Title

How far can a commercial decision model's probabilities be trusted? A calibration audit of Jev against open and general-purpose classifiers

Author

Meng, Lingsen

Publication Date

2026-09-24

How far can a commercial decision model's probabilities be trusted? A calibration audit of Jev against open and general-purpose classifiers

Lingsen Meng
Department of Earth, Planetary, and Space Sciences, University of California, Los Angeles

Preprint draft, 2026-09-24.

Abstract

Commercial "decision models" return a choice or a probability for a closed question about a passage of text, and they are marketed as calibrated judgments that software can act on directly. We audit one such system, TypeSafe's Jev 1.13.0, with 3,244 evaluations of 2,372 distinct texts from two public classification tasks, and compare it with two open zero-shot classifiers (one of them trained with examples from these tasks) and two small general-purpose language models. The top-label calibration of Jev's probabilities differs between tasks and question forms, so it has to be checked for each: the calibration error changes by 0.071 between two ways of posing the same question, the middle of the four-class scale is overconfident (stated 0.750, observed 0.612), and 71.6% of four-class answers sit at exactly 1.00, which leaves a threshold on the top-class probability 53 operating points. One fitted temperature removes most of the miscalibration where it exists, and split conformal prediction reaches its marginal target, although its single-label answers are wrong 5.93% of the time at a nominal 5%. Against an open checkpoint trained without examples from these benchmarks, Jev is more accurate in all 16 combinations of label wording and decision rule on both tasks, with an exploratory full-test-set estimate of +5.72 points (95% interval [+4.70, +6.75]) on AG News. On AG News it is less accurate than a checkpoint trained with examples from these tasks (−2.30 points [−3.33, −1.28]), and on sentiment the two are not distinguishable. The benchmarker's choices matter as much: the label descriptions alone moved the gap from +6.93 to +9.67 points, and the choice of open checkpoint moved it by ~8 points and reversed its sign. Returned scores should therefore be validated for the target task, question form and decision rule, and recalibrated where that validation shows it is needed.

1. Introduction

A new class of commercial model is sold to programs and not to readers. An application sends a passage together with a closed question, and the model returns a choice, a score or the probability of a yes/no proposition, along with a number called `confidence`. The value proposition is that these outputs are calibrated, i.e. that software can route confident cases automatically and send the rest to a person. Whether this holds depends on whether the probabilities behave as probabilities, which is an empirical question that the accuracy and latency benchmarks usually reported for such systems do not answer.

Calibration of neural classifiers has been studied extensively (e.g. Guo et al., 2017; Naeini et al., 2015), and so have the tools for using an imperfect score safely, including temperature scaling

(Guo et al., 2017), selective classification (Geifman and El-Yaniv, 2017) and conformal prediction (Vovk et al., 2005; Angelopoulos and Bates, 2021). However, these tools are usually applied to models whose logits are available. A commercial decision API exposes only what it chooses to return, here probabilities rounded to 0.01 and a derived `confidence` statistic, and the vendor's own documentation advises users to choose thresholds themselves and to test on their own data. We are not aware of a published calibration measurement of such a system whose per-item outputs are released for checking.

Our questions also touch four related lines of work. The probabilities of language models have been found to be informative but imperfectly calibrated, and confidence stated in words is often overconfident (Kadavath et al., 2022; Xiong et al., 2023). Classification by prompting is sensitive to label wording and formatting (Zhao et al., 2021; Sclar et al., 2023), an effect we measure here on the side of the benchmark and not of the model. Comparisons between models are confounded when a model has seen a task's data in training (Sainz et al., 2023; Deng et al., 2023). Commercial language-model APIs have also been audited from the outside, e.g. for changes in behaviour over time (Chen et al., 2023). However, none of these studies examines a decision API whose product is a quantised probability, which is the object of this study.

In this study we ask three questions. First, are the returned probabilities calibrated, and what does it take to make them usable? Second, how much more accurate is the commercial system than free alternatives asked the same questions? Third, can any of the systems support an automatic tier with a stated error rate? Our analysis of the second question shows that the answer depends as much on the benchmarker's choices as on the systems, and we report that dependence as a result in its own right.

2. Systems and data

2.1 Tasks

We use two public English text-classification tasks. SST-2 (Socher et al., 2013), in the GLUE development split (Wang et al., 2018), contributes 872 movie-review sentences. AG News (Zhang et al., 2015) contributes a sample of 1,500 items drawn with a fixed seed from the 7,600-item test set, with the title and description joined as `title. description`. For the calibration analysis of Jev the SST-2 items were asked in two forms. The yes/no form uses the question type `noul` with the instructions "Is the overall sentiment of this movie-review excerpt positive?". The choice form uses the question type `choice` with the instructions "What is the overall sentiment of this movie-review excerpt?" and the option criteria "favourable overall" (positive) and "unfavourable overall" (negative). The two forms therefore differ in the instructions and the option criteria as well as in the question type, and we compare them as complete question forms, not as a controlled change of the type field alone. AG News was asked as a four-way choice (Section 2.3). These three conditions give 3,244 evaluations of 2,372 distinct texts (872 sentences asked in both forms, and 1,500 news items). In addition, the second set of label descriptions in Section 2.3 adds 2,372 further calls, for 5,616 Jev calls in total.

2.2 Systems

Table 1 lists the five systems. Jev was called through its public API with the model version pinned. The two DeBERTa-v3-large checkpoints (He et al., 2021) are zero-shot classifiers built on natural

language inference (Yin et al., 2019; Laurer et al., 2023). They differ in their training data in a way that matters for any comparison: according to its model card, the `-c` checkpoint was trained without supervised examples from these benchmarks, whereas the released weights of the other checkpoint were trained with up to 500 examples per class from each of 28 datasets, AG News and Rotten Tomatoes among them (the source of the SST-2 sentences), and with additional NLI data. We therefore call the latter task-exposed. The card states that no test data were used for either, and the two checkpoints differ in more than task exposure, so the difference between them is not a controlled measurement of task exposure alone. Neither pretraining exposure nor Jev's own training data could be examined in this study. For the two OpenAI models, class probabilities were read from the `top_logprobs` of a single output token.

Table 1: Systems compared.

system	description
Jev 1.13.0	commercial decision model (TypeSafe), returns probabilities on a 0.01 grid and a `confidence` statistic
DeBERTa-v3-large zeroshot v2.0 **`-c`**	open NLI classifier, trained without supervised examples from these benchmarks according to its model card
DeBERTa-v3-large zeroshot v2.0	same family, trained with up to 500 examples per class from 28 datasets including these tasks (task-exposed)
gpt-5.4-nano, gpt-4.1-nano	small general-purpose models, probabilities from `top_logprobs`

2.3 Label descriptions

Every system is told what the labels mean. For the open classifiers the description is a hypothesis sentence, and for Jev it is a short criterion per option. We use two sets for each of Jev and the two open checkpoints: our own descriptions (e.g. "business, economy, markets, companies" for the business class), and the descriptions the open model's author used to produce the model card (e.g. "This example news text is about business news"), which we apply to Jev by taking the phrase after "is about". With the author's descriptions, our pipeline reproduces all four model-card figures on the author's own 2,000-item evaluation split to within 0.001 in macro-F1 (`-c`: 0.8192 on AG News against the card's 0.819, and 0.8695 on Rotten Tomatoes against 0.869; the task-exposed checkpoint: 0.8978 against 0.898, and 0.9080 against 0.908). The author's descriptions are thus a shared reference that neither we nor the vendor chose, although not a demonstrably neutral or optimal one. The OpenAI models were run with one fixed prompt. Sections 4.1 and 4.2 use our descriptions for Jev.

3. Methods

Calibration. We report accuracy and the top-label expected calibration error (ECE) over 15 equal-width bins of the top-class probability, each with a 95% item-bootstrap percentile interval (5,000 resamples for accuracy and 1,000 for ECE), whereas accuracy within a subset of items (a reliability bin or a pooled band) carries a Wilson interval (Wilson, 1927). We checked that the conclusions do not depend on the binning scheme, and Appendix A gives the exact definitions and conventions used throughout. Accuracy uses the choice the API reports, not an argmax recomputed from the rounded probabilities, which differ on two AG News items returned as exact ties. Temperature scaling (Guo et al., 2017) is fitted on a random half and evaluated on the other, 200 times.

Selective prediction. We compare systems by the error rate on the accepted share of items at fixed coverage, accepting items in decreasing order of the top-class probability and breaking ties at random in expectation, and by the area under the risk–coverage curve (AURC), with paired item-bootstrap intervals (2,000 resamples). Comparing the lowest error rate each system can reach is not a fair comparison, because the minimum over thresholds is biased downward by an amount that grows with the number of distinct scores, and because different systems reach their minima at very different coverage.

Certification of an automatic tier. To test whether a system supports a promised error rate on automatically routed traffic, each random split divides the labelled items into three disjoint thirds. On the first third we choose the widest threshold whose observed error meets the target; on the second we test that single threshold with a one-sided Wilson upper bound at $z = 1.96$, which is nominally a 97.5% bound, more conservative than a 95% bound, and an approximate, not an exact finite-sample, bound; the last third is used for evaluation. The observed error is compared with the target exactly, in integers. We report the number of certifying splits out of 2,000 and the expected share of traffic automated when every failed split sends all items to a person. A scan that keeps the widest threshold whose individual bound meets the target is not a valid certification: on an uninformative score with a true error rate of 5.01% it certifies a 5% tier in 9.9% to 13.2% of simulated samples, against 1.7% to 2.1% for a single test of one pre-chosen threshold.

Conformal prediction. Split conformal sets (Vovk et al., 2005; Angelopoulos and Bates, 2021) are built on a random half and evaluated on the other, over 4,000 splits, with the nonconformity score one minus the probability of the true class. Because the probabilities are quantised, we report both the deterministic rule and the randomised (smoothed) tie-break on the same splits, and the error rate inside the subgroup of items that receive a single-label set.

Cross-system accuracy. Paired differences use the exact McNemar test (McNemar, 1947). We cross two label descriptions with two decision rules for each system, giving 16 cells per comparison. The first rule is the answer as delivered. The second is an unsupervised prior correction in the spirit of Batch Calibration (Zhou et al., 2023) and contextual calibration (Zhao et al., 2021): each class's probability is divided by that class's mean probability over a disjoint random half of the batch, and no labels are used. Zhou et al. write their correction as subtracting a batch mean, and the two forms are not equivalent in general; we use the multiplicative form because it needs no constant for the exact zeros that Jev and the OpenAI models return. A cell is counted as significant if $p < 0.01$ in at least 95% of 200 random splits; since all cells share the same items, the count is a stability statistic, not a set of independent confirmations.

Full-test-set estimate. The 1,500-item sample turned out to be somewhat harder for `-c` than the full test set. Since both open checkpoints were run on all 7,600 items and Jev only on the sample, we estimate the full-test-set difference with a regression (GREG) estimator (Särndal et al., 1992), using both checkpoints' per-item correctness and the class labels as auxiliary variables, with a linearisation standard error and finite-population correction. Every system receives exactly the text Jev received. The comparison for this estimate (both systems on the author's descriptions, answers as delivered) was fixed before the corresponding Jev run, and the protocol and analysis code for that run and for the full-set open-model run were hashed before they were executed.

Table 2: Jev's accuracy and calibration, with 95% item-bootstrap intervals.

task	accuracy	ECE (15 equal-width bins)
SST-2, yes/no form	0.948 [0.933, 0.962]	0.091 [0.078, 0.104]
SST-2, choice form	0.961 [0.947, 0.974]	0.020 [0.014, 0.035]
AG News, four-way	0.867 [0.850, 0.884]	0.089 [0.076, 0.108]

4. Results

4.1 Jev's probabilities as returned

The question form changes the calibration. The two SST-2 rows are the same 872 sentences answered by the same model under the two complete question forms of Section 2.1, which differ in their instructions and option criteria as well as in the question type. The observed difference in ECE is 0.071, with a bootstrap mean of 0.067 and an interval of [0.047, 0.086], and its direction does not change under any binning scheme we tried. The yes/no form is underconfident (Figure 1a), since its probabilities are not sharp enough, whereas the choice form is close to calibrated as returned. Thus the calibration measured on one question form cannot be assumed for the other, even though the model and the items are identical, and whether a calibrator fitted on one form transfers to the other would itself need validation.

On the four-way task the middle of the scale is overconfident as a block. Every AG News bin below the top one, i.e. below a maximum probability of 0.9333, pools 258 items with a mean stated confidence of 0.750 against an observed accuracy of 0.612, interval [0.552, 0.670]. The ordering within that range is not resolved: the eight pooled bins hold 2, 18, 24, 31, 37, 29, 48, 69 items, and a monotonicity test returns $p = 0.76$ with power 0.16 against the observed deviation, so a non-monotonic reliability curve can be neither shown nor ruled out.

The probabilities are quantised. Every returned probability is an integer multiple of 0.01. On AG News 71.6% of items come back with the top class at exactly 1.00, and those items are 94.3% correct; on the SST-2 choice task 61.7% claim 1.00 and are 99.3% correct. A threshold cannot cut inside a bucket, so a threshold on the maximum probability reaches 53 distinct operating points on 1,500 four-class items (Figure 1b). Thresholding on `confidence` instead reaches 68 operating points, and ordering by maximum probability with `confidence` as the tie-break reaches 108, so the quantisation constrains the obvious routing rule, although it is not a hard limit on what the API makes routable.

The `confidence` field carries information the rounded probabilities do not. The vendor's documentation describes `confidence` only as a statistic computed from the probability distribution (TypeSafe, 2026). However, it is not a function of the probabilities as returned: on AG News, 51 distinct probability vectors map to more than one `confidence` value, and its correlation with the maximum class probability is 0.99976, with 9 discordant item pairs out of 544,089; on the SST-2 choice task it is exactly monotone in the maximum probability (0 of 229,233 pairs) and the correlation is 0.99957. Inside the saturated 1.00 bucket it is the only available handle (Table 3). The split runs in the useful direction on AG News, where the subgroup is small and the difference is not significant (Fisher $p = 0.069$), and it runs the other way on SST-2. We do not attempt to identify how `confidence` is computed, since the vendor's post-processing is not observable; the evidence above establishes only that it is not redundant with the rounded probabilities. Since `confidence` is itself on a 0.01 grid, the extra resolution it offers is likely a factor of between 1.3

and 2.

Table 3: Error rate inside the 1.00 bucket, split by `confidence`.

subgroup of the 1.00 bucket	n	error rate
AG News, confidence = 0.99	28	14.3%
AG News, confidence = 1	1046	5.45%
SST-2 choice, confidence = 0.99	60	0.0%
SST-2 choice, confidence = 1	478	0.84%

One fitted parameter removes most of the miscalibration. Temperature scaling fitted on held-out halves (Table 4) largely repairs both tasks that start miscalibrated, whereas the SST-2 choice task was already calibrated and does not improve (its interval, [−0.001, +0.026], spans zero). The fitted temperature is, however, not a property of the model. On AG News 66.9% of the returned probability entries are exactly 0.00 and must be clipped before taking logarithms, and moving the clipping threshold walks the fitted temperature from 1.28 through 1.82, 2.36, 3.47 and 5.20 to 6.95, so only its direction relative to one is trustworthy. The SST-2 yes/no form has no exact zeros and its T = 0.47 is unaffected.

Table 4: Held-out ECE before and after temperature scaling (200 random halves).

task	as returned	after temperature scaling	fitted T
SST-2 yes/no	0.091	0.023	0.47
SST-2 choice	0.024	0.034	1.63
AG News	0.092	0.028	3.46

4.2 Conformal prediction

Split conformal prediction reaches its marginal target, but the quantisation puts many items exactly on the threshold, so the tie-breaking convention matters (Table 5; Figure 1c). The deterministic rule over-covers and the randomised rule sits on target, and their singleton rates differ by 25.8 points (Monte-Carlo standard error 0.44) on the same splits. Neither rule is wrong, but a singleton rate quoted for a quantised system is uninterpretable without the rule. More importantly, a 95% marginal guarantee does not mean that the single answers are right 95% of the time: coverage is an average over all items, the items that receive a single label are a selected subgroup, and here they are wrong 5.93% of the time under the randomised rule and 6.38% under the deterministic one. On the binary task 1.5% of the sets are empty, which is legitimate in conformal prediction but requires a fallback in a two-way decision.

Table 5: Split conformal prediction on AG News at a 95% target (4,000 shared splits).

rule	coverage	mean set size	single-label sets	error inside single-label sets
randomised (smoothed)	0.951	1.40	66.2%	5.93%
deterministic	0.970	2.51	40.4%	6.38%

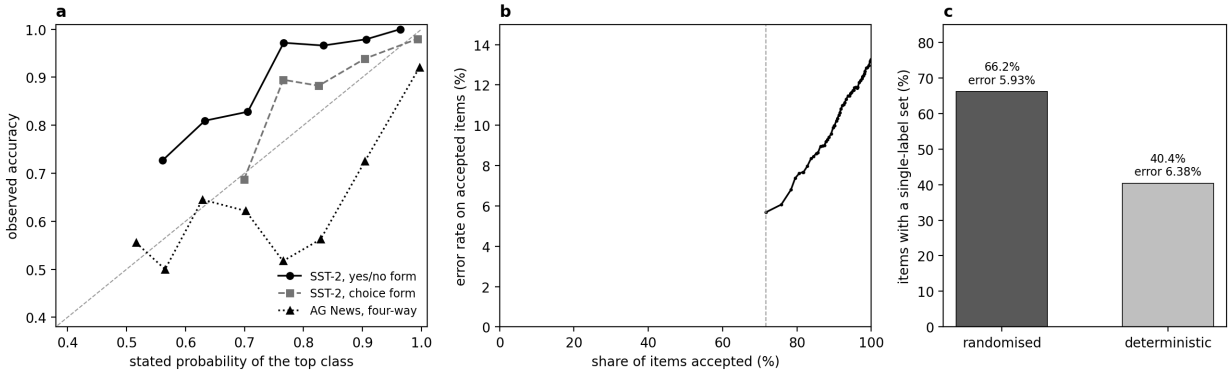

Figure 1: Jev's probabilities as returned. **a**, Reliability for the three tasks; the yes/no form lies above the diagonal and the four-way task below it across the middle of the scale. **b**, Error rate against coverage on AG News for a threshold on the returned top-class probability; no non-empty threshold of this kind reaches coverage below 72%, because the 1.00 bucket holds 71.6% of items. **c**, The same predictor at a 95% target under the two conformal tie-breaking rules, with the error rate inside single-label sets.

Table 6: Accuracy on the evaluation samples.

accuracy	SST-2	AG News
Jev, our label wording	0.961	0.867
Jev, the open model author's label wording	0.960	0.877
`-c`, our wording	0.905	0.771
`-c`, its author's wording	0.911	0.808
task-exposed, our wording	0.959	0.901
task-exposed, its author's wording	0.955	0.907
gpt-5.4-nano	0.948	0.811
gpt-4.1-nano	0.938	0.851

4.3 Accuracy against open and general-purpose classifiers

The label descriptions move the comparison. Much of what AG News files under Sci/Tech is news about technology companies, so the extra nouns in our business description pull Sci/Tech items into Business: under our wording `-c` sent 196 of the 393 Sci/Tech items there, and Jev sent 101. Changing only the label descriptions moved `-c` by 3.7 points on AG News and Jev by 1.0 point (21 items gained against 6 lost, exact McNemar $p = 0.0059$; the two Jev runs were made four days apart with the version pinned, so part of that change may reflect the date rather than the wording). Our wording therefore hurt the open model more than the commercial one, and the measured gap depends on which system received which description (Table 7). The cell with both systems on our wording is about 40% larger than the cell in which neither side's wording was chosen by us.

Table 7: Jev minus `-c`, AG News, both uncorrected, 1,500-item sample.

label descriptions	Jev minus `-c`
both with our wording	+9.67 points
both with the author's wording	+6.93 points
`-c` with the author's, Jev with ours	+5.93 points

The direction is stable across all 16 configurations. Table 8 summarises the full grid of label descriptions and decision rules (Figure 2a). Jev is more accurate than `-c` in every cell on both tasks, and less accurate than the task-exposed checkpoint in every cell on AG News, whereas on SST-2 no cell resolves a difference from the task-exposed checkpoint.

Table 8: Jev minus each open checkpoint over 16 wording × rule cells, on the evaluation samples.

Jev minus…	range over 16 cells	cells with $p < 0.01$ in ≥95% of splits
`-c`, AG News	+4.16 to +11.14 points	16 of 16
task-exposed, AG News	−4.06 to −1.92 points	16 of 16
`-c`, SST-2	+4.93 to +5.74 points	16 of 16
task-exposed, SST-2	+0.08 to +0.69 points	0 of 16

The size, estimated for the full test set. For the comparison fixed in advance (both systems on the author's wording, answers as delivered), with every system given identical input text, Jev is more accurate than `-c` by +5.72 points, 95% interval [+4.70, +6.75], and less accurate than the task-exposed checkpoint by −2.30 points [−3.33, −1.28], and both are exploratory estimates in the sense set out in Section 6. With Jev on our wording against the open models on the author's wording, the same estimates are +4.82 and −3.21. When the open models were instead run on the full set with the text HTML-unescaped and joined with a space, and thus with input different from Jev's, the estimates were +5.27 and −2.27, so the input formatting alone moves the estimate by ~0.5 points. SST-2 has no larger pool to extrapolate to, and there the sample difference is +4.93 points against `-c`.

4.4 Selective prediction at matched coverage

The lowest error rates the systems reach on AG News are not comparable: `-c` reaches 1.9% by accepting 13% of the items, whereas Jev's lowest rate comes at 69% coverage, since a deterministic threshold on its top-class probability cannot single out any smaller slice (69.0% of items sit at

1.00 under the author's wording). At matched coverage the picture is different (Table 9; Figure 2b). `-c` is cleaner than Jev only on the smallest slices (by 3.87 points at 13% coverage), and on the sample curve Jev makes fewer errors at every coverage from 34% upward, with an area under the risk–coverage curve lower by 1.81 points [0.84, 2.79]. The task-exposed checkpoint's curve lies below Jev's at every coverage from 5% upward (at 13% the difference is not distinguishable from zero), and its area is lower by 2.48 points [1.45, 3.44]. The 13% and 72% columns mark where the systems' minima lie on the two wordings and 34% is where the curves were observed to cross, so none is a prespecified threshold, and the intervals on the areas do not imply that the ordering holds simultaneously at every coverage. Ranking Jev's answers by its `confidence` field instead of the top-class probability shrinks the tied block to 67.1% and lowers its error at 13% coverage and its area from 5.41% and 6.30% to 5.16% and 6.13%, respectively, which changes none of these comparisons. The comparison with the task-exposed checkpoint is therefore a ranking, and the comparison with `-c` is a statement about resolution: the zero-shot checkpoint can isolate a small, clean slice, and a threshold on Jev's returned scores cannot isolate a slice that small.

Table 9: AG News error rate on the accepted share, all systems on the author's wording.

AG News, error rate on the accepted share	13%	50%	72%	all	area under the curve
Jev, author's wording	5.41%	5.41%	5.90%	12.27%	6.30%
`-c`, author's wording	1.54%	7.20%	11.39%	19.20%	8.11%
task-exposed, author's wording	3.08%	2.80%	3.80%	9.27%	3.82%

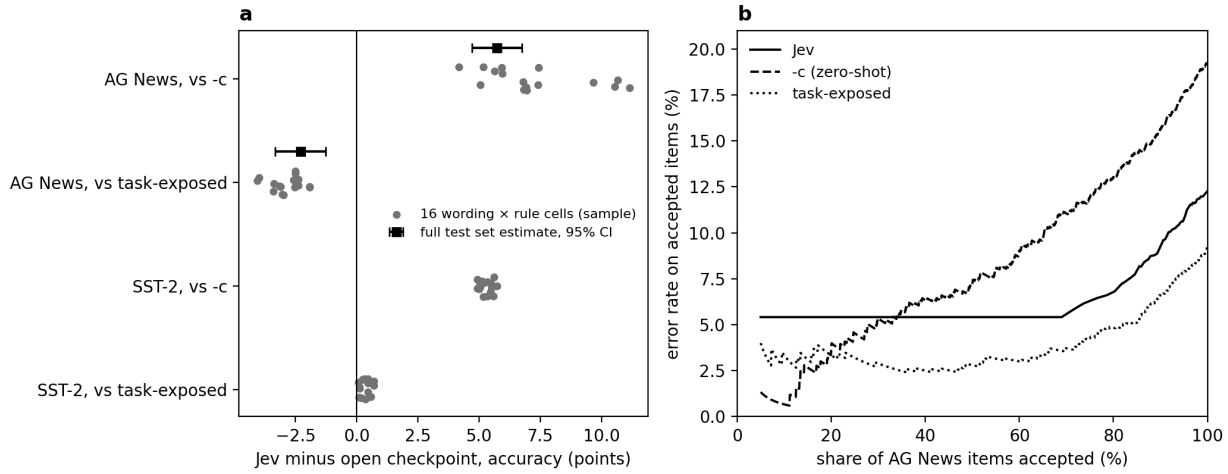

Figure 2: Jev against the open checkpoints. **a**, Jev's accuracy minus each checkpoint's, in points. Grey dots are the 16 combinations of label description and decision rule on the evaluation sample; black squares are the full-test-set estimates for AG News (both systems on the author's descriptions, answers as delivered, identical input text) with nominal 95% intervals. **b**, AG News error rate on the accepted share of the sample, accepting items in order of the top-class probability with ties broken at random, all systems on the author's descriptions; curves are plotted from 5% coverage upward.

4.5 Certification of an automatic tier

With the procedure of Section 3, over 2,000 random splits of AG News, a 5% tier certifies for Jev in 0 splits under either wording, for `-c` in 23, and for the task-exposed checkpoint in 95 (both open

checkpoints on the author's wording). Since the splits that fail send everything to a person, the expected share of traffic automated is 0.2% for `-c` and 3.3% for the task-exposed checkpoint, and a 2% tier certifies for no system in any configuration. On the easier SST-2 task, Jev on the author's wording certifies a 5% tier in 191 of 2,000 splits and automates 9.4% of traffic in expectation (220 splits and 10.9% on our wording), against 125 splits for the task-exposed checkpoint and 61 for `-c`.

These rates belong to the procedure as much as to the systems. Choosing the widest threshold that just meets the target on the first third leaves little statistical margin for the test on the second. In an exploratory sensitivity check that changed only this selection rule, to the threshold with the lowest upper bound on the first third, the rates changed substantially and not uniformly. On AG News the task-exposed checkpoint rose from 95 to 423 certifying splits and its expected automation from 3.3% to 10.5%, `-c` fell from 23 to 13, and Jev rose from 0 to 4 under either wording; in those four splits Jev's held-out error on the author's wording averages 6.1%, above the 5% target. On SST-2 Jev on the author's wording rose from 191 to 1,776 splits and its expected automation from 9.4% to 61.9%, the task-exposed checkpoint from 125 to 1,920, and `-c` from 61 to 251. The 2% result is also bounded by the data budget, since even with no errors a bound at $z = 1.96$ needs at least 189 accepted items in the test third, and a 13% slice of a ~500-item third holds ~65. Low success of one threshold-learning rule therefore does not establish that no certifiable operating point exists.

The object of such a certification is the population selective risk of the chosen threshold, that is, its error rate on future items from the same distribution, under the sampling assumptions of the bound (Geifman and El-Yaniv, 2017; Angelopoulos et al., 2021). A valid bound certifies a threshold whose population risk exceeds the target only with small probability, and it does not require every finite evaluation batch to meet the target. On the sample, Jev's error in its 1.00 bucket under the author's wording is 5.41% (56 of 1,035 items, 69.0% coverage), above the 5% target. However, a sample error rate is not the population risk, and 4 passing splits out of 2,000, below the nominal 2.5%, cannot validate the confidence level of our approximate Wilson procedure. Four splits are also too few, and too conditionally selected, to say anything about deployment performance, so we treat these certification rates as descriptions of one procedure on this sample and not as finite-sample guarantees.

The conformal results of Section 4.2 give a related comparison. Averaged over 4,000 random splits, the error rate inside the single-label subgroup at 95% coverage on AG News is 5.93% for Jev, 7.43% for `-c` and 5.21% for the task-exposed checkpoint, all on the author's wording (pooling all singletons across splits instead gives 6.05%, 7.53% and 5.25%, respectively). The three systems commit to a single label on 69.4%, 51.6% and 83.6% of items, so these are different subgroups and not a comparison at equal coverage.

4.6 The OpenAI models depend on a convention

For the OpenAI models the probabilities are read from the top-ranked output tokens, and the conventional recipe folds those onto the option set and renormalises. This step is not neutral: gpt-4.1-nano certifies a 5% AG News tier in 0 of 2,000 splits with the probabilities left raw and in 70 with them renormalised, and its error rate on the top 13% of items is 9.23% raw against 2.84% renormalised, on either side of Jev's 5.41%. We therefore report no directional selective-prediction result for it. gpt-5.4-nano returns a single candidate token on 89% of calls, and on 216 of the 284 AG News items it gets wrong the true class receives exactly zero visible probability, so its 95% conformal set averages 3.02 of 4 classes. This is a property of what the API exposes under our

request settings, and it cannot be ruled out that the underlying model is better behaved.

5. Discussion

For a deployer. Our results suggest that the probabilities a decision API returns should be validated for the target task, question form and decision rule before they are acted on, and recalibrated where that validation shows it is needed. The top-label calibration moved with the task and with the form of the question, whereas the SST-2 choice form was close to calibrated as returned and temperature scaling did not improve it. Where there was miscalibration, a temperature fitted on a few hundred labelled items removed most of it. However, the fitted value depended on how exact zeros were handled, so it belongs to a version, a task and a question form, not to the model. We note that top-label ECE is not the loss of any particular decision rule, so a low ECE does not by itself show that a given high-stakes subgroup is safe, and a high ECE does not by itself make the scores useless for a low-stakes rule. A few hundred labelled items from the target distribution buy the temperature fit and a conformal construction at once. We note that a 95% conformal guarantee is marginal, and that the answers it commits to were wrong ~6% of the time here; the tie-breaking rule must be stated for any quantised system, since the two conventions differ by ~26 points of singleton rate. The 0.01 quantisation also restricts a threshold rule to a few dozen operating points, which is consistent with Jev's flat risk curve below ~70% coverage and with its failure to certify a 5% tier on the four-class task under our procedure; the `confidence` field offers a modest amount of extra resolution.

For a benchmarker. The comparison with open models was more sensitive to the benchmarker's choices than to anything else we varied. The choice between two checkpoints of the same family, one of them trained with examples from these tasks, moved the comparison by ~8 points and changed its sign, so the training data of any open baseline should be established before a comparison is reported. The label descriptions alone changed the sample gap by about 40%, and they did so without looking unreasonable, which argues for using the open model author's published descriptions and giving the commercial system the same ones. A gap that is stable in direction and variable in size is best reported as a range over a grid of reasonable configurations, with a population estimate for one configuration fixed in advance and with identical input text for every system. Selective prediction should be compared at matched coverage, and a certification rate should be reported together with the threshold-selection rule, the confidence level and the data budget, since each of them moves it.

What the comparison does and does not establish. Within the configurations examined here, Jev was more accurate than an open checkpoint that, according to its model card, saw no examples from these benchmarks, and less accurate on AG News than one that did. This is consistent with a commercial decision model offering a real, moderate accuracy advantage over free zero-shot classification on generic tasks. However, the advantage is of the same order as the effect of a single label description, and it did not translate into a certifiable automatic tier on the harder task under our procedure.

6. Limitations

The study covers two English text-classification tasks and one version of each system; we did not test domain data, other languages, long contexts, latency or cost. Jev and the two open checkpoints were each tried under exactly two label wordings, and the OpenAI models under one fixed prompt;

the sentence that frames Jev's question and the OpenAI prompts were written by us and not varied. The OpenAI probabilities come from a single output token under a short completion budget, which is a proxy for the model's belief and not the object Jev returns. Conformal coverage is marginal and the calibration splits are within-dataset, so none of the results transfers to a shifted domain without remeasurement. The routing intervals come from a bootstrap over a fixed table of scores, and the downstream analyses of the prior-corrected rows estimate the class means leave-one-out over the whole sample, whereas the grid uses disjoint halves, so their uncertainty statements are approximate for the full procedure. The population estimate covers AG News only. Finally, the analysis is exploratory. The evaluation items were used while the analysis was developed, and the decision to give every system identical input text in the full-test-set estimate was made after an exploratory estimate for the same target (+5.74) was available; the +5.72 is therefore a transparently specified exploratory estimate rather than a preregistered confirmation, and its nominal interval does not include the uncertainty from earlier choices of method, wording or task. A confirmatory test would need items not used here, under a protocol fixed in advance.

Broader impact statement

This study evaluates a commercial product through its public interface on public benchmarks, and it reports results that are both favourable and unfavourable to the product. The main risk is that narrow results on two English classification tasks are read as a general verdict that the product is safe or unsafe to deploy, whereas our measurements concern one version, two tasks and the configurations described in Section 6. We therefore state the scope next to each headline number, release the per-item outputs and the cached responses so that any claim can be checked, and recommend validation on the target task in preference to reliance on either the vendor's numbers or ours. The released predictions omit the source texts, which are subject to the datasets' licences, and no personal data were used.

Data and code availability

All API calls are cached on their request body, every random seed is fixed, and a rerun from the cache is bit-identical. The code, per-item predictions, response caches and the hashed protocols are available at https://doi.org/10.5281/zenodo.22935043. SST-2 and AG News carry unspecified licences, and the original AG corpus is restricted to non-commercial research use, so the released predictions omit the source texts and keep row indices, which allows anyone who downloads the datasets to rejoin them. All 5,616 Jev calls in this study cost US$0.086 in total.

AI-assisted research statement

Anthropic's Claude was used as a coding and analysis assistant throughout this project. It implemented and executed the data collection, analysis and statistical code, prepared documentation, and assisted in preparing manuscript text. The author reviewed and takes responsibility for all claims, code and text in this paper.

References

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., & Lei, L. (2021). Learn then Test: Calibrating predictive algorithms to achieve risk control. arXiv:2110.01052.

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv:2307.09009.

Deng, C., et al. (2023). Investigating data contamination in modern benchmarks for large language models. arXiv:2311.09783.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. Advances in Neural Information Processing Systems 30. arXiv:1705.08500.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning. arXiv:1706.04599.

He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543.

Kadavath, S., et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023). Building efficient universal classifiers with natural language inference. arXiv:2312.17543.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157.

Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. Proceedings of the AAAI Conference on Artificial Intelligence, 29.

Sainz, O., et al. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. arXiv:2310.18018.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying language models' sensitivity to spurious features in prompt design. arXiv:2310.11324.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of EMNLP 2013.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). Model Assisted Survey Sampling. Springer.

TypeSafe (2026). Confidence. TypeSafe AI documentation, https://docs.typesafe.ai/confidence (accessed 24 September 2026).

Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic Learning in a Random World. Springer.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158), 209–212.

Xiong, M., et al. (2023). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv:2306.13063.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. Proceedings of EMNLP-IJCNLP 2019.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems 28. arXiv:1509.01626.

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. Proceedings of the 38th International Conference on Machine Learning. arXiv:2102.09690.

Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K., & Roy, S. (2023). Batch calibration: Rethinking calibration for in-context learning and prompt engineering. arXiv:2309.17249.

Appendix A. Definitions and conventions

Top-label ECE. Each item is scored by its top-class probability c and its correctness z, where correctness uses the answer the API reports and, for the yes/no form, c = max(p, 1 − p) with p the returned probability of "yes". The items are assigned to 15 bins of equal width on [0, 1], each closed on the left and open on the right except the last, which includes 1.00, and ECE is the sum over non-empty bins of the bin's share of items times the absolute difference between its mean correctness and its mean c. The question-form difference in Section 4.1 is the difference between the two ECEs on the same 872 sentences, and its bootstrap mean and interval resample sentences jointly for both forms (2,000 resamples).

Temperature scaling. Probabilities are clipped below at a constant before taking logarithms, the rescaled probabilities are proportional to the clipped probabilities raised to the power 1/T, and T is chosen on a grid from 0.20 to 12.00 in steps of 0.01 by minimising the negative log-likelihood of the true class. Table 4 uses a clipping constant of 1e-6 and fits T on 200 random halves, evaluating on the other half. The clip sensitivity in Section 4.1 refits T on all items for constants of 1e-2, 1e-3, 1e-4, 1e-6, 1e-9 and 1e-12. The transformation is monotone within each item, so it never changes which class is chosen.

Risk–coverage curve. Items are accepted in decreasing order of the score. Within a group of tied scores a partial acceptance is charged the group's mean error per accepted item, which is the expectation under random tie-breaking. AURC is the mean of this curve over the n possible acceptance counts, and its intervals come from a paired item bootstrap (2,000 resamples) over the fixed table of returned scores.

Certification. Each of 2,000 splits permutes the items and divides them into thirds A, B and C (the remainder goes to C). The candidate thresholds are the ends of tie groups in decreasing score that accept at least 15 items of A. The widest candidate whose empirical error on A meets the target, compared exactly in integers, is chosen (in the sensitivity check, the candidate with the lowest Wilson upper bound on A). The chosen threshold must then accept at least 15 items of B, and the one-sided Wilson upper bound at z = 1.96 on their error must not exceed the target. C gives the held-out error and coverage of a certified threshold. A split that fails at any step automates nothing, and the expected automation is the mean over all 2,000 splits of the share of C accepted.

Conformal prediction. Each of 4,000 splits divides the items into two halves. The nonconformity score of a class is one minus its probability. With n calibration items and target coverage 1 − a, the threshold is the k-th smallest calibration score of the true class, k = ceil((n + 1)(1 − a)), and

a class enters the prediction set if its score does not exceed the threshold. The randomised rule adds independent uniform noise of width 1e-9 to every calibration and test score, which breaks the ties created by the 0.01 grid without reversing any untied comparison. The deterministic rule adds none, and both rules use the same splits. The singleton error is the error rate among test items whose set has exactly one class, averaged over splits (Section 4.5 also reports it pooled over splits).

The full-test-set estimator. Let j be Jev's correctness on the n = 1,500 sampled items, a simple random sample of the N = 7,600 test items, and let x be the auxiliary vector (both checkpoints' correctness and indicators for three of the four classes), which is known for all N items. With b the least-squares coefficients of j on (1, x) in the sample, Jev's population accuracy is estimated as the sample mean of j plus the difference between the population and sample means of x, weighted by b. The gap subtracts the checkpoint's accuracy on all 7,600 items, which is known exactly. The standard error is the square root of $(1 - n/N)$ times the residual variance (with $n - 6$ degrees of freedom) divided by n, and the interval is ± 1.96 standard errors.

Kinds of interval. Four kinds of uncertainty statement appear in this paper. Item-bootstrap percentile intervals are used for Table 2, the question-form difference and the areas under the risk curves. Wilson intervals are used for accuracy within a subset of items, and a Wilson bound for the certification procedure. Spreads and counts over random splits are used for Tables 4 and 5, the certification counts and the stability count of the 16-cell grid, with a Monte-Carlo standard error where one is given. In addition, the full-test-set estimate carries a linearisation interval. None of these intervals includes variation from querying the systems again, and all except the last are also conditional on the fixed table of returned scores.

Systems and requests. Jev was called with the model version pinned to 1.13.0 and the question forms of Section 2.1, and the label descriptions of Section 2.3 enter as the option criteria. The two DeBERTa checkpoints (MoritzLaurer/deberta-v3-large-zeroshot-v2.0-c, revision b2730f1, and MoritzLaurer/deberta-v3-large-zeroshot-v2.0, revision cf44676) were run in half precision on one NVIDIA RTX 5070 Ti with PyTorch 2.11.0 and Transformers 5.17.0, with each premise–hypothesis pair truncated at 384 tokens. The class probabilities are the softmax over classes of the entailment logits, and the predicted class is the one with the largest. The OpenAI models were called through the chat completions endpoint with a completion budget of 4 tokens and log-probabilities enabled, requesting the top 20 alternatives for gpt-4.1-nano and 5, the limit for that model, for gpt-5.4-nano. The probability of an option is the summed probability of the first-token candidates that match it as a prefix, left raw by default (renormalised over the options in the variant of Section 4.6), and an option absent from the candidates receives zero. Every response is cached on its request body, so the analyses can be replayed without new calls.