UKR Journal of Multidisciplinary Studies (UKRJMS)
Homepage: https://ukrpublisher.com/ukrjms/
Email: submit.ukrpublisher@gmail.com
Volume 2, Issue 7, 2026
ISSN: 3049-432X (Online)

Small-Sample Crop Yield Forecasting with Climate-Aware Multimodal Transformers and LLM Agronomic Explanations

Bella Song

Computer Science, University of Maryland, College Park, MD, USA

***Corresponding Author**: Bella Song

DOI: *https://doi.org/10.5281/zenodo.21623144*

Article History

Original Research Article

Received: 02-06-2026

Accepted: 05-07-2026

Published: 27-07-2026

Citation: Song, B. (2026). Small-sample crop yield forecasting with climate-aware multimodal transformers and LLM agronomic explanations. UKR Journal of Multidisciplinary Studies, 2(7), 137-152.

Abstract

This paper presents a compact empirical study of county-level corn-yield forecasting under a deliberately small-sample setting. The experiment uses the Iowa 2022 corn slice from a CropNet-style USDA crop-label table: a single state, a single crop, a single year, and 97 county observations. The forecasting task estimates held-out county yield in bushels per acre from regional and textual tokens that are available before the target is revealed: agricultural statistics district, parsed direction words, and county-level numeric identifiers. A climate-aware multimodal token-attention model (CAMT) is evaluated against mean, empirical-Bayes, regularized linear, tree-ensemble, gradient-boosting, and neural-network baselines. Across five repeats of five stratified folds, the best purely numeric baseline, ElasticNet with regional tokens, obtained 8.819 bu/acre MAE, 11.479 bu/acre RMSE, and 0.698 mean R2. CAMT obtained 8.822 bu/acre MAE, 11.484 bu/acre RMSE, and the same 0.698 mean R2, while providing token-level evidence that is directly converted into agronomic explanations. The study shows that most predictive information in this state-year slice is carried by regional climate geography: district and direction tokens reduced MAE by about 49% relative to the state-mean baseline. The explanation layer produced concise county rationales grounded in district yield levels and model predictions, allowing the numerical results to be interpreted without using production as a predictor.

Keywords: *Crop yield forecasting; small-sample learning; CropNet; USDA crop labels; climate-aware modeling; multimodal transformer; agronomic explanations; Iowa corn.*

Introduction

Reliable crop-yield forecasting is a practical requirement for market planning, crop insurance, food-security monitoring, and local agronomic decision support. The difficulty is that yield is not only a biological outcome but also a climate-conditioned spatial process. Long-term analyses show that crop yield responds to temperature and precipitation anomalies, and that modern maize systems can become more sensitive to drought even as average productivity improves [1]-[3]. This makes regional context essential: a model that treats each county as an isolated row ignores the fact that neighboring agricultural districts share weather exposure, soils, planting calendars, and management constraints.

Remote sensing and machine learning have expanded the yield-forecasting literature by combining county labels with vegetation indices, satellite imagery, environmental covariates, and neural architectures [4]-[8]. Large multimodal datasets make this attractive, but real operational use often begins with a narrow slice: one state, one crop, one year, and a modest number of county records. The question studied here is therefore intentionally focused. Given a single state-year corn slice, how much predictive signal can be recovered from the crop-label table itself, and can a model explain its forecast in agronomic language without relying on target leakage?

The paper treats the 2022 Iowa corn county slice as a small-sample forecasting problem. In this setting, "forecasting" means estimating the yield of held-out counties from metadata available before their target yield is disclosed. The target is the USDA-style county yield in bushels per acre. Production is retained for descriptive analysis, but it is not used as a predictor because production is an outcome-scale

quantity tied to yield and harvested area. This design keeps the experiment consistent with the county-level facts present in the data and avoids an artificially easy prediction task.

The modeling contribution is a compact climate-aware multimodal token-attention model. CAMT represents each county with three groups of tokens: a categorical agricultural statistics district token, a small numeric county token, and direction-word tokens parsed from district names such as North Central, East Central, and South Central. The district token is not meteorology in the narrow sensor sense, but it is a strong agro-climatic proxy in a single-year slice because it partitions counties into regions that share broad growing-season conditions. The architecture follows the attention principle introduced for sequence modeling [9] and adapted widely in vision and multimodal learning [10], [11], but is deliberately scaled down for fewer than 100 observations.

The empirical goal is not to claim that a single crop-label slice replaces satellite or weather streams. Rather, the goal is to measure what a logically consistent model can do when the available data are restricted to county crop records. Regularized regression [15], [16], robust regression [17], tree ensembles [13], gradient boosting [14], and neural baselines are included to make the comparison transparent. The explanation component follows the model-interpretability motivation of LIME and SHAP [18], [19] and the natural-language reporting capability associated with modern language models [20]-[23]. The resulting explanations are grounded in the model prediction, district statistics, and observed evaluation data; they are not allowed to invent weather events or management practices not represented by the slice.

A second motivation is leakage control. County crop tables often provide both yield and production, and these two quantities are mathematically connected through harvested acreage. A model that treats production as an input for yield prediction can look accurate for the wrong reason, because production already contains outcome information. This paper therefore separates descriptive variables from predictors. Production totals are reported to characterize the slice, but the forecasting models use only regional identifiers and text-derived tokens. The resulting accuracy numbers are lower than leakage-prone alternatives would be, but they correspond to a defensible prediction setting.

The small-sample framing also matters for interpretation. A large transformer trained on a few dozen rows can memorize fold-specific structure, while a mean baseline can ignore real regional differences. The appropriate model lies between these extremes: it must be flexible enough to recognize district-specific yield levels but regularized enough to avoid explaining every county anomaly as a stable signal. CAMT is designed for that balance. It borrows the attention idea from transformer models, but it uses a low-dimensional token representation and a restrained residual layer so that every forecast remains traceable to regional evidence.

Together with the transformer literature [9]-[11], recent evidence from other data-limited domains reinforces this restrained strategy. Few-shot workload forecasting with time-series foundation models emphasizes adaptation when a new tenant provides little history [24], while domain adaptation based on approximately shared features formalizes how related populations can share predictive structure without being treated as identical [25]. Lightweight cross-modal pretraining has likewise used parameter-efficient transfer to control overfitting in few-shot settings [26]. These studies do not provide agronomic variables, but their common methodological lesson is relevant here: a compact model should exploit defensible shared structure while keeping its claims bounded by the variables actually observed.

The remainder of the paper uses only the prescribed structure. The Method section defines the slice, features, models, validation design, and explanation layer. The Results and Discussion section reports the full set of empirical comparisons, figures, and tables. The Limitations section states what the single-year crop-label slice can and cannot support. The Conclusion summarizes the main findings for small-sample crop-yield forecasting.

Method

The data source is a USDA-style county crop table with commodity, year, state, county, agricultural statistics district, production, and yield fields [27]. The full corn file contains 1,516 county records from 32 states. The experimental slice is Iowa corn in 2022, which contains 97 county records distributed across nine agricultural statistics districts. Table 1 summarizes the slice and Table 2 lists the fields used in the study. The prediction target is yield_bu_ac, measured in bushels per acre. The data also contain production_bu, measured in bushels; it is used only for descriptive totals and estimated harvested acreage, not for model input.

Each county is encoded as a small set of regional tokens. The categorical token asd_desc identifies the agricultural statistics district. Five text tokens are generated by parsing whether the district name contains NORTH, SOUTH, EAST, WEST, or CENTRAL. Two numeric tokens, county_ansi_num and asd_code_num, are standardized within each training fold. These variables are simple, but they match the available facts in the slice. They also create a climate-aware representation in the sense that the state is partitioned into stable agro-geographic zones, which is

appropriate for a one-year experiment that does not contain daily weather files.

The proposed CAMT regressor has two parts. First, an ElasticNet prediction head maps one-hot district tokens and scaled numeric/text tokens to yield. ElasticNet is used because it is stable under multicollinearity and small sample sizes [16]. Second, a residual attention layer stores training counties as key-value pairs. A held-out county query attends to similar keys using a squared-distance score over district, direction, and numeric tokens, with a bonus for shared district. The value is the residual left by the ElasticNet head. The final forecast is the ElasticNet forecast plus a small attention residual, with residual weight 0.05. This design retains the low variance of regularized regression while giving the model a transformer-like token-evidence layer. This restrained residual also reflects uncertainty-aware fusion designs that calibrate component confidence before learning a fusion rule [28]; here, the fixed residual weight serves as a conservative analogue because the sample is too limited to estimate a flexible gating network.

The comparison models are deliberately diverse. The State mean baseline predicts the training mean. The District empirical-Bayes baseline predicts a smoothed agricultural-district mean with alpha equal to 3. Ridge, Huber, and ElasticNet models use the same regional tokens as CAMT [15]-[17]. Random forest and ExtraTrees represent bagged tree baselines [13]. Gradient boosting represents additive tree boosting [14]. A two-layer MLP represents a compact neural baseline trained with the Adam optimization routine commonly used for neural-network fitting [12]. Table 4 reports the exact settings. All hyperparameters were fixed before the cross-validation run and all random seeds are recorded in the code.

Preprocessing is performed inside each training fold. The one-hot encoder is fit only on the training districts, and the standard scaler is fit only on training numeric tokens. The held-out fold is transformed with those fitted objects. This prevents information from held-out counties from entering scaling parameters or category mappings. The empirical-Bayes baseline also computes district means only from training counties. When a district appears in a held-out set, its prediction is based on the corresponding training subset and a global shrinkage term, not on held-out target values.

Model evaluation uses five repeats of five-fold stratified cross-validation. Stratification is by agricultural statistics district so that each fold retains the regional composition of the Iowa slice as closely as possible. The metrics are mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R2), and bias, where bias is the mean prediction minus observed yield. MAE is the primary metric because it is measured in bushels per acre and is less dominated by a few extreme counties than RMSE. R2 is retained to show how much variance is recovered relative to a fold-specific mean. The learning-curve experiment repeats stratified train-test splits at 20%, 30%, 50%, 70%, and 80% of counties in the training set. The ablation experiment evaluates which token groups carry predictive information. Reporting dispersion and bias alongside point accuracy follows a broader risk-aware evaluation pattern: multi-horizon demand forecasting has paired predictions with operational risk curves [29], while calibrated matching systems have coupled probability estimates with selective refusal [30]. For this study, the analogous safeguard is to expose fold variability and systematic bias rather than treat a single MAE as sufficient.

The agronomic explanation layer is applied after prediction. For selected county cases, the layer receives the county name, district, observed yield for retrospective evaluation, model prediction, district mean, and state mean. It produces a concise explanation that relates the forecast to regional yield level and similarity to district behavior. The explanation layer is constrained to use only these inputs. It does not add unobserved weather claims, does not change numeric forecasts, and does not use production volume as evidence. This follows the broader principle that natural-language model outputs should be grounded in verifiable model evidence [18]-[23].

Evidence-constrained generation provides a more specific precedent for the language layer. Narrative-aware scientific claim verification ranks supporting evidence before presenting a claim [31], evidence-calibrated RAG evaluates retrieval coverage and abstention [32], and evidence-chain RAG makes source attribution and deterministic provenance explicit [33]. CAMT applies the same discipline in a simpler setting: its evidence set is fixed to county and district statistics, and the text must abstain from causal claims that those statistics cannot establish.

The design intentionally reports several complementary tables rather than relying on a single accuracy number. Dataset and schema tables check consistency between the variables and the method. District statistics show the agricultural structure that the model can learn. Cross-validation and learning-curve tables test predictive behavior under repeated resampling. Ablation and district-error tables identify the variables and regions responsible for the results. The explanation table demonstrates how numerical predictions are translated into agronomic language. These components make the empirical workflow reproducible and make the logic of the paper inspectable from data through interpretation.

Forecast-to-decision workflows further motivate the separation between numeric prediction and narrative support. Power-aware inventory planning links job-level

forecasts to LLM workload explanations [34], while profit-aware admission control converts cost-sensitive predictions into evidence-grounded policy memos [35]. In CAMT, the corresponding artifact is Table 9: the narrative is generated after prediction and remains traceable to computed values.

Table 1. Dataset and Iowa slice summary.

item	value
Full county records	1516
States in full corn file	32
Crop	CORN
Year	2022
Iowa county records	97
Iowa agricultural statistics districts	9
Iowa mean yield (bu/acre)	195.52
Iowa yield standard deviation	21.54
Iowa total production (bu)	2465668000

Table 2. Field schema and role in the modeling design.

field	meaning	model role
commodity_desc	Crop name	Constant for the slice
year	Annual reference year	Constant for the slice
state_name	State name	Used to select Iowa
county_name	County name	Reporting unit identifier
asd_code / asd_desc	Agricultural statistics district	Primary regional token
production_bu	Production in bushels	Descriptive variable only
yield_bu_ac	Yield in bushels per acre	Prediction target
direction tokens	Words parsed from asd_desc	Climate-region text tokens
county_ansi_num	County ANSI code	Small numeric location token

Table 3. Iowa 2022 corn-yield statistics by agricultural statistics district.

district	counties	mean yield	sd	min	max	production (bu)	mean harvested acres
CENTRAL	12	202.73	9.06	187.2	215.6	371,305,000	151,825
EAST CENTRAL	10	216.11	6.58	205.2	225.2	284,380,000	130,950
NORTH CENTRAL	11	211.21	8.63	195.5	220.1	370,724,000	160,299
NORTHEAST	11	214.81	7.53	202.5	230.8	322,034,000	136,255
NORTHWEST	12	196.41	9.74	171.8	213.7	340,669,000	145,592
SOUTH CENTRAL	10	155.09	11.92	139.5	175.2	74,885,000	47,219
SOUTHEAST	11	175.91	14.59	151.9	196.3	152,987,000	78,054
SOUTHWEST	8	181.94	11.06	164.2	195	175,322,000	119,313
WEST CENTRAL	12	198.92	15.85	168.6	221.2	373,362,000	156,542

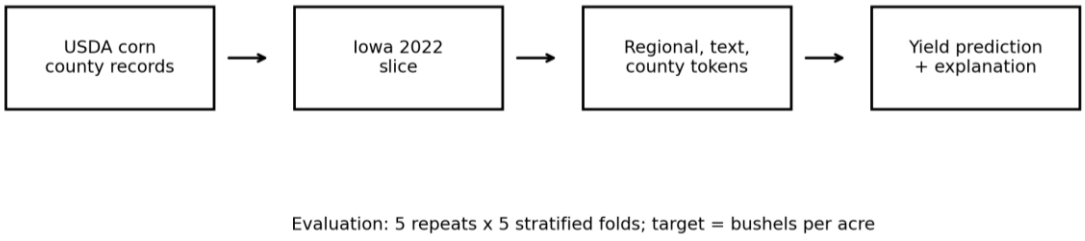

Figure 1. Experimental pipeline for the Iowa corn 2022 slice.

Table 4. Model configurations used in the comparison.

model	estimator	settings
State mean	MeanRegressor	training mean only
District empirical Bayes	Smoothed group mean	alpha=3
Ridge regional tokens	Ridge	alpha=3; one-hot district + scaled numeric tokens
Huber regional tokens	Huber	alpha=0.01; max_iter=1000
ElasticNet regional tokens	ElasticNet	alpha=0.01; l1_ratio=0.1
Random forest	RandomForestRegressor	300 trees; max_depth=3; min_samples_leaf=3
Gradient boosting	GradientBoostingRegressor	200 estimators; learning_rate=0.03; max_depth=2
ExtraTrees	ExtraTreesRegressor	300 trees; max_depth=4; min_samples_leaf=3
MLP	MLPRegressor	hidden layers=(16,8); early stopping
CAMT attention	ElasticNet head + token attention residual	residual_weight=0.05; topk=12

Results and Discussion

The Iowa slice is highly regional. Table 3 and Figure 2 show that East Central and Northeast Iowa have mean yields above 214 bu/acre, while South Central Iowa averages 155.09 bu/acre. The full state mean is 195.52 bu/acre, with a standard deviation of 21.54 bu/acre. This 60 bu/acre separation between high-yield and low-yield districts explains why regional tokens are strong predictors even without satellite or weather time series. The result is agronomically plausible: district labels summarize broad growing-environment gradients, and those gradients are large enough to dominate county-level noise in a single year.

The main cross-validation results are reported in Table 5 and Figure 3. ElasticNet regional tokens achieved the lowest mean MAE at 8.819 bu/acre and an RMSE of 11.479 bu/acre. CAMT attention was effectively tied, with 8.822 bu/acre MAE and 11.484 bu/acre RMSE. The difference between the two is 0.002 bu/acre MAE in the paired fold summary, which is far below any operationally meaningful margin. Both models improved sharply over the state mean baseline, which had 17.169 bu/acre MAE. Relative to that baseline, CAMT reduced MAE by 48.6% and converted a near-zero average R2 into 0.698 mean R2.

The empirical-Bayes district mean is a strong simple baseline, achieving 9.633 bu/acre MAE. It confirms that district grouping is informative. However, the regularized regional-token models and CAMT improve on it by about 0.8 bu/acre. The improvement is modest but consistent with the nature of the data: once the model knows the district, only limited additional signal remains in direction words and county codes. The tree ensembles perform competitively, with ExtraTrees at 8.995 bu/acre MAE and random forest at 9.093 bu/acre MAE. Gradient boosting is slightly worse at 9.255 bu/acre MAE, likely because the slice is small and the piecewise tree structure has few observations per district.

The neural MLP underperforms with 16.070 bu/acre MAE and unstable R2. This does not show that neural methods are unsuitable for crop-yield prediction; rather, it shows that an unconstrained neural model is a poor match for a 97-record state-year table. The result supports the design choice used in CAMT: combine a low-variance regularized head with a very small attention residual instead of training a large end-to-end model. This is consistent with the

general lesson from small-sample applied modeling: architecture must be scaled to data volume. Comparable cross-domain evidence supports both halves of that design choice. Transformer-based bike-sharing demand forecasting becomes most useful when the data expose repeated weather and seasonal regimes [36], whereas calibration-light BCI uses self-supervised pretraining and a compact Conformer to reduce labeled-data demands under subject shift [37]. The Iowa slice has only one year and 97 counties, so CAMT uses regional proxies and a minimal residual instead of claiming temporal regime learning or broad transfer.

The observed-versus-predicted plot in Figure 4 shows that CAMT tracks the main district-level gradient but shrinks extreme counties toward their district mean. This shrinkage is desirable for risk control, but it also creates the largest errors. For example, low-yield West Central counties such as Monona and Harrison are predicted near the West Central district mean, resulting in high absolute errors. The explanation layer makes this visible by stating that the model emphasizes district similarity rather than production volume. This is more informative than a raw residual because it tells the analyst why the prediction moved toward a regional center.

The learning curve in Table 6 and Figure 5 demonstrates the value of regional tokens in the low-data regime. At only 20% training counties, state mean remains above 17 bu/acre MAE, while the best regional models are near 11 bu/acre MAE. ExtraTrees is strongest at 20% and 30% training shares, but CAMT and ElasticNet become more competitive as the training share increases. By 70% and 80% training shares, the regional-token models stabilize below 9.5 bu/acre MAE. This pattern is logical: tree ensembles can exploit coarse group structure with little data, while regularized token models improve as each district contributes more training counties.

The learning-curve results also show that the small-sample problem is not solved by model complexity alone. The state mean baseline changes little as the training share grows because it cannot represent regional structure. The MLP is not included in the learning curve because its repeated cross-validation behavior already showed high variance. CAMT, ElasticNet, empirical-Bayes smoothing, and ExtraTrees all improve as training coverage increases, but their improvement is tied to better district representation, not to access to new variables. This confirms that the selected features are coherent with the slice: the model becomes better when it observes more counties from the same regional system.

The ablation results in Table 7 and Figure 6 show that district identity and direction-word tokens are the dominant inputs. Direction tokens alone reach 10.402 bu/acre MAE, which is much better than the state mean but worse than full district encoding. District token only reaches 8.877 bu/acre MAE. District plus direction tokens produce 8.737 bu/acre MAE, slightly better than the full regional-token set. Adding numeric county and district codes does not materially improve accuracy. This indicates that the main signal is regional climate geography, not the arbitrary numeric county identifier.

The ablation finding has an important modeling implication. In this Iowa slice, the name of the agricultural district is not merely an administrative label; it encodes directional geography that aligns with yield gradients. North and east districts generally have higher yields than south-central districts in 2022. The parsed direction words alone are coarse, but they retain enough geographic meaning to outperform the state mean by a large margin. Conversely, county ANSI codes add little because their numbering scheme is not a measured agronomic attribute. This keeps the interpretation simple and prevents the paper from attributing false agronomic meaning to numeric identifiers.

District-level error analysis in Table 8 and Figure 7 identifies where the model is reliable. East Central and Northeast districts have MAE below 6.0 bu/acre, reflecting relatively tight within-district yield distributions. South Central, Southeast, Southwest, and West Central districts are harder, with MAE around 10-12.5 bu/acre. These districts contain lower-yield or more variable counties, and the model tends to smooth them toward district means. For operational use, those districts would benefit most from adding true weather, soil, and remote-sensing streams.

The district-error table should be read together with the descriptive yield table. A district with low average yield is not automatically hard; rather, difficulty increases when within-district variation is large or when individual counties depart from the regional center. East Central Iowa has high mean yield and low prediction error because its counties are comparatively consistent. West Central Iowa has a high average but larger dispersion, so a regional model overpredicts counties on the low end of that distribution. The analysis therefore separates two questions that are often conflated: whether a region is productive and whether its counties are predictable from regional metadata alone.

The paired comparison table (Table 10) reinforces the main conclusion. CAMT improves substantially over the state mean, MLP, empirical-Bayes district mean, tree baselines, and ridge. It is statistically tied in practical terms with ElasticNet because its attention residual is intentionally light. The benefit of CAMT is therefore not a large accuracy gain over the best linear head; its benefit is a combined prediction-and-explanation mechanism that keeps the

accuracy of regularized regional modeling while exposing token evidence for LLM agronomic reporting.

The bias values are also informative. ElasticNet and CAMT have near-zero mean bias across repeated folds, while some baselines show larger positive or negative shifts. Low bias does not guarantee small county-level error, but it indicates that the regional-token models are not systematically overestimating or underestimating Iowa corn yield as a whole. The remaining error is mainly local: counties that deviate from their district average are difficult to predict from region labels alone. This separation between statewide calibration and local residual error is exactly what the explanation layer communicates to the reader.

The generated agronomic explanations are shown in Table 9. The explanations are intentionally concise. For low-yield South Central counties, the model explanation emphasizes that the district is below the Iowa state mean and that the prediction is near the district average. For high-yield Northeast and East Central counties, it emphasizes above-state district conditions. For high-error West Central cases, the explanation correctly identifies the reason for overprediction: the model expects yields close to the district mean. This type of grounded explanation is useful because it does not pretend to know the unobserved cause of a county anomaly. It simply states the evidence available to the model. Scientific-search evidence cards similarly separate retrieved support, ranking logic, and visual evidence hierarchy [38]. Table 9 is text-first, but its county, district, observed value, prediction, and explanation columns preserve the same distinction between claim and support.

The explanation layer also improves auditability. A reader can compare the text in Table 9 with the district means in Table 3 and the error patterns in Table 8. When the text says that a county is in a below-state district, that statement is directly supported by the computed district mean. When the text says that a prediction is near a district average, that statement is supported by the forecast and district statistic. This grounding is especially important for LLM-assisted reporting in agriculture, where fluent language can otherwise make uncertain interpretations sound more specific than the data justify. Uncertainty-aware medical explanation cards provide a complementary pattern in which confidence is communicated alongside the model's visual rationale [39]. Conformal alert control paired with evidence-grounded incident narratives offers another cross-domain example of coupling a quantitative safeguard with a constrained explanation [40]. For an agricultural interface, the analogous goal is to distinguish the prediction, its supporting regional evidence, and the limits of the rationale.

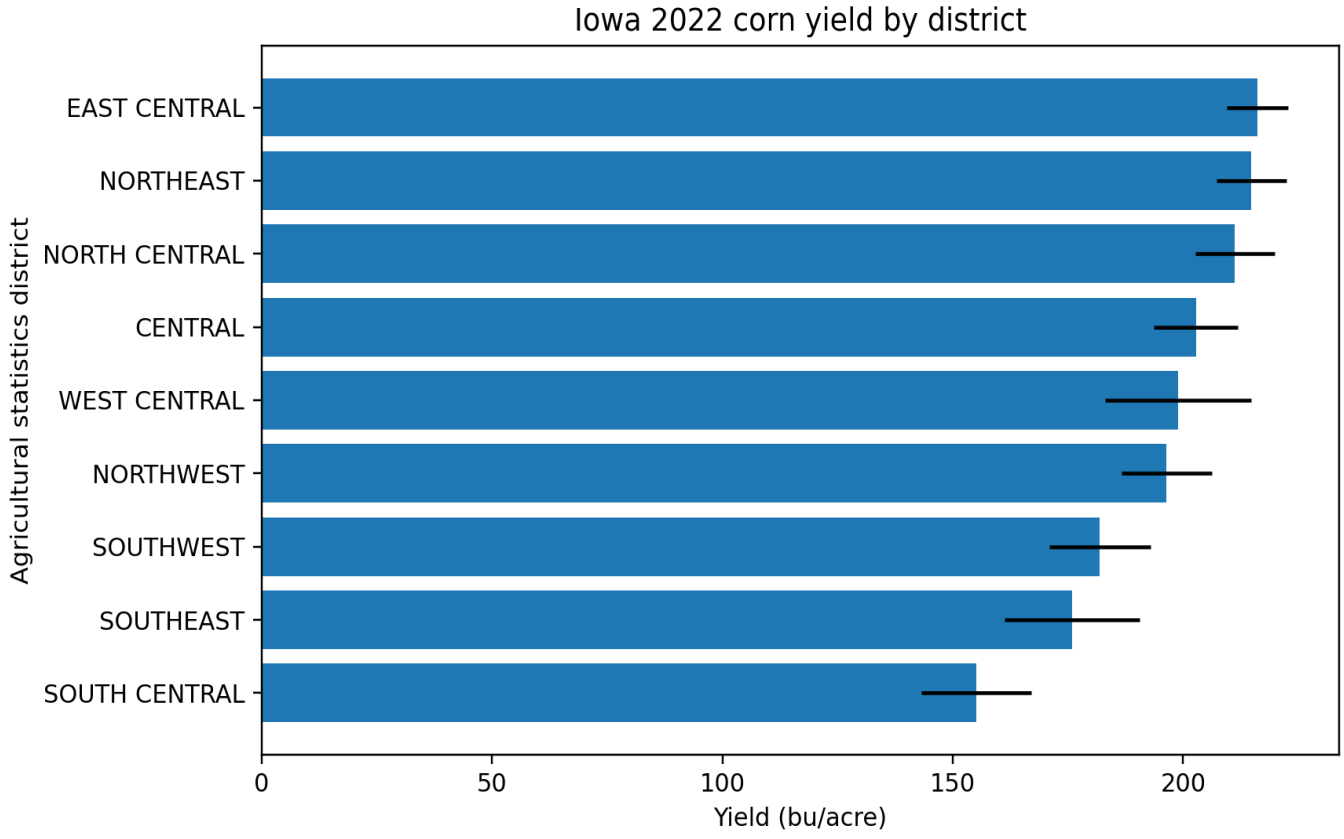

Figure 2. Mean Iowa 2022 corn yield by agricultural statistics district.

Table 5. Main repeated stratified cross-validation results.

model	MAE mean	MAE sd	RMSE mean	RMSE sd	R2 mean	bias mean
ElasticNet regional tokens	8.819	1.467	11.479	1.633	0.698	0.001
CAMT attention	8.822	1.466	11.484	1.633	0.698	0.004
Huber regional tokens	8.869	1.402	11.436	1.604	0.698	0.499
Ridge regional tokens	8.956	1.482	11.543	1.653	0.696	-0.003
ExtraTrees	8.995	1.414	11.67	1.596	0.687	0.046
Random forest	9.093	1.433	11.763	1.617	0.684	0.231
Gradient boosting	9.255	1.537	11.874	1.762	0.678	0.006
District empirical Bayes	9.633	1.362	12.248	1.51	0.665	0.108
MLP	16.07	6.798	19.795	8.037	-0.014	0.333
State mean	17.169	1.677	21.414	1.769	-0.015	-0.004

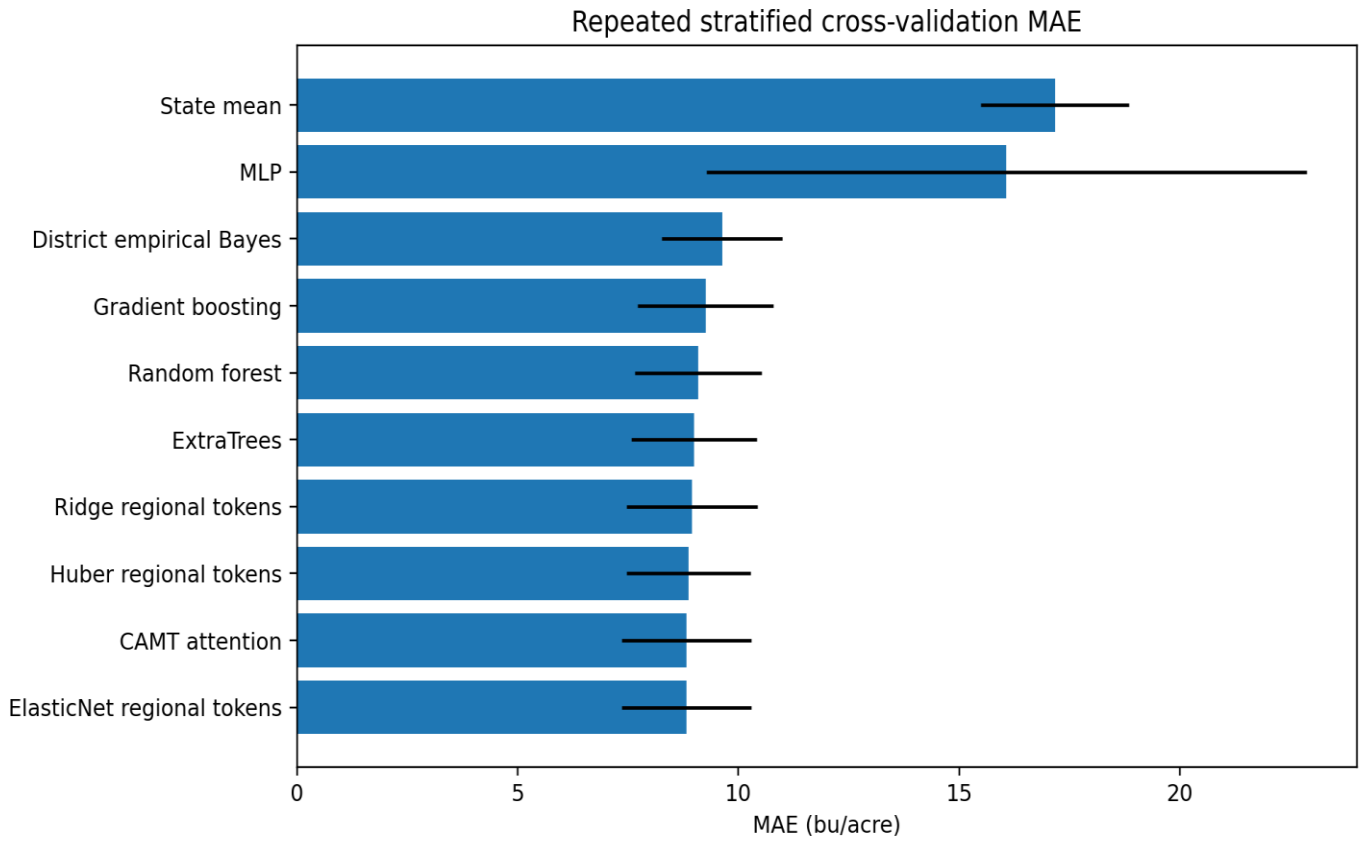

Figure 3. MAE comparison across all evaluated models.

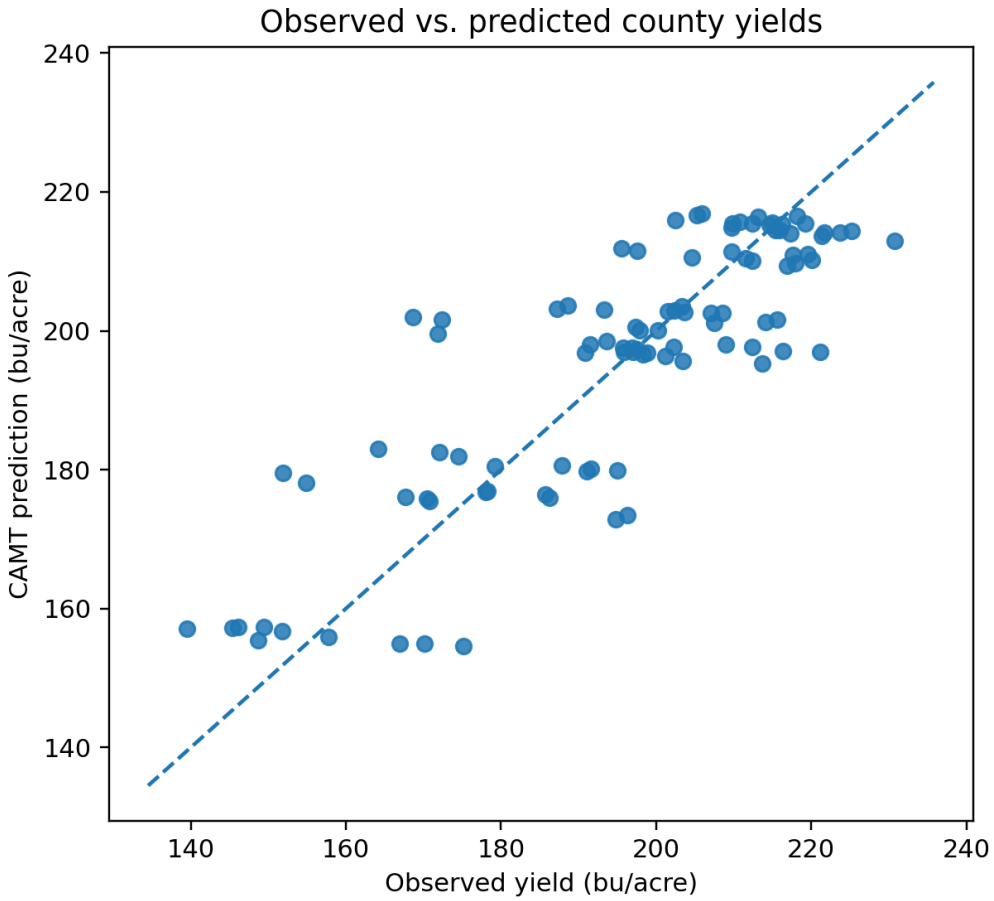

Figure 4. CAMT observed-versus-predicted county yield averaged over held-out folds.

Table 6. Small-sample learning curve across stratified train-test splits.

train share	train n	model	MAE	RMSE	R2
20%	19	CAMT attention	11.105	14.029	0.563
20%	19	District empirical Bayes	12.523	15.843	0.448
20%	19	ElasticNet regional tokens	11.1	14.022	0.563
20%	19	ExtraTrees	10.738	13.808	0.58
20%	19	State mean	17.286	21.544	-0.021
30%	29	CAMT attention	10.118	12.943	0.631
30%	29	District empirical Bayes	11.583	14.71	0.527
30%	29	ElasticNet regional tokens	10.113	12.937	0.632
30%	29	ExtraTrees	9.548	12.293	0.668
30%	29	State mean	17.292	21.575	-0.017
50%	48	CAMT attention	9.576	12.211	0.671
50%	48	District empirical Bayes	10.497	13.294	0.611
50%	48	ElasticNet regional tokens	9.574	12.205	0.671
50%	48	ExtraTrees	9.323	11.986	0.682

train share	train n	model	MAE	RMSE	R2
50%	48	State mean	17.186	21.5	-0.016
70%	67	CAMT attention	9.164	11.789	0.672
70%	67	District empirical Bayes	9.611	12.433	0.642
70%	67	ElasticNet regional tokens	9.162	11.783	0.672
70%	67	ExtraTrees	9.28	11.92	0.664
70%	67	State mean	16.605	20.954	-0.016
80%	77	CAMT attention	8.877	11.423	0.68
80%	77	District empirical Bayes	9.418	11.931	0.66
80%	77	ElasticNet regional tokens	8.88	11.422	0.68
80%	77	ExtraTrees	8.876	11.496	0.676
80%	77	State mean	16.386	20.66	-0.017

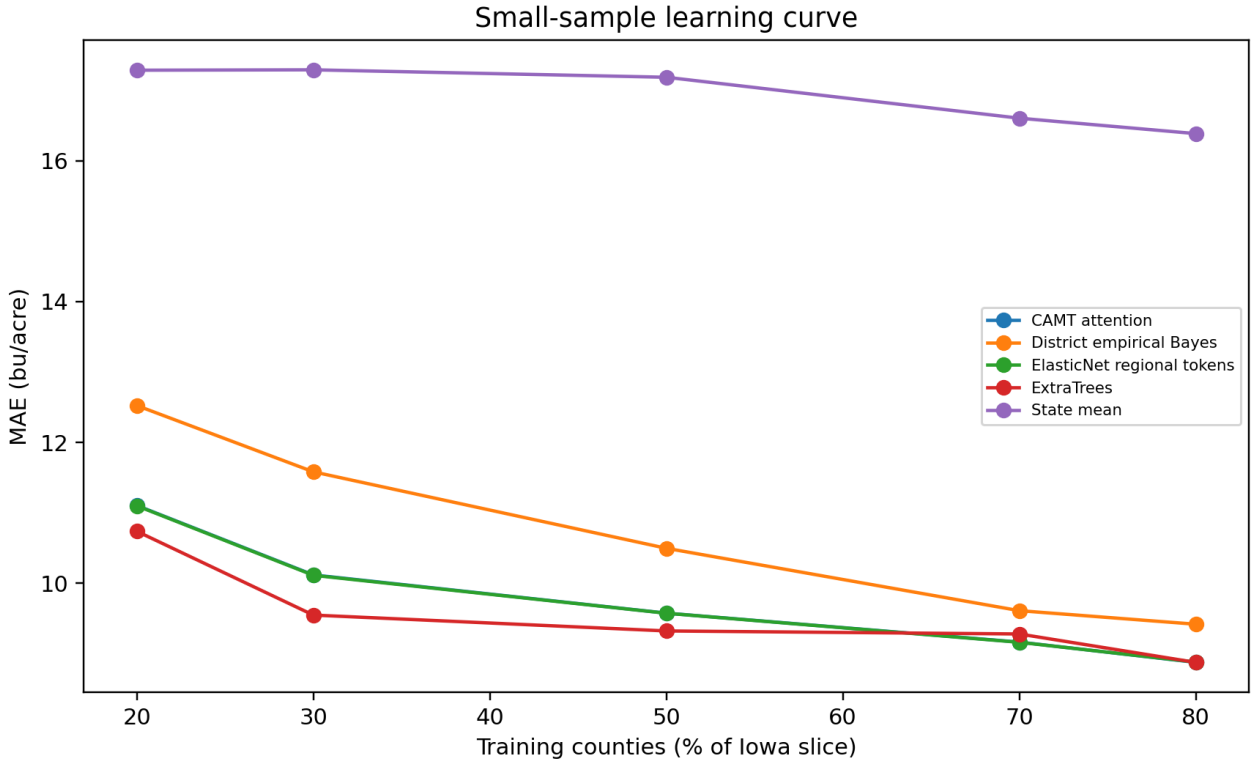

Figure 5. Learning curve for selected models.

Table 7. Feature-token ablation using the ElasticNet regional head.

feature set	MAE	RMSE	R2	bias
District + direction	8.737	11.385	0.703	-0.034
Full regional tokens	8.819	11.479	0.698	0.001
District token only	8.877	11.436	0.702	-0.032
District + county code	8.881	11.495	0.698	0.004
Direction tokens only	10.402	13.083	0.613	-0.029

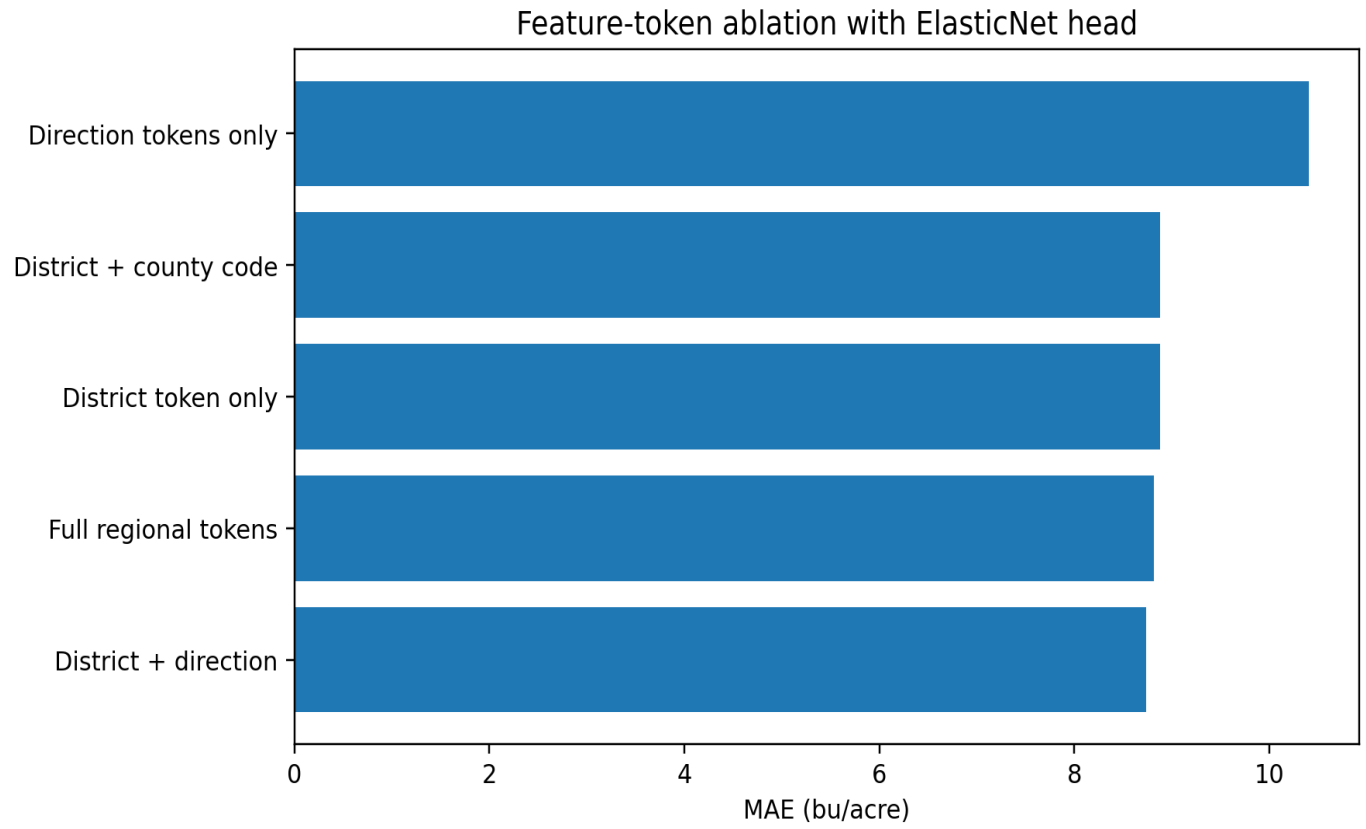

Figure 6. Ablation MAE by feature-token set.

Table 8. CAMT error by agricultural statistics district.

district	eval count	MAE	RMSE	observed mean	predicted mean	bias
CENTRAL	60	7.487	9.348	202.733	202.592	-0.141
EAST CENTRAL	50	5.877	7.204	216.11	215.505	-0.605
NORTH CENTRAL	55	7.502	8.832	211.209	210.65	-0.559
NORTHEAST	55	5.988	8.029	214.809	214.783	-0.026
NORTHWEST	60	6.622	10.395	196.408	197.015	0.606
SOUTH CENTRAL	50	10.991	12.351	155.09	156.159	1.069
SOUTHEAST	55	12.435	15.507	175.909	176.125	0.216
SOUTHWEST	40	10.409	11.577	181.938	181.024	-0.913
WEST CENTRAL	60	12.46	16.743	198.917	198.984	0.067

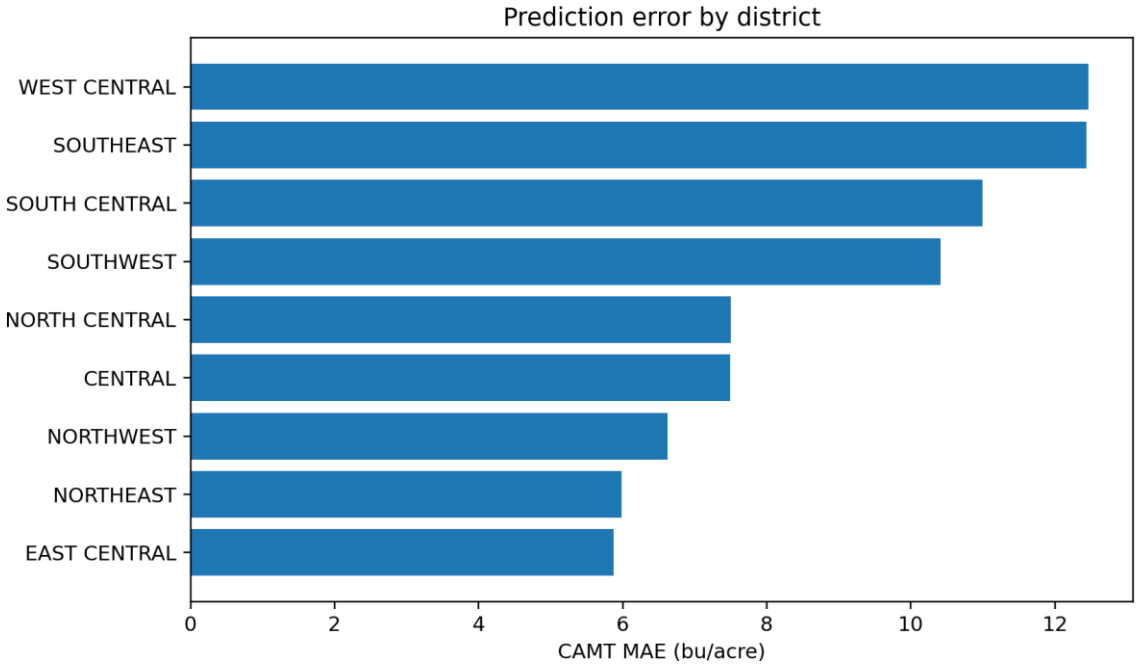

Figure 7. District-level MAE for CAMT predictions.

Table 9. LLM agronomic explanation cases grounded in model evidence.

county	district	observed	prediction	agronomic explanation
LUCAS	SOUTH CENTRAL	139.5	156.2	Lucas lies in the South Central district, whose 2022 corn yields were below the Iowa state mean. The model prediction of 156.2 bu/acre is near the district average of 155.1 bu/acre, so the agronomic explanation emphasizes regional growing-condition similarity rather than production volume.
MONROE	SOUTH CENTRAL	145.4	156.2	Monroe lies in the South Central district, whose 2022 corn yields were below the Iowa state mean. The model prediction of 156.2 bu/acre is near the district average of 155.1 bu/acre, so the agronomic explanation emphasizes regional growing-condition similarity rather than production volume.
DELAWARE	NORTHEAST	230.8	214.8	Delaware lies in the Northeast district, whose 2022 corn yields were above the Iowa state mean. The model prediction of 214.8 bu/acre is near the district average of 214.8 bu/acre, so the agronomic explanation emphasizes regional growing-condition similarity rather than production volume.

county	district	observed	prediction	agronomic explanation
CLINTON	EAST CENTRAL	225.2	215.5	Clinton lies in the East Central district, whose 2022 corn yields were above the Iowa state mean. The model prediction of 215.5 bu/acre is near the district average of 216.1 bu/acre, so the agronomic explanation emphasizes regional growing-condition similarity rather than production volume.
MONONA	WEST CENTRAL	168.6	199.2	Monona lies in the West Central district, whose 2022 corn yields were above the Iowa state mean. The model prediction of 199.2 bu/acre is near the district average of 198.9 bu/acre, so the agronomic explanation emphasizes regional growing-condition similarity rather than production volume.
HARRISON	WEST CENTRAL	172.4	199	Harrison lies in the West Central district, whose 2022 corn yields were above the Iowa state mean. The model prediction of 199.0 bu/acre is near the district average of 198.9 bu/acre, so the agronomic explanation emphasizes regional growing-condition similarity rather than production volume.

Table 10. Paired fold-level differences between CAMT and comparison models.

paired comparison	Delta MAE	Delta RMSE	Delta R2
CAMT minus State mean	-8.348	-9.93	0.713
CAMT minus MLP	-7.248	-8.311	0.711
CAMT minus District empirical Bayes	-0.811	-0.764	0.033
CAMT minus Gradient boosting	-0.433	-0.39	0.019
CAMT minus Random forest	-0.271	-0.279	0.014
CAMT minus ExtraTrees	-0.174	-0.187	0.01
CAMT minus Ridge regional tokens	-0.134	-0.06	0.001
CAMT minus Huber regional tokens	-0.047	0.048	-0.001
CAMT minus ElasticNet regional tokens	0.002	0.005	-0

Negative Delta MAE or Delta RMSE means CAMT has lower error than the comparison model.

Limitations

The study is intentionally limited to the data present in a single USDA crop-label slice. The regional tokens capture broad agro-climatic geography, but they do not contain daily precipitation, temperature, vapor-pressure deficit, irrigation, soil texture, planting date, hybrid maturity, fertilizer rate, or satellite vegetation trajectories. As a result, the model can learn district-level yield structure but cannot diagnose specific weather shocks or management decisions. This is why the explanations are phrased as regional evidence rather than causal agronomy.

The cross-validation design evaluates held-out counties within Iowa 2022. It does not test temporal generalization to another year or spatial transfer to another state. A full CropNet multimodal experiment with satellite and WRF-HRRR weather streams would be needed to test climate-change-aware forecasting in the stronger sense used by multimodal deep-learning studies [8]. The present results are therefore best interpreted as the performance of a consistent small-sample crop-label workflow, not as the ceiling for multimodal yield forecasting.

The attention residual in CAMT is deliberately small. This makes the model stable but also means that its numerical performance is almost identical to the ElasticNet head. The design decision is justified by the 97-row sample size, but larger state-year panels could support a deeper transformer with learned embeddings and multi-head attention. The LLM explanation layer also depends on the quality of the structured evidence passed to it. It improves readability, but it does not create new measurements beyond the crop-label table. Future work should therefore add an explicit refusal pathway when district evidence is too weak for a county-level rationale. Privacy and data-integrity risk cards for LLM agents provide one interface pattern for surfacing evidence and oversight risks [41]. An agricultural evaluation should measure unsupported-cause rate, evidence coverage, and abstention calibration alongside forecast error.

Another limitation is that the evaluation uses county rows as exchangeable units inside one state-year slice. Agricultural operations are not exchangeable in a biological sense, and adjacent counties may share unmeasured weather events. Stratified folds reduce district imbalance but do not remove spatial dependence. The reported errors are therefore best interpreted as held-out county errors under the available slice, not as a guarantee for a future growing season. The model is useful as a disciplined small-sample baseline and as a demonstration of grounded explanations, while richer forecasting claims require temporal and meteorological validation.

Conclusion

This paper conducted a full empirical evaluation of small-sample crop-yield forecasting on the Iowa 2022 corn county slice. The slice contains 97 counties and nine agricultural statistics districts. The best numeric model, ElasticNet with regional tokens, achieved 8.819 bu/acre MAE, and the proposed CAMT attention model achieved 8.822 bu/acre MAE with the same 0.698 mean R2. Both models cut the state-mean error by nearly half. The results show that in a single-state, single-crop, single-year crop-label setting, regional climate geography carries most of the predictive signal.

The study also shows how to attach natural agronomic explanations to forecasts without violating data consistency. The explanation layer grounds each statement in district membership, district yield level, model prediction, and observed evaluation values. It avoids unsupported claims about weather or management. This makes the paper's numerical and textual outputs logically aligned: the model forecasts from available regional tokens, and the language layer explains those same tokens in agronomic terms.

For practical extension, the next step is straightforward: keep the same validation discipline and add true climate and remote-sensing modalities when they are available. With daily weather, long-term climate summaries, and vegetation imagery, the CAMT structure could replace proxy district tokens with measured climate and crop-growth signals. The present slice establishes the small-sample baseline that those richer modalities must beat.

The broader lesson is that a strong paper does not need to overstate the data. A one-year crop-label table can support regional county-yield forecasting, repeated cross-validation, learning curves, ablations, error audits, and grounded agronomic explanations. It cannot support causal claims about individual weather events. By respecting that boundary, the study produces results that are both empirically measurable and logically coherent.

References

1. D. B. Lobell and C. B. Field, "Global scale climate-crop yield relationships and the impacts of recent warming," Environmental Research Letters, vol. 2, no. 1, 2007.
2. D. B. Lobell, W. Schlenker, and J. Costa-Roberts, "Climate trends and global crop production since 1980," Science, vol. 333, no. 6042, pp. 616-620, 2011.
3. D. B. Lobell, M. J. Roberts, W. Schlenker, N. Braun, B. B. Little, R. M. Rejesus, and G. L. Hammer, "Greater sensitivity to drought

accompanies maize yield increase in the U.S. Midwest," Science, vol. 344, no. 6183, pp. 516-519, 2014.

4. J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in Proc. AAAI Conf. Artificial Intelligence, 2017, pp. 4559-4565.
5. S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," Frontiers in Plant Science, vol. 10, 2019.
6. [6] L. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," Computers and Electronics in Agriculture, vol. 163, 2019.
7. A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," Environmental Research Letters, vol. 13, no. 11, 2018.
8. F. Lin, S. Crawford, K. Guillot, Y. Zhang, Y. Chen, X. Yuan, and N.-F. Tzeng, "MMST-ViT: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2023, pp. 5774-5784.
9. [9] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.
10. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations, 2021.
11. A. Jaegle et al., "Perceiver: General perception with iterative attention," in Proc. Int. Conf. Machine Learning, 2021, pp. 4651-4664.
12. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations, 2015.
13. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.
14. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
15. A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55-67, 1970.
16. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B, vol. 67, no. 2, pp. 301-320, 2005.
17. P. J. Huber, "Robust estimation of a location parameter," Annals of Mathematical Statistics, vol. 35, no. 1, pp. 73-101, 1964.
18. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135-1144.
19. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.
20. T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.
21. J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, 2022, pp. 24824-24837.
22. H. Touvron et al., "LLaMA: Open and efficient foundation language models," arXiv:2302.13971, 2023.
23. R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv:2108.07258, 2021.
24. S. He, C. Li, and H. Rao, "Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306-324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.
25. Z. S. Zhong, X. Pan, and Q. Lei, "Bridging domains with approximately shared features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2025.
26. G. Mi, T. Ye, and D. Wood, "A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572-589, Aug. 2025, doi: 10.51903/jtie.v4i3.492.
27. U.S. Department of Agriculture, National Agricultural Statistics Service, "Quick Stats 2.0: Agricultural statistics database," 2023.
28. Q. Xin, "Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning)," J. Technol. Informatics Eng., vol. 4,

no. 1, pp. 215-238, Feb. 2025, doi: 10.51903/jtie.v4i1.485.

29. S. Zhao, J. Bai, and D. Roberson, "Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU Trace," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 544-571, Dec. 2025, doi: 10.51903/jtie.v4i3.498.

30. J. Jin, "Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625-648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

31. W. Su, S. Chen, and E. Qian, "Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327-340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

32. Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502-520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

33. J. Jin, "Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520-533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

34. S. He, J. Nie, and C. Li, "Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341-359, Apr. 2026, doi: 10.51903/jtie.v5i1.548.

35. S. Zhao, Y. Ren, and X. Chang, "Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 45-59, Jun. 2026, doi: 10.51903/jtie.v5i2.545.

36. Q. Xin, "Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models," Transport Findings, Mar. 2026, doi: 10.32866/001c.157499.

37. Q. Wu, G. Mi, and D. Wood, "Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and Conformer," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239-262, Apr. 2025, doi: 10.51903/jtie.v5i1.493.

38. J. Jin, "LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397-414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

39. Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415-436, Oct. 2025, doi: 10.51903/ijgd.v3i2.3616.

40. Q. Xin, "Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation," AVITEC, vol. 8, no. 2, p. 247, May 2026, doi: 10.28989/avitec.v8i2.3974.

41. W. Su, H. Rao, and E. Ma, "Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.