their overall consistency.

2.5 - Critically there are no baselines or comparison to alternatives. The only comparison reported is to earlier versions of the system, but this is meaningless because we aren't told what was changed between versions. Minimally I would have expected comparison to human performance and comparison to meaningfully ablated versions of the system (and comparison to other systems if there were any).

Response: As mentioned, we are unaware of any automatic system capable of holding a full debate except for Project Debater, preventing a standard comparison to baseline systems. Thus, to address this comment, we added an evaluation focusing on the more limited task of generating a debate opening speech, which is clearly the first step any debating system should be capable of. For this task, we compared the performance of Project Debater to the performance of expert human debaters, as well as to six baseline systems. These included two semi-automatic systems that essentially exploited a large amount of manually curated arguments; two systems that relied on the GPT-2 transformer language model, fine-tuned to the task at hand; a system relying on a state-of-the-art argument mining engine; and finally a system that relied on a state-of-the-art summarization method. See Section 6 in the Supplementary for more details. Extensive evaluation of all these methods was performed by crowd annotators. The results are reported in Lines 190-215 and in Figure 3a, as well as in Section 7.1 in the Supplementary.

2.6 - The fact of the system participating in a live debate is appropriately not presented as evaluation, yet it is indicated as evidence in the conclusion. I'm ok reading past this, but do want to emphasize that this event alone does not establish any scientific contribution.

Response: We agree. The participation of the system in public debates serves as anecdotal analysis regarding the system capabilities, and we do not intend to claim it represents an evaluation of Project Debater. The focus of this paper is on the results reported in the *Evaluation*

and Results Section. In particular, the newly added results for *Evaluation of the Final System*, serve to support our main claim, mentioned in our response to Comment 2.1.1 above.

3. Response to Referee-3

*3.1 - **Summary of the key results** - This is a very large AI project, combining many different components into a slick final product. The utility of a grand challenge for AI is undeniable, and Project Debater is ambitious in scale, scope and profile. Its argument mining components are amongst the best extant algorithms, and as a cross-domain, if not cross-genre project, the NLP techniques are extraordinarily robust. The paper summarises the architecture of the system, describes how debate contributions are assembled and lays out preliminary evaluation.*

*3.2 - **Originality and significance** - The work described here is a novel assembly of many parts, some of which are the state of the art in themselves. To be able to engage in debate is indeed a major milestone in AI evolution — but...*

3.2.1 … it is not clear either that the system here does in fact produce all of its own contributions...

Response: In response to this comment, and also per the suggestion of the first referee, we added new analysis that reports the fraction of content generated by each part in the system pipeline. As now described in Figure 4, and in more detail in Section 10 in the Supplementary, the component that contributed most of the content is the Argument Mining one, and this trend is even more evident for motions on which the system performs relatively well (Figure 4a). Further, in the added detailed description of the Argument Knowledge Base (Section 5.2 in the Supplementary) we explain that even assigning content from this component is far from trivial, and at least for some classes may be quite nuanced and challenging. That is, even though the text there is written in advance in the context of a pre-specified class of motions, it is still a challenge to understand whether it will be applicable for a newly introduced motion. Finally, as now indicated in Figure 4b, considering the 78 motions in our evaluation set, conventional canned text represents less than 18% of the content generated by the system.

3.2.1 - … and more concerningly, the methodology of the evaluation does not convince that the evaluation would distinguish success from failure.

Response: In response to this issue, also raised by the second referee, we have added two extensive evaluations, and rewrote the *Evaluation and Results* Section accordingly. In the newly added *Comparison to Baseline Systems*, for the task of generating a debate opening speech, our results demonstrate that Project Debater significantly outperforms all baselines, while still lagging behind expert human debaters (Figure 3a), suggesting that this new evaluation can distinguish success from failure, and further identify the shades of gray in between. Similarly, the results reported in the newly added *Evaluation of the Final System*, Lines 216-231, further qualify and support our claim that Project Debater can decently perform in a debate. We further expanded the *Error Analysis* section, now referred to as *In Depth Analysis*, and in Lines 279-308 provide a more detailed comparison between motions with satisfactory performance, versus motions over which the system fails to perform well.

3.3 - ***Data & methodology*** *- The conception of argument structure and organisation that underlies the system is weak, and the results can be seen in the stilted and halting presentation of material that is created. Evidence cannot be argued for; relevance cannot be supported; counterarguments cannot undercut. These problems could have been headed off at the pass if the foundational work had connected to the literature (or even textbook accounts) from argumentation theory [2].*

Response: We agree that from the perspective of argumentation theory our approach was rather simplistic. This was mainly dictated by the pragmatic constraints of building a working system within a reasonable time frame, coupled with the understanding that more involved argumentation models will naturally increase the complexity of the project, placing additional burden in various aspects associated with developing the system. In addition, we were not certain what would be the impact of relying on more subtle models on the expected non-expert audience of our Grand Challenge debate. We hope that future work will highlight the potential role of deeper models in our context. Beyond that, to address this comment in the paper, we added Section 8 and Section 9 in the Supplementary, describing the development process which started from a relatively conventional – albeit minimalistic - view of argumentation, and ended in the simplified definitions for claim and evidence and for how to combine them. In addition, we note that our new evaluation suggests that when compared to what currently can be taken as baselines for generating an opening speech, the results of Project Debater as perceived by crowd annotators significantly outperform the performance of all baselines (Figure 3a), perhaps also in terms of fluency.

*3.3.1 - The problems become manifest in the footnote at 104, in which the definitions are worryingly incoherent. A Claim is defined as having a clear stance towards the motion. Evidence is defined as something that clearly supports or contests the motion yet is not a Claim. A Claim must furthermore be 'concise', whilst Evidence must be 'a single sentence'. Assuming that single sentences are more or less concise, and that having a clear stance is more or less the same as clearly supporting or contesting, these two definitions are *identical*. All that remains is the final sentence: '[Evidence] provides an indication for whether a … Claim is true' — suggesting that the only way to tell Evidence from Claim is to see whether or not it supports something else.*

Response: We added Section 8 in the Supplementary to describe the rationale for the definitions we used, and the focus on single sentences, and we refer to this section in Footnote 7. It is true that the ultimate definitions we used for claims and evidence were similar. However, as the referee points out, they were not identical. E.g., as now described in the Supplementary, while an evidence was always represented by a full sentence – and rarely two, a claim was typically a phrase within a sentence. Correspondingly, the average length of evidence and claims across all Project Debater speeches in the 78 motions in our evaluation set were 26 and 11.5 tokens, respectively. Moreover, the far more elaborate guidelines that we initially used have defined the initial dataset on which our classifiers were originally trained (Aharoni et al., 1st Argument Mining Workshop, 2014), where the distinction between claim and evidence was even more apparent. These differences propagated to the final classifiers used by Project Debater, due to the retrospective labeling paradigm we applied to collect the labeled data to train the classifiers, as described by Ein-Dor et al. (AAAI, 2020). Finally, we note that that candidate evidence sentences

were constrained by the queries described in the same paper (see Section 5.1.2 in the Supplementary), resulting with additional differences from candidate claims.

3.3.2 - (On a minor note, there is also a comment that gives pause for thought - what is sentence-level argument mining? [103] Arguments can surely be as long as a book or as short as a word?)

Response: In Section 5.1 in the Supplementary we now describe in more detail the sentence-level argument mining and the rationale for taking this approach. Indeed, a detailed argument can be much longer than a single sentence. However, focusing on single-sentence arguments made the task of some of the underlying argument mining components more feasible. Perhaps more importantly, we felt that enabling Project Debater to extract arguments consisting of several sentences, full paragraphs, or even full articles, and pasting them into the speech in their entirety would have missed part of the goal we were aiming for. Rather, in our view, building a narrative based on short arguments - or argument parts - is a more challenging and interesting task.

3.3.4 - More significantly, the SM reveals at least some of the genesis of the output. Presumably, the black text is templated. It seems that material in classes 4 thru 8 and 21 thru 23 is automatically processed; in classes 9 thru 20 it is extracted from a database of manually created arguments. The problem is twofold:

3.3.4.1 first, the proportion of manually created argument that is just matched and dropped in is very high (almost exactly 50% by word count on the preschool first speech; a little over 50% on the preschool second speech).

Response: In Figure 4 in the main paper and in Section 10 in the supplementary we now report systematic results regarding the break-down of the generated content by source. We stress that using content from the Argument Knowledge Base – Types 9 thru 20 in the transcripts – is more involved than simple topic matching, as now described in Section 5.2 in the Supplementary. In fact, in the transcribed debate about subsidizing preschool, which was matched to a Knowledge Base class pertaining to Welfare, one can see that this match is not perfect, and that the resultant arguments are slightly awkward in the context of the debate. Other examples for the impact of such an error are discussed in Lines 268-269, and in Lines 297-302.

3.3.4.2 - But much more significant is that almost all of the organisational material lies amongst this manually prepared material that is being selected on the basis of simple topic matching. That is to say, the parts of the speeches that are 'clever' from a linguistic or rhetorical point of view are all canned text.

Response: As discussed above, and also in Section 5.2 in the Supplementary and by Bilu et al. (ACL, 2019), matching a motion to the appropriate classes is a subtle and quite challenging task. Moreover, the cost of an error is high, since it may result with less relevant – or even completely irrelevant - content throughout the entire debate, implying a requirement for very high precision. In addition, we note that the organization of the speech goes beyond the content of Types 9-20 in the transcripts, as these content units should be integrated alongside mined arguments and over three speeches.

3.3.5 - There is not enough detail in either paper or SM to be sure, but it seems as though the structure of the speech is predetermined, with (admittedly variable number or amount) of mined

data being inserted into fairly narrow and tightly constrained slots. There is a lot of work in argumentation theory, rhetoric and cognitive science on the organisation of debates (see, e.g, [1],

[2] and rhetoric handbooks from Cicero onwards), which seems to have been missed entirely.

Response: As the referee suggested, the general structure of each of the three speeches was relatively predetermined, as now described in detail in Section 5.4 in the Supplementary. As mentioned in our response to Comment 3.3 above, this design decision was mainly motivated by an attempt to simplify the problem to enable developing a decently performing system in a time frame of several years. Correspondingly, we focused on speech structures inspired by the know-how and experience of expert debaters, and not on formal theory of rhetoric. In hindsight, it would have been better to explore both directions simultaneously. We hope that future work will further elucidate this direction.

3.4 - Appropriate use of statistics - The most significant concern surrounds evaluation. One of the advantages of previous IBM challenges, Deep Blue and Jeopardy, was that they had crystal-clear success conditions. Whilst, as the authors argue, this is partly as a result of lying inside AI's 'comfort zone' (which may be at least in part easier to say with the benefit of twenty or thirty years of downstream research). Here, there are two aspects to evaluation: polling and annotation.

3.4.1.1 - The polling reported in SM§6 is extremely superficial. There is very little statistical power in the results, and it is not even clear that it is measuring the right thing. The evaluation should be telling the reader the extent to which there is an effect.

Response: The polling reported in the Supplementary, for the three public debates, is indeed anecdotal in nature, and was not intended to be presented as evaluation. In Lines 181-186 we further discuss the limitations of this measure. Our main results, including new results added to the revised version, are described in the updated *Evaluation and Results* Section, and in detail in Section 7 in the Supplementary, including a description of how the results were analyzed and the statistical significance tests we performed.

3.4.1.2 - If there was a machine, for example, that performed a google search on, say, "telemedicine facts" and reads out the web page of the top hit, might that have at least as much success on the "knowledge enrichment" poll, and quite possibly the "stance change"? We need to see some kind of baseline or comparator — and preferably a series of benchmarks — in order to formulate and interpret evaluation.

Response: In response to this comment, and to Comment 2.5 made by the second referee, we performed additional extensive evaluation, that included comparison to baseline systems. Since we are unaware of any other automatic system capable of holding a full debate, we focused this comparison on the more limited task of generating a debate opening speech. As reported in Lines 190-215 and in Figure 3a, the average scores by crowd annotators for Project Debater speeches are significantly higher compared to all baselines examined, including methods that use manually curated high quality arguments as input.

3.4.2 - The year-on-year improvement described in the paper (186-203) is similarly thin:

Response: Below we answer the specific questions raised, but add here as a more general note, that following this comment and related comments by the second referee we performed

additional evaluations, the results of which are included in the revised version. First, we added *Comparison to Baseline Systems*, as mentioned in our response to Comment 3.4.1.2 above. In addition, we added a new *Evaluation of the Final System*, in Lines 216-231 and in Figure 3b, where we use a classical Likert scale in a scenario-based task, and add two controls, aiming to estimate the level to which humans perceive Project Debater as able to decently participate in a debate. The original *Progress Over Time* evaluation is now reported in Lines 232-247 and in Figure 3c, with additional details of the rationale behind its design, which aimed to assess progress and reflect how prepared the system is for a Grand Challenge debate.

3.4.2.1 - what is the questionnaire?

Response: In Sections 7 in the Supplementary, we now describe the full instructions given to the annotators for all evaluation schemes.

3.4.2.2 - What is the annotator agreement?

Response: In Section 7 in the Supplementary we now report inter-annotator agreement for all three evaluations reported in the revised version. As we discuss in this section, due to the subjective nature of scoring a debate or a speech, these agreement scores may be expected to be relatively low. Correspondingly, in the same section we report the results of multiple analyses, that support the reliability of the reported annotation results, and their overall consistency.

3.4.2.3 - What are the grounds for thinking that the subjective assessment of five people is reliable?

Response: The evaluation of advanced NLG systems - such as automatic translation, summarization, or dialog systems - is a notoriously difficult problem, precisely due to the fact pointed by the referee - that the assessment is often subjective in nature. This is especially true in our context and represents another aspect of the scientific challenge of our work. To address this issue, we applied several measures. First, in the two new evaluations included in the revised version, we used 15-20 annotators to evaluate each instance, which is beyond what is typically applied in the literature. In addition, we included various controls and analyses, reported in Section 7 in the Supplementary, to further support the reliability of our annotation results. Finally, in all the evaluations we now report the standard error of the mean and analyze the statistical significance of the results.

3.4.2.4 - Are the objective measures of quality that can be used?

Response: As mentioned, following the above comments as well as the related comments by the second referee, we added more evaluations that are reported in the revised version. While we are unaware of well-defined objective measures that can be used, we added a comparison to the results over speeches recorded by human expert debaters as a reference, depicted in Figure 3a.

3.4.2.5 - Why a five point scale?

Response: A five point scale is a standard choice for a Likert scale (http://rube.asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html), often from 'Strongly Agree' to 'Strongly Disagree'. In the new evaluations, we adopt this exact scale, while in the *Progress Over*

Time evaluation, we designed a specific scale intended to reflect the ability of the system to perform in a live debate, enabling us to track progress towards that goal.

3.4.2.6 - What is meant by 'decently taking part in a live debate'?

Response: This phrase was used as part of the instructions for the *Progress Over Time* analysis. A similar phrase was used in the new *Evaluation of the Final System*, and in both cases, we expected the annotators to interpret that literally. To ensure reasonable interpretation, in the new evaluations we used scenario-based tasks, in which we ask the annotators to imagine they are in the audience of a competitive debate. The full instructions for all three evaluations are now shared in Section 7 in the Supplementary.

3.4.2.7 - Do the annotators remain the same?

Response: In all three evaluations the annotators were selected from a fixed set that was larger than the number of annotations per motion. Thus, there is variability of annotators between motions, with some overlap. This is now described in more detail in Section 7 in the Supplementary.

3.4.2.8 - How are they insulated from bias?

Response: In the *Progress Over Time* evaluation we aimed to obtain a reliable estimate of how prepared the system is for its public debut. Thus, we wanted the annotators to be aware of the fact that they asses the performance of an automatic system built for the purpose of debating in front of an audience, as this would be known to the audience at the actual event. This is now explicitly stated in Lines 236-238. In the two new evaluation schemes, the crowd annotators were blind to the system and goals of the project, and further were not even told that they are evaluating the performance of an automatic system.

3.4.2.9 - Finally, in presenting the results, why do we see the top quartile performance? This seems to be saying nothing other than that best performance is better than the average performance.

Response: We agree, and we have now excluded the top quartile report altogether.

3.4.2.10 - Ultimately, neither the main paper nor SM deliver evaluations with either sufficient detail or sufficient rigour to count as such.

Response: We appreciate the thorough feedback regarding the evaluation included in our original submission. As detailed above, we have made significant changes to both the main paper and the Supplementary to address those concerns. The changes include both extensive evaluations that were not present in the original submission, as well as adding more details regarding the original evaluation. We hope these changes address the concerns raised by the referee in this concluding comment.

3.4.2.11 - (One example of this superficiality is a comment in SM§4: 'The Unseen Set [was] … by and large hidden from developers'. This is a worrying phrase: was it hidden or not?)

Response: As described in Sections 3 and 4 in the Supplementary, some of the developers were involved in the process of motion collection, that also gave rise to the Unseen motion set. Thus,

at least these developers were to some extent exposed to motions in this set. That said, the motions in this set were not included in the routine development and assessment of the system. We rephrased the relevant line in Section 4 in the Supplementary accordingly.

*3.5 – **Conclusions** - It is certainly the case that something like debating lies beyond typical, older AI systems, but a call to arms to tackle generalized AI or composite AI is not new. (See for example hybrid AI systems from Minsky onwards; the HAIS workshop series; the current debate in the popular press about AGI; and proponents of avoiding narrow focus from McCarthy onwards). And whilst it is certainly reasonable to argue that debates often 'have no clear winner' this seems to be equated with not needing quantitative evaluation which detracts from the significance of the paper significantly.*

Response: As detailed above, in response to this concern, also expressed in other comments, in the *Evaluation and Results* section we added two new quantitative and extensive evaluations, and further elaborated on the evaluation included in our original submission.

*3.6 - **Suggested improvements** -*

3.6.1 - For the paper

3.6.1.1 … it would be important to have much more detailed SM, allowing the reader a more detailed understanding of how the system is put together, and where its limitations lie.

Response: We agree, and correspondingly added Section 5 in the Supplementary that provides a detailed account of the main components of the system.

3.6.1.2 - It would also be important to develop a much more rigorous evaluation rubric.

Response: We agree, and as mentioned, added new evaluations, elaborated on the results included in our original submission, and rewrote the *Evaluation and Results* Section accordingly.

3.6.2 - For the work itself, it would be good to see better connection to the literature on rhetoric and on persuasion (at least by linguistic means) to try to address the large proportion of the delivered arguments that are currently simply canned text.

Response: We added Section 8 and Section 9 in the Supplementary to better explain how our methodology is related to argumentation theory. In addition, in Section 5.2 in the Supplementary, we added a more detailed description of the Argument Knowledge-Base component, aiming to convey that incorporating the knowledge therein is more involved than simply inserting canned text. Finally, in the newly added Figure 4 in the main paper and the corresponding Section 10 in the Supplementary, we now report a detailed breakdown of the generated content according to the text's source, showing that a relatively moderate amount of text is actually canned.

*3.7 - **References** - Referencing in general is very good, though there is something of a lack of referencing outside the authoring team. The system architecture necessarily connects the pieces of the authors' work reported elsewhere, but makes little reference to alternative ways of accomplishing the same tasks. (The reference list as a whole has good balance but generally only in setting up context, not in the detail of, for example, argument mining or rebuttal generation. I*

know that space is limited, but there needs to be explication of how the work presented here goes beyond the state of the art — and where it doesn't).

Response: In the new *Comparison to Baseline Systems* evaluation we compare our system to several alternative methods for generating a debate opening speech, that do not rely on our work. These methods are described and referenced in Section 6 in the Supplementary. In addition, we added Section 9 in the Supplementary where we aim to explain the connection between our modeling choices and argumentation theory, with hopefully appropriate reference there as well.

*3.8 - **Clarity & Context** - The paper is clear and well written, with an excellent abstract that is crystal clear and exciting. The conclusions are much more fuzzy and hand-wavy and don't leave up to the promise of the abstract.*

Response: We have tried to improve the presentation in the *Discussion* Section, and in particular rewrote the opening of this section, in Lines 312-323.

References provided by R3 –

[1] Cialdini, R. Influence: The Psychology of Persuasion, Harper Business, 2006.
[2] van Eemeren, F.H. et al. (eds) Handbook of Argumentation Theory, 2014.

4. Additional Updates

4.1 We corrected the description of how stance detection is done in Lines 110-112.

4.2 The title and opening of the *Error Analysis* Section in Lines 258-260 were updated to better reflect the updated content of this section, and better connect it to the updated content in the previous section.

4.3 Pipeline-set-1, described in the Supplementary, eventually included 78 motions and not 80 as stated in our original submission. This is corrected in Line 194, and in Section 4 in the Supplementary.

4.4 We corrected an error in describing the task presented for the *Progress Over Time* evaluation. Originally, we wrote - "For each motion, every annotator was asked to read and listen to three speeches", while in fact, annotators were only presented with audio files. This is now stated explicitly in Lines 235-236.

4.5 We updated the abstract in Lines 19-22, and Lines 65-67, to reflect the new analyses added to the revised version.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

The authors did an excellent job in responding to the concerns raised in the first round of reviewing. I have no remaining or new concerns that require modifications to the manuscript.

Referee #2 (Remarks to the Author):

I have now evaluated the revision and response letter. I appreciate the detailed and constructive responses. I will largely restrict myself to the issues I raised in my initial review.

First, regarding replicability. I appreciate that a great deal more information is now included in the appendices. I'm still not completely certain that I (or another AI expert) could replicate the system from this description. (Though it's a bit hard to tell without going through this and previous papers in a great deal of detail.) It seems pretty likely that this is by design since there is no intention to open source (and key components are to be available as a service-for-fee). This hazy line between openness and competitive advantage is probably the inevitable result of doing science that is also directly part of a for-profit endeavor. I am personally uncomfortable with this, but I will leave it to the editor to decide if it is acceptable for Nature.

Second, regarding evaluation. I appreciate the work that was put into adding evaluations. I think they are vastly better than the previous evaluations, though I have some questions about the new additions. First some statistical questions: The error bars in fig 3 are surprisingly small for 15 responses on a 5 point scale. Partly this may be because they are standard errors, which are not appropriate for this type of data (stderr assumes normal data; use eg bootstrapped CI instead). But even so, they look small to me, and the low kappas indicate higher variance as well. (Any reason not to provide the anonymized data, so people can explore such questions?) Relatedly it is not reported what test the significance values come from. Please describe the statistics in more detail!

The evaluation of fig3a seems to me like a simple but sound test of the overall quality of the initial statement. The baseline models are a good-faith attempt to put together reasonable representatives and the comparison to human experts is important. Overall this does look like it supports the claim that the system provide "decent performance" at generating initial statements. The test of the full model, ie fig 3b, seems a bit less compelling. For one thing, if I understand correctly, the S2 human response is not a response to the particular S1 generated by the model. These dialogues are thus already questionably coherent and the model's ability to coherently respond in S3 might not be well tested. The baselines also strike me as much weaker for this experiment.. I'm not sure why we'd expect them to be coherent at all. Further there is no human control for this experiment making it more difficult to tell how well the model is actually doing. (Being judged "decent" in a little over half the motions doesn't seem all that great.) Just to make clear what is probably obvious, this direction of evaluation is a positive step, but it's much weaker than "human level" performance — coherence, persuasion, fluidity.

A final minor point on evaluation: with these more clear evaluations added, I would think the "Progress Over Time" Eval could be cut — it's an internal development metric that isn't properly controlled, so doesn't provide much information.

Finally, regarding clear statement of the scientific claims. As indicated in the response letter, the claim is "results indicating this system can decently perform in a debate with a human expert debater." This is clear, which I appreciate, though to me it isn't the most interesting kind of scientific claim. It is an existence claim for an artifact with a certain property. It isn't a claim that, for instance, certain principles or ideas lead to such an artifact. The value then comes down to whether the existence of the artifact, or the techniques described, are of profound scientific interest. Since most of the components going into the system have been reported elsewhere,

and I'm still not entirely sure this system could be replicated from what's provided, I feel like it really comes down to whether the existence of the artifact is sufficiently important. If someone documented a machine that was indistinguishable at debating from an expert human, I would say something very important had happened. These results are weaker than that, and I think it's open for debate (so to speak) how impressed we should already be.

Author Rebuttals to First Revision (please note that the authors have quoted the reviewers in italic font and responded in plain font):

2. Response to the Second Referee

2.1 I have now evaluated the revision and response letter. I appreciate the detailed and constructive responses. I will largely restrict myself to the issues I raised in my initial review.

2.2 First, regarding replicability. I appreciate that a great deal more information is now included in the appendices. I'm still not completely certain that I (or another AI expert) could replicate the system from this description. (Though it's a bit hard to tell without going through this and previous papers in a great deal of detail.) It seems pretty likely that this is by design since there is no intention to open source (and key components are to be available as a service-for-fee). This hazy line between openness and competitive advantage is probably the inevitable result of doing science that is also directly part of a for-profit endeavor. I am personally uncomfortable with this, but I will leave it to the editor to decide if it is acceptable for Nature.

Response: Project Debater is a complex system, integrating multiple components. The Supplementary Materials describe these components and their integration in detail and provide references to previous publications which give even further details. Moreover, as noted in the code availability statement, most of the components will be made **freely** available for academic research purposes as part of this publication. These include Claim and Evidence detection, Claim boundary detection, Stance detection, Theme extraction, and Narrative generation, as well as additional NLP capabilities described in the Supplementary, such as the Wikification tool, the Term Relatedness scorer, and an associated text clustering utility. It is true that trying to reproduce the system will probably not yield an exact copy as some fine details may still be not described in full in this manuscript and require making an educated decision. Yet, we believe it is fair to suggest that the details and access provided here will suffice to construct a close approximation of Project Debater, that will also decently perform in a debate with a human expert debater. While the speeches obtained by such an approximation may not be identical to those generated by Project Debater, we expect that assessing these speeches will yield very similar results to those reported in Figure 3 of the present work.

2.3 Second, regarding evaluation. I appreciate the work that was put into adding evaluations. I think they are vastly better than the previous evaluations, though I have some questions about the new additions.

2.3.1 First some statistical questions: The error bars in fig 3 are surprisingly small for 15 responses on a 5 point scale. Partly this may be because they are standard errors, which are not appropriate for this type of data (stderr assumes normal data; use eg bootstrapped CI instead). But even so, they look small to me, and the low kappas indicate higher variance as well. (Any reason not to provide the anonymized data, so people can explore such questions?) Relatedly it is not reported what test the significance values come from. Please describe the statistics in more detail!

Response: We thank the referee for this comment. After careful consideration, we have accepted the suggestion to use bootstrapped CI, and we now show the 95% CIs as the error bars. We note that each bar is based not on 15 responses but rather on 78 times 15 responses (78 times 20 in Panel b), since we consider 78 motions in our evaluation. It is therefore not surprising that the error bars are not very large. We have also updated the text in the main paper (Figure 3 legend) and in the Supplementary to describe the calculations of the error bars and significance analysis in more detail. Finally, per the suggestion of the referee we will share the anonymized data as part of this paper.

2.3.2 The evaluation of fig3a seems to me like a simple but sound test of the overall quality of the initial statement. The baseline models are a good-faith attempt to put together reasonable representatives and the comparison to human experts is important. Overall this does look like it supports the claim that the system provide "decent performance" at generating initial statements.

2.3.3 The test of the full model, ie fig 3b, seems a bit less compelling. For one thing, if I understand correctly, the S2 human response is not a response to the particular S1 generated by the model. These dialogues are thus already questionably coherent and the model's ability to coherently respond in S3 might not be well tested. The baselines also strike me as much weaker for this experiment.. I'm not sure why we'd expect them to be coherent at all. Further there is no human control for this experiment making it more difficult to tell how well the model is actually doing. (Being judged "decent" in a little over half the motions doesn't seem all that great.) Just to make clear what is probably obvious, this direction of evaluation is a positive step, but it's much weaker than "human level" performance — coherence, persuasion, fluidity.

Response: Evaluating an interactive debate is indeed more challenging than evaluating the opening statement alone. We therefore opted for keeping a single fixed S2 human speech per motion, enabling a stable comparison between Project Debater and the two controls, and a stable evaluation of the progress over time. Since the S3 speech only

rebuts S2, and does not directly consider S1, we believe this is a fair approximation for this purpose. Regarding the two controls in Figure-3b, we note that since to the best of our knowledge Project Debater is the only available system that can interactively debate, we could not compare to any relevant baselines. However, we did find it important to present controls, i.e. "systems" that as the referee noted, are not expected to perform well and serve only to confirm the overall validity of the annotations provided by the crowd annotators. We tried to better clarify this point in the text.

Regarding human control, unfortunately, we do not have enough human S3 speeches, to enable a meaningful comparison to human controls. It is worth noting, that the evaluation of the opening speech of Project Debater reported in Figure-3a is well aligned with its overall score reported in Figure-3b (in both evaluations, Project Debater achieves a score around 4). Based on this observation, as well as on the fact that the overall score is highly dependent on the level of the opening speech, we do not expect the human performance to very much surpass 4.2, although this is not a sound statement and hence not included in the manuscript.

2.3.4 A final minor point on evaluation: with these more clear evaluations added, I would think the "Progress Over Time" Eval could be cut — it's an internal development metric that isn't properly controlled, so doesn't provide much information.

Response: We accepted the suggestion by the referee and correspondingly moved Figure-3c and the associated analysis to the Supplementary.

2.4 Finally, regarding clear statement of the scientific claims. As indicated in the response letter, the claim is "results indicating this system can decently perform in a debate with a human expert debater." This is clear, which I appreciate, though to me it isn't the most interesting kind of scientific claim. It is an existence claim for an artifact with a certain property. It isn't a claim that, for instance, certain principles or ideas lead to such an artifact. The value then comes down to whether the existence of the artifact, or the techniques described, are of profound scientific interest. Since most of the components going into the system have been reported elsewhere, and I'm still not entirely sure this system could be replicated from what's provided, I feel like it really comes down to whether the existence of the artifact is sufficiently important. If someone documented a machine that was indistinguishable at debating from an expert human, I would say something very important had happened. These results are weaker than that, and I think it's open for debate (so to speak) how impressed we should already be.

Response: We agree with the referee that a machine capable of debating at a level and manner indistinguishable from those of an expert debater would be an incredible achievement, and that the work presented does not fully attain this goal. Nonetheless, we believe that the results depicted in Figure-3 show that this work is a considerable step toward this goal. Consider, for example, GPT-2, the current state-of-the-art in text

generation that is widely available. The quality of Project Debater speeches is much closer to human performance than to this benchmark, even after fine-tuning GPT-2 over thousands of speeches by human debate experts, or over tens of thousands of high- quality arguments authored by humans. Moreover, Project Debater also generates effective rebuttal speeches, which are an even greater challenge than opening speeches, as they need to address the main issues raised by the opponent. Hence, we believe that even at its current state, Project Debater is breaking new grounds and suggests not only that the lofty goal of human-level performance may be attainable, but also a route towards attaining it.

3. Additional Updates

3.1 We made additional small updates and edits to the Supplementary Materials, including updating the title to match the new title of the main paper, adding a Table of Contents, etc.

Reviewer Reports on the Second Revision:

Referee #2 (Remarks to the Author):

Thank you for clarifying and updating the statistics for fig 3a. The result of 3a seems fine, and supports the claim the project debater is decent at generating first statements. Great result!

Regarding the evaluation in fig 3b, using fixed S2 per motion still troubles me. If indeed as claimed "the S3 speech only rebuts S2, and does not directly consider S1" then the correct evaluation would be an evaluation of S3 given S2, not of overall performance (given also S1). But more problematic for me is the controls: it's fine to include some ablations of the system as controls, but you also have to include some stronger controls. In particular, in a setting where your system is the only one you can think of to do a task, you really need to compare to human performance. It is a weak justification to say "we do not have enough human S3 speeches, to enable a meaningful comparison to human controls" — gather this data! Thus, the rather critical result of fig 3b is still underwhelming, and doesn't convince me that "Project Debater also generates effective rebuttal speeches, which are an even greater challenge than opening speeches".

To summarize, the paper has been clarified on many points and important details added to the supplement. The evaluations have been improved so that at this point I see support for the claim that this system can decently perform at generating first statements in a debate; but I don't see strong enough support for the claim that "this system can decently perform in a debate with a human expert debater".

Author Rebuttals to Second Revision (please note that the authors have quoted the reviewers in italic font and responded in plain font):

1. Response to the Second Referee

Thank you for clarifying and updating the statistics for fig 3a. The result of 3a seems fine, and supports the claim the project debater is decent at generating first statements.

Great result! Regarding the evaluation in fig 3b, using fixed S2 per motion still troubles me. If indeed as claimed "the S3 speech only rebuts S2, and does not directly consider S1" then the correct evaluation would be an evaluation of S3 given S2, not of overall performance (given also S1). But more problematic for me is the controls: it's fine to include some ablations of the system as controls,

but you also have to include some stronger controls. In particular, in a setting where your system is the only one you can think of to do a task, you really need to compare to human performance. It is a weak justification to say "we do not have enough human S3 speeches, to enable a meaningful comparison to human controls" — gather this data! Thus, the rather critical result of fig 3b is still underwhelming, and doesn't convince me that "Project Debater also generates effective rebuttal speeches, which are an even greater challenge than opening speeches". To summarize, the paper has been clarified on many points and important details added to the supplement. The evaluations have been improved so that at this point I see support for the claim that this system can decently perform at generating first statements in a debate; but I don't see strong enough support for the claim that "this system can decently perform in a debate with a human expert debater".

Response: We agree with the concern regarding whether the evaluation fully supports the statement "this system can decently perform in a debate with a human expert debater". Correspondingly, and following the editor's advice, we revised our paper in two ways to address this issue. First, when making this statement (Line 41), we softened the claim by changing the verb 'indicating' to 'suggesting'. Second, we added a new paragraph (Lines 195-200), detailing the limitations of the evaluation and the conclusions that can be drawn from it. We hope that softening the language and being clear and transparent about the limitations of the evaluation, addresses the remaining concern, and we are open to any additional suggestions in this context.

2. Additional Updates

2.1 We added an Acknowledgments statement and made additional minor updates and edits to the main paper and to the Supplementary Materials.