Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can General-Purpose LLMs Predict 24-Hour Drug Release from PLGA Microparticles? An Exploratory Benchmark Against Conventional Models

Seyed Masih Abtahi Naeini

masih.abtahi@srbiau.ac.ir

Independent Researcher

Research Article

Keywords: PLGA microparticles, drug-release prediction, general-purpose large language models, in-context learning, regression benchmark, target interpolation, reproducibility

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11590653/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose

To evaluate whether general-purpose large language models (LLMs), prompted in zero-shot, few-shot, and full-training-context settings, can predict 24-h cumulative in-vitro release fractions for drug-loaded PLGA microparticles relative to conventional regressors.

Methods

Nine methods were compared on 67 retained cases using a literature-derived endpoint. Three targets were direct 24-h measurements and 64 were linearly interpolated; T045 used a figure-supported midpoint with two sensitivity alternatives. The original DOI-group split manifest and case-to-source map were unavailable, and a candidate source-group overlap remains unresolved; DOI-independent separation therefore cannot be confirmed.

Results

The full-training-context LLM yielded MAE 12.269 percentage points (pp), RMSE 15.770 pp, and R^2 0.217; Random Forest yielded 12.259 pp, 15.973 pp, and 0.196, respectively. Archived paired intervals for metric differences included zero, but because the source mapping and bootstrap records were not recovered, they are conditional summaries.

Conclusions

The retained results show near-tied performance, not a decisive winner. Because 64/67 targets were interpolated and only three were direct measurements, results represent agreement with a dataset-derived endpoint, not validation against direct 24-h observations. Missing model/version receipts and incomplete split provenance preclude complete reproducibility.

Summary

Summary (submission metadata; approximately 100 words): This exploratory benchmark compares three prompting conditions for general-purpose large language models with conventional regressors in predicting 24-h cumulative in-vitro release fractions from PLGA microparticles. On 67 retained cases, the full-training-context LLM and Random Forest showed nearly identical MAE. However, 64 targets were interpolated, only three were direct 24-h measurements, and one candidate source-group overlap remains unresolved. The original case-to-source map, split manifest, bootstrap execution records, and provider-native model metadata were not recovered. Results therefore describe agreement with a dataset-derived endpoint and do not establish DOI-independent generalization or LLM superiority.

Teaser (25–30 words): An exploratory PLGA benchmark compares general-purpose LLM prompting with conventional regression while documenting interpolated endpoints, unresolved source-group provenance, and limits on reproducibility and superiority claims.

1. Introduction

Poly(lactic-co-glycolic acid) (PLGA) microparticles are widely investigated as long-acting drug-delivery systems because their polymer matrix can support sustained in-vitro release. Release behaviour depends on interacting drug, polymer, and formulation properties, and the interpretation of release profiles is complicated by heterogeneous experimental protocols and incomplete reporting across primary studies. In this setting, curated datasets can help organize the available formulation space and support data-driven hypothesis generation, but predictions remain conditional on the quality, comparability, and provenance of the underlying measurements (Grizić and Lamprecht [1]; Bao et al. [2]).

Bao et al. [2] assembled a literature-derived dataset comprising 321 in-vitro PLGA microparticle release experiments, 4,913 release time-point records, and 89 drugs collected from 113 publications. Separately, the processed project table contains 321 formulation indices; that project-specific representation should not be conflated with the source paper's count of release experiments. This resource makes comparative modelling possible while inheriting limitations of literature-mined data, including sparse endpoint sampling and heterogeneous experimental protocols. Subsequent studies have developed machine-learning approaches for PLGA release prediction using different formulation collections and endpoint definitions. Catapano et al. [3] modelled release profiles from 113 PLGA formulations; Alabsi et al. [4] evaluated a harmonized dataset of 347 formulations and 5,265 release measurements using formulation descriptors and early-release information. Tang et al. [6] subsequently developed MC-NODE using the Bao et al. dataset itself (321 release curves and 4,913 observations), with DOI-grouped five-fold cross-validation to evaluate complete release trajectories. These studies establish that PLGA release modelling with conventional and specialized machine-learning methods is already an active area. Their reported performance measures should not be ranked directly against the present results because the datasets, prediction targets, information available to the models, and validation designs differ.

General-purpose large language models (LLMs) offer a different prediction interface: numeric examples and query cases can be placed in a prompt, allowing a model to produce predictions without fitting a conventional task-specific estimator in the usual way. Yet this flexibility does not guarantee reliable tabular regression. Recent work has documented substantial sensitivity of general-purpose LLM predictions to task-irrelevant changes in variable labels, row/column order, or representation, and has advised caution when LLMs are used as black-box tabular predictors [5]. For formulation science, the practical question is therefore not whether an LLM can emit a plausible release fraction, but whether its predictions are competitive against transparent baselines under a controlled split and a clearly defined target.

This study presents an exploratory comparison of zero-shot, few-shot, and full-training-context LLM prediction conditions with conventional regression baselines for a 24-h cumulative in-vitro release target derived from the public PLGA microparticle dataset. The intended contribution is narrower than proposing a new PLGA-specific predictor: it is to describe how general-purpose LLM prompting conditions compare with conventional regressors on a fixed 24-h endpoint under explicitly documented data and execution constraints. The retained test set contains 67 cases and is associated with 23 nominal DOI labels in the audit mapping. Although archived protocol materials describe an intended DOI-grouped split and an exact ten-feature purge, the original pre-run case-to-formulation/DOI map and split manifest were not recovered; DOI-independent separation therefore cannot be independently authenticated. The analysis reports MAE, RMSE, R^2, bias, target-construction subgroups, and a cluster-bootstrap comparison conditional on the retained nominal labels. Because most 24-h labels are interpolated, one target required a figure-supported reconstruction, and historical LLM execution metadata are incomplete, results are interpreted conservatively as exploratory agreement with a defined dataset-derived endpoint.

2. Methods

Dataset, endpoint, and target construction

The processed PLGA microparticle dataset contained 4,913 rows representing 321 formulation indices, 113 DOI groups after DOI normalization, and 89 drug names after trimming whitespace. The ten numerical predictors were drug molecular weight (Da), topological polar surface area (Å²), drug LogP, reported polymer molecular weight (kDa), LA:GA molar ratio, initial drug-to-polymer ratio (w/w), particle size (μm), drug loading capacity (%), drug encapsulation efficiency (%), and solubility-enhancer concentration (%). Formulation Index and sampling time were not columns in the blinded predictor workbooks. Release sampling time was expressed in days, and the primary endpoint was Target_24h_fraction, the cumulative in-vitro release fraction at 24 h (Time = 1.0 day). When an observed measurement at exactly 1.0 day was available, that value was used directly. Otherwise, the target was linearly interpolated between the nearest observed timepoints that strictly bracketed 1.0 day:

$$\hat{y}_{24} = y_1 + [(1 - t_1)/(t_2 - t_1)] \times (y_2 - y_1), \text{ where } t_1 < 1 < t_2$$

Across the 321 formulation-level endpoints, 7 were direct 24-h measurements and 314 were interpolated targets. In the fixed test set, 3 targets were direct measurements and 64 were interpolated. All model metrics were calculated against these defined dataset-derived targets; consequently, agreement with the endpoint should not be interpreted as validation against directly measured 24-h outcomes for most test cases.

Duplicate timepoints and endpoint freeze

Four formulation/timepoint groups contained repeated timestamps: two conflicting pairs and two exact same-value duplicates. Bao et al. [2] report extracting release points from published profiles using

WebPlotDigitizer. I inspected the source PDF by Grizić and Lamprecht [1]: Fig. 5 (PDF p. 30; printed p. 29) is captioned mean ± SD (n = 3), and the early erythromycin point's visible vertical error bar is consistent with the two same-time release values 0.278226 and 0.403226 at t = 0.236842 days being lower/upper error-bar bounds. Their midpoint (0.340726) is used as the central-curve estimate; interpolation to the next point (t = 1.026316 days, release = 0.584677) gives Target_24h_fraction = 0.576546. The first-record target (0.574462) and second-record target (0.578629) remain sensitivity scenarios. This is figure-supported reconstruction, not direct recovery of the authors' replicate-level raw mean. The other conflicting pair is Guanosine, but its duplicated time is not part of the 24-h bracket; exact duplicate rows for Prilocaine and progesterone also do not alter their 24-h targets. The canonical scoring vector is frozen as midpoint_T045_figure_supported_2026-10-18 (T045 target = 0.576545698924732 fraction; 67-case target SHA-256 = 13216501dba145da6e5f399671eb13b77864e11e83af25271d4cff8d0650f246). The first- and second-record target vectors remain sensitivity scenarios; the same canonical primary target was applied across all nine methods.

Retained holdout and split provenance

Archived protocol and fitting-script materials describe an intended 67-case DOI-grouped holdout, an initial 254-row training set, and removal of one training row with an exact match to a held-out case across all ten predictor features, leaving 253 training formulations. The retained audit mapping associates the 67 cases with 23 nominal DOI labels. A source-group audit identified one test descriptor vector that can map to two candidate source formulations; one candidate DOI group is represented by 18 rows in the 253-row full-training context. Because the original pre-run case-to-formulation/DOI map and frozen split manifest were not recovered, the actual source identity and whether this candidate overlap applies to the test case cannot be resolved. Thus, DOI-independent separation for the full-training condition is not verified, and the results must not be interpreted as a confirmed leakage-free DOI holdout.

Prediction conditions and evaluation

Three prompt-based LLM conditions were evaluated on the same 67 retained inputs: zero-shot (no labelled examples in context), few-shot (12 labelled training examples in context), and full-training-context (253 labelled training examples in context). Conventional comparators were Random Forest, Gradient Boosting, Extra Trees, Ridge regression, and constant mean/median baselines. LLM context learning and conventional supervised model fitting are not identical procedures; only the full-training LLM condition had 253 labelled examples in its prompt context, so zero-shot and few-shot are not matched-data comparisons.

Execution provenance and implementation reproducibility

The retained project records contain the three condition prompts and input/context workbooks. The zero-shot prompt uses only the test input; the few-shot prompt includes 12 labelled examples; the full-training-context prompt includes 253 labelled formulations. The prompts request JSON-only numeric predictions and instruct the model not to browse. The three retained numerical prediction vectors match

the retained case-wise scoring workbook. They are not provider-native execution receipts: provider/API request logs, exact provider/model name and version/snapshot, endpoint, original invocation timestamps, inference settings, retry history, and other hidden system/developer messages were not recovered and are not inferred retrospectively.

The retained historical baseline-fitting script records feature definitions, preprocessing, model constructors, hyperparameters, random states, an intended DOI-grouped split, and a purge step for exact ten-feature matches. Ridge used median imputation, standardization, and alpha = 1.0. Random Forest and Extra Trees each used 500 estimators, min_samples_leaf = 2, max_features = 1.0, random_state = 42, and n_jobs=-1 after median imputation. Gradient Boosting used 200 estimators, learning_rate = 0.03, max_depth = 2, loss='huber', and random_state = 42 after median imputation; mean and median baselines used constant DummyRegressor strategies. The historical script selects the first record at a duplicated timestamp, while the canonical scoring vector uses the figure-supported T045 midpoint. The original pre-run split manifest, exact case-to-formulation map, historical software versions, frozen environment, fit-time logs, and fitted estimator objects were not recovered. The script is retained as an audit artifact and is not represented as a complete end-to-end reproduction of the stored benchmark.

Mean absolute error (MAE) and root mean squared error (RMSE) were calculated on release fractions and multiplied by 100 to report percentage points (pp). R^2 was calculated on the same 67 cases. Predictions were not clipped to [0, 1]. The retained scoring workbook reports paired uncertainty estimates from 10,000 bootstrap replicates resampling 23 nominal DOI labels in the retained audit mapping (seed 20261018), keeping cases assigned to each sampled label together. The original bootstrap implementation, draw-level output, and execution log were not recovered, so the intervals are reported as retained results and were not independently regenerated during manuscript finalization. Because the original case-to-formulation/DOI mapping and split manifest are unavailable, the intervals are conditional on an unverified label mapping and do not authenticate DOI-independent separation or substitute for external validation. The retained procedure reports percentile limits using linear interpolation and the same sampled labels for the primary T045 target and sensitivity scenarios.

Use of generative AI

ChatGPT (OpenAI) was used as the interactive interface for the retained zero-shot, few-shot, and full-training-context prediction requests described in this study. The exact model name/version or snapshot, backend endpoint, original invocation timestamps, inference settings, and provider-native request/response receipts were not preserved or recovered; the historical executions therefore cannot be independently authenticated or fully reproduced. ChatGPT also assisted with protocol and documentation review, consistency checks of retained numerical outputs, figure preparation from retained case-wise data, and manuscript structuring and drafting. ChatGPT is not an author and was not a source of experimental measurements. The sole author is responsible for source verification, interpretation, accuracy, and final approval of the manuscript.

3. Results

Each retained numerical prediction vector contains the 67 expected case identifiers exactly once (201 case-condition predictions in total), with no missing or duplicate case-condition keys and no values outside [0, 1]. The vectors match the retained case-wise scoring workbook. The canonical target vector is frozen to midpoint_T045_figure_supported_2026-10-18 (SHA-256: 13216501dba145da6e5f399671eb13b77864e11e83af25271d4cff8d0650f246); prediction values were retained unchanged, and the saved workbook reports scores against this version.

On the retained set of 67 cases, the zero-shot and few-shot LLM conditions had MAEs of 16.732 pp and 17.150 pp, respectively, and negative R^2 values. The full-training-context LLM achieved MAE 12.269 pp, RMSE 15.770 pp, and R^2 0.217. Random Forest had a marginally lower MAE (12.259 pp), whereas the LLM had a modestly lower RMSE (15.770 vs 15.973 pp) and higher R^2 (0.217 vs 0.196). The MAE difference was + 0.009 pp for LLM minus Random Forest, too small by itself to support a meaningful ranking. Table 1 summarizes these descriptive metrics.

Table 1
Primary fixed-holdout performance

Method	N	MAE (pp)	RMSE (pp)	R^2	Bias (pp)
Zero-shot LLM	67	16.732	20.552	-0.331	-3.886
Few-shot LLM (12 examples)	67	17.150	20.446	-0.317	4.965
Full-training-context LLM (253 examples)	67	12.269	15.770	0.217	-1.294
Random Forest	67	12.259	15.973	0.196	-1.008
Gradient Boosting	67	12.510	16.238	0.169	-3.438
Extra Trees	67	12.698	16.170	0.176	-0.366
Ridge regression	67	13.553	16.971	0.093	0.141
Mean baseline	67	14.425	17.998	-0.020	-2.545
Median baseline	67	15.206	20.389	-0.310	-9.914

Bias is mean(prediction − target); positive values indicate overprediction on average. Lower MAE/RMSE and higher R^2 are better. These are descriptive results on one fixed holdout.

Paired comparison: full-training LLM vs Random Forest

The paired differences and nominal-label cluster-bootstrap intervals are summarized in Table 2.

Table 2
Paired nominal-label cluster-bootstrap comparison of full-training-context LLM versus Random Forest

Metric (LLM − RF)	Observed difference	95% Nominal-label cluster-bootstrap CI	Interpretation
MAE difference (pp)	+ 0.009	−0.884 to + 0.909	Includes 0
RMSE difference (pp)	−0.203	−1.156 to + 0.638	Includes 0
R^2 difference	+ 0.020	−0.061 to + 0.154	Includes 0

All three 95% bootstrap intervals crossed zero. For full-training LLM minus Random Forest, the intervals were − 0.884 to + 0.909 pp for MAE, − 1.156 to + 0.638 pp for RMSE, and − 0.061 to + 0.154 for R^2, based on 10,000 paired resamples of the 23 nominal DOI labels (seed 20261018). These conditional intervals do not establish a clear advantage for either method and do not verify the original split provenance. The retained absolute-error comparison was similarly balanced: the full-training LLM had lower absolute error on 33 cases, and Random Forest on 34, with no ties. Target-construction subgroup results are summarized in Table 3.

Table 3
Descriptive performance by target-construction subgroup (primary midpoint target)

Subset	N	LLM MAE	LLM RMSE	RF MAE	RF RMSE
All test cases	67	12.269	15.770	12.259	15.973
Direct 24-h measurement	3	8.270	9.554	9.231	11.205
Interpolated bracket ≤ 24 h	26	14.229	18.269	13.366	17.475
Interpolated bracket > 24 h	38	11.243	14.245	11.741	15.198
Direct or interpolated bracket ≤ 24 h	29	13.613	17.569	12.938	16.934

MAE/RMSE are in percentage points. The 29-case subset comprises 3 direct measurements plus 26 interpolated targets with bracketing intervals no wider than 24 h; it is not a 29-case direct-measurement evaluation.

Performance differed across target-construction subgroups. In the 26-case short-bracket interpolation subgroup (bracketing interval ≤ 24 h), Random Forest had a lower MAE than the full-training LLM (13.366 vs 14.229 pp). In the 38-case longer-bracket subgroup (interval > 24 h), the full-training LLM had a modestly lower MAE (11.243 vs 11.741 pp). The combined 29-case direct-or-short-bracket analysis also favored Random Forest by MAE (12.938 vs 13.613 pp). The direct-only subgroup contains just three test cases and supports descriptive reporting only.

Table 4 summarizes rescoring against three T045 target definitions (0.574462, figure-supported midpoint 0.576546, and second-record alternative 0.578629), with predictions unchanged. Across nine

methods, rank ordering remained unchanged for MAE, RMSE, and R^2. The Run 3 minus Random Forest MAE difference remained + 0.00905 pp across scenarios, and all retained paired intervals included zero. This sensitivity analysis addresses T045 only, not broader uncertainty in interpolated targets.

Table 4
Sensitivity of full-holdout performance to the three T045 target scenarios

T045 target scenario	Target fraction	Full LLM MAE	Full LLM RMSE	Full LLM R^2	RF MAE	RF RMSE	RF R^2
First raw record	0.574462366	12.265	15.765	0.217	12.256	15.968	0.196
Figure-supported midpoint	0.576545699	12.269	15.770	0.217	12.259	15.973	0.196
Second raw record	0.578629032	12.272	15.775	0.217	12.263	15.977	0.196

Only the T045 reference target changed; all predictions and test-case assignments were held fixed. MAE/RMSE are percentage points; R^2 is unitless. Target fraction is on a 0–1 scale. All values were recomputed from the saved case-wise prediction records.

4. Discussion

The main finding is a near tie between the full-training-context LLM and Random Forest on the retained 67-case set. The LLM produced a slightly lower RMSE and a slightly higher R^2, whereas Random Forest produced a marginally lower MAE; the MAE difference was only + 0.009 percentage points for LLM minus Random Forest. The archived paired bootstrap intervals based on nominal DOI labels included zero, but their group mapping and original computation records were not independently recovered; they should be read as conditional summaries, not verified DOI-cluster inference. These data do not support a decisive winner. Zero-shot and 12-example few-shot conditions performed less favourably than the 253-example context condition, but this is descriptive and cannot isolate the effect of context size from the differing prediction procedures.

The results should be viewed alongside, but not numerically ranked against, published PLGA release-prediction studies. Catapano et al. [3] reported machine-learning prediction of release profiles from a separate collection of 113 PLGA formulations. Alabsi et al. [4] used a harmonized set of 347 formulations and 5,265 release measurements, incorporating both formulation descriptors and early-release information. Most directly relevant to the source data used here, Tang et al. [6] developed MC-NODE using the same Bao et al. dataset of 321 release curves and 4,913 observations, evaluating complete trajectories with DOI-grouped five-fold cross-validation. MC-NODE therefore provides an important reference for modelling this formulation domain, but its trajectory-level task and cross-validation design differ from the present fixed 67-case, 24-h endpoint comparison. The studies collectively show that conventional and specialized machine-learning approaches to PLGA release

prediction already exist. Accordingly, the present work should not be presented as the first AI-based PLGA release study or as a new validated PLGA predictor. Its narrower purpose is to describe the retained outputs of general-purpose LLM prompting conditions relative to conventional regressors, while making endpoint construction and provenance limitations explicit.

The learning conditions are not equivalent in the strict algorithmic sense. The zero-shot LLM received no labeled examples in context, the few-shot condition received 12 examples, and the full-training-context condition received 253 labeled formulations; conventional models were fitted to the designated training data. The comparison is useful as a practical description of the retained procedures, but it cannot isolate model architecture from information exposure or learning mechanism. The three retained numerical prediction vectors match the scoring workbook; however, provider-native execution receipts, model identity/version, runtime settings, and original invocation times remain unavailable. Potential prior model exposure to the public source data also cannot be evaluated without model identity and training-cutoff information.

The archived fitting code describes an intended DOI-grouped split and an exact-feature purge, but the original pre-run case-to-formulation/DOI mapping and split manifest were not recovered. The retained audit mapping contains 23 nominal DOI labels. In a source-group audit, one test descriptor vector mapped to two candidate source formulations; one candidate DOI group is present in the 253-row full-training context, with 18 associated training rows. The original mapping is required to determine whether this is actual overlap for the test case. A single retained set also provides a limited estimate of generalization, and the literature-mined dataset combines heterogeneous experimental protocols. Future evaluations should use a prospectively frozen split manifest, an independent or carefully harmonized dataset, and preferably directly measured 24-h targets. Sensitivity to row order, feature order, feature names, prompt wording, repeated generations, and alternative training-context selections was not evaluated here. This is a material limitation because general-purpose LLM tabular predictions have been shown to vary under task-irrelevant changes in representation and ordering [5].

Endpoint construction is a central interpretive limitation. Only 3 of the 67 test cases have an observed value at exactly 24 h; the other 64 targets are linearly interpolated from bracketing observations. Accordingly, the test scores measure agreement with a specific dataset-derived 24-h label, not accuracy against 67 independent direct measurements at 24 h. The direct-only subgroup (n = 3) is too small to support a reliable comparative claim. Moreover, the interpolation interval varies among cases, so the target can carry different levels of temporal reconstruction uncertainty. The T045 analysis transparently addresses one identifiable ambiguity, but the three-scenario result only demonstrates stability to the three specified T045 values; it cannot quantify uncertainty in every interpolated target.

Taken together, the retained predictions suggest that the full-training-context LLM was competitive with Random Forest on this defined endpoint, but the differences were small and uncertain. The results do not establish broad LLM superiority, reliable performance against direct 24-h measurements, robustness to irrelevant prompt changes, or translational value for formulation design. The defensible contribution is

an exploratory, transparently qualified comparison that makes the endpoint-construction and execution-provenance limitations explicit.

5. Limitations

- Constructed endpoint: 64/67 test targets are interpolated rather than directly measured at 24 h. The primary claim must therefore concern agreement with a defined, dataset-derived 24-h target—not validated accuracy against direct 24-h measurements.
- Sparse direct endpoint: only 7 formulations in the full dataset have direct 24-h values, including 3 in the test split. The direct-only results are too small for reliable generalization claims.
- Exploratory single retained set: 67 cases are associated with 23 nominal DOI labels in the retained audit mapping. A test descriptor vector maps to two possible source formulations, and one candidate DOI group is represented by 18 rows in the full-training context. The original pre-run case-to-formulation/DOI mapping and split manifest were not recovered, so actual overlap status and DOI-independent train/test separation cannot be independently verified. The original bootstrap implementation and draw-level logs were also not recovered; archived intervals are conditional summaries only and do not substitute for external validation.
- Training-condition mismatch: zero-shot, few-shot, full-training-context LLM, and conventional models do not have identical learning procedures or labeled-data exposure. Avoid causal attribution to training-set size or claims of architecture superiority.
- Duplicate-timepoint / error-bar ambiguity: source Fig. 5 was visually inspected. Its early erythromycin curve has a visible vertical SD error bar consistent with the conflicting T045 pair being its lower/upper bounds. The midpoint target is the primary figure-supported estimate; first-record and second-record targets remain sensitivity scenarios. This does not recover author-level replicate data, and crowded overlapping markers prevent claiming exact pixel-level reconstruction. A second conflicting pair occurs in Guanosine but lies outside its 24-h interpolation bracket; exact duplicate pairs for Prilocaine and progesterone also do not alter their 24-h targets. Do not midpoint-average all tied-time pairs automatically without source evidence.
- Execution provenance: the three retained numerical prediction vectors each contain 67 values and match the scoring workbook. Provider-native execution receipts, model/provider name and version, original invocation timestamps, and runtime settings were not recovered. The historical conventional-model fitting script is retained, but its exact software versions, locked environment, and fit-time logs remain unavailable.
- Baseline-fit reproducibility: the historical baseline-fitting script documents model constructors, hyperparameters, preprocessing, feature names, and seeds. Exact Python/library versions, a frozen environment specification, fit-time execution logs, and saved estimator objects remain unavailable; therefore, exact environment-level reproduction cannot be claimed.

6. Conclusion

On the retained 67-case set, the full-training-context LLM and Random Forest showed nearly identical MAE (12.269 vs 12.259 pp); the LLM had a modestly lower RMSE (15.770 vs 15.973 pp). The archived paired bootstrap intervals based on nominal DOI labels included zero, but their group mapping and original computation records were not independently recovered; they should be read as conditional summaries, not verified DOI-cluster inference. Because 64/67 targets were interpolated and only three test cases had direct 24-h measurements, these results concern agreement with constructed endpoints rather than validated accuracy against direct observations. One candidate source-group overlap for the full-training condition remains unresolved, and historical LLM execution metadata and the exact classical-model software environment are incomplete. In light of existing PLGA-specific machine-learning studies, including work using the same underlying source dataset, this manuscript is best interpreted as an exploratory comparison of general-purpose LLM prompting conditions rather than as a new or validated PLGA release predictor.

Declarations

Competing interests:

The author declares no relevant financial or non-financial interests to disclose.

Funding:

No funding was received for conducting this study.

Ethics approval: Not applicable. This study analysed previously published in-vitro formulation and release data and did not involve human participants or live animals.

Author Contribution

Seyed Masih Abtahi Naeini was responsible for conceptualization, methodology, data curation, formal analysis, investigation, writing—original draft, and writing—review and editing. The author reviewed and approved the final manuscript and accepts responsibility for its content.

Data Availability

The data analyzed in this study were derived from previously published sources, as described and cited in the manuscript. The processed datasets and prediction outputs supporting the findings are available from the corresponding author upon reasonable request, subject to source licensing and the limitations of the archived records. Historical raw LLM responses and complete execution metadata were not recoverable and are therefore not available.

References

1. Grizić D, Lamprecht A (2020) Predictability of drug encapsulation and release from propylene carbonate/PLGA microparticles. Int J Pharm 586:119601. https://doi.org/10.1016/j.ijpharm.2020.119601
2. Bao Z, Kim J, Kwok C, Le Devedec F, Allen C (2025) A dataset on formulation parameters and characteristics of drug-loaded PLGA microparticles. Sci Data 12:364. https://doi.org/10.1038/s41597-025-04621-9
3. Catapano AF, Zheng L, Yuan X, Yuan K (2024) Machine learning for predicting drug release behavior of PLGA microspheres. AMIA Annual Symposium Proceedings. ;2024:212–217. https://pmc.ncbi.nlm.nih.gov/articles/PMC12919544/
4. Alabsi H, Bahlool AZ, Bizzari A (2026) Predictive modelling of drug release from PLGA microparticles using artificial intelligence and machine learning. Next Res 10:101842. https://doi.org/10.1016/j.nexres.2026.101842
5. Liu H, Yang M, Adomavicius G (2026) Robustness is important: Limitations of LLMs for predictions on tabular data. PNAS Nexus 5(6):pgag197. https://doi.org/10.1093/pnasnexus/pgag197
6. Tang Z, Li H, Li T, Xue F (2026) MC-NODE: A Mechanism-Decomposed Neural Differential Model for PLGA Microsphere Drug Release Prediction and Attribution. Pharmaceuticals 19(8):1227. https://doi.org/10.3390/ph19081227

Figures

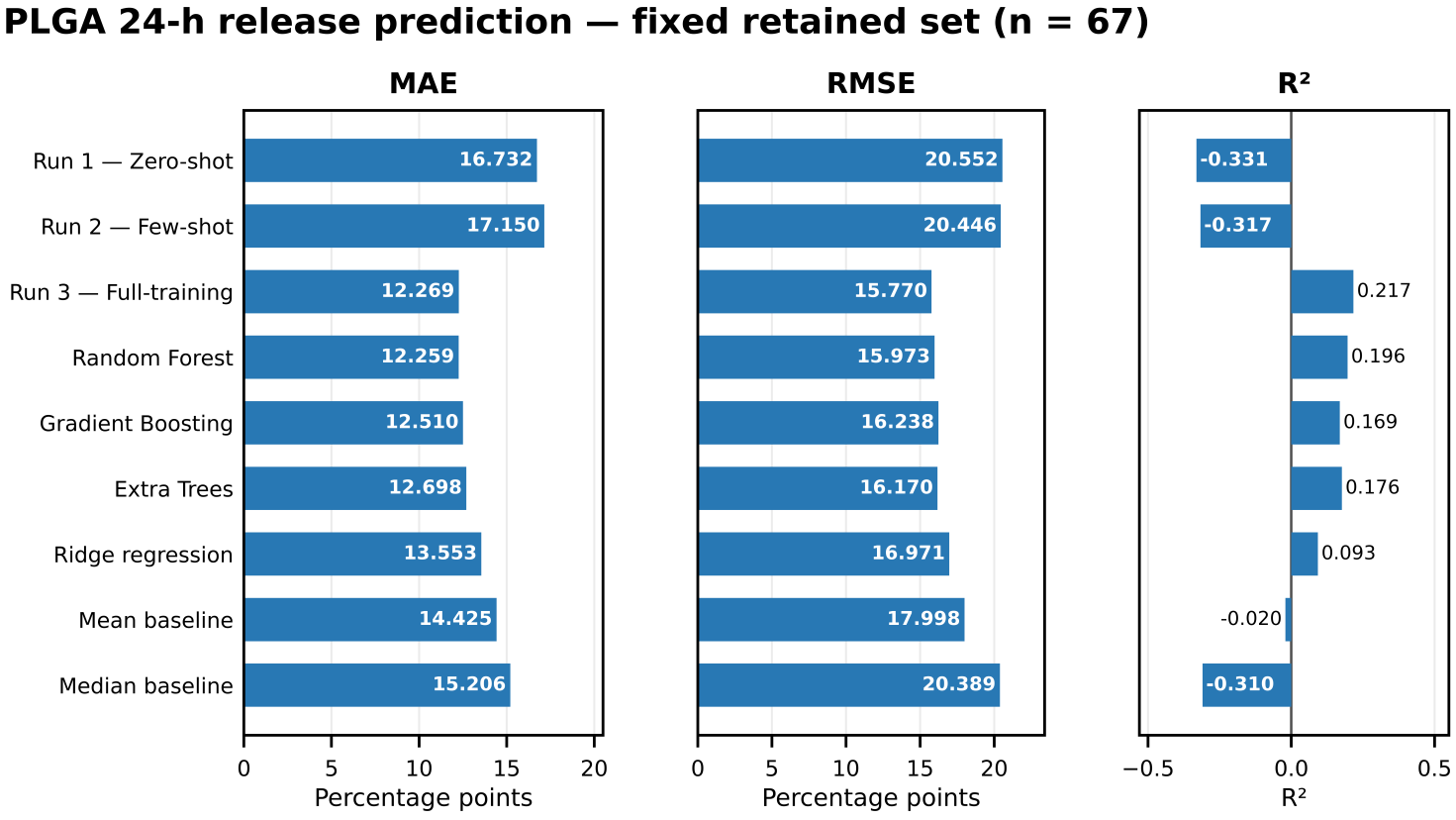


Figure 1

Retained-set performance of three prompt-based LLM conditions and six conventional/baseline methods (n = 67). MAE and RMSE are expressed in percentage points of the release fraction; R^2 is unitless. Values are reported from the retained case-wise predictions against the primary figure-supported T045 midpoint target (0.576545699). The figure summarizes the fixed retained records and does not independently authenticate the original test split or LLM execution metadata.

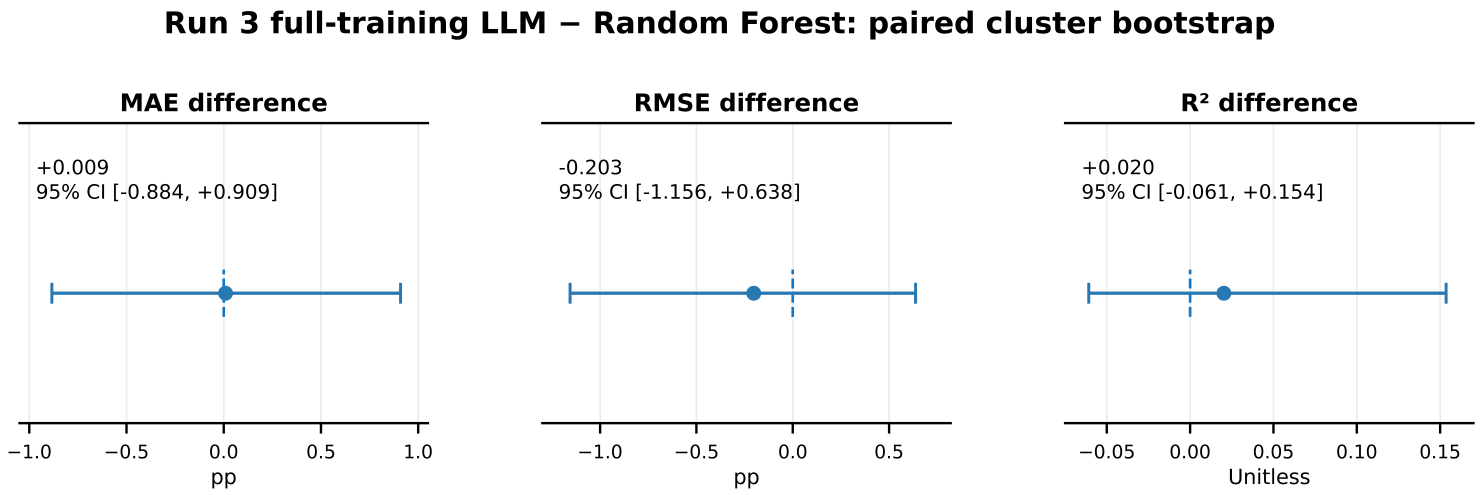


Figure 2

Paired comparison of full-training-context LLM (Run 3) minus Random Forest. Points show the observed differences; horizontal bars show 95% percentile intervals from 10,000 bootstrap resamples of the 23 nominal DOI labels in the retained audit mapping (seed 20261018). Negative MAE/RMSE differences and positive R^2 differences favour Run 3. All intervals include zero. The original pre-run mapping and split manifest were not recovered, so this analysis is conditional on retained nominal labels and does not verify DOI-independent separation.

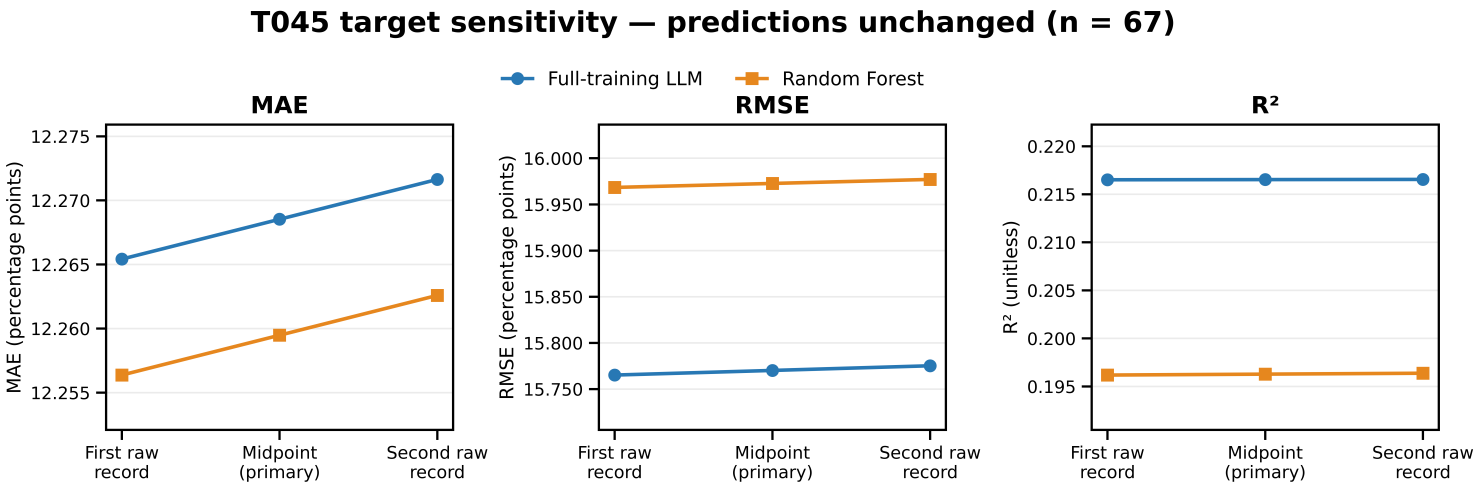


Figure 3

Sensitivity of retained-set performance to three alternative 24-h target definitions for T045. The target was set to 0.574462366 (first raw-record alternative), 0.576545699 (figure-supported midpoint; primary working value), or 0.578629032 (second raw-record alternative); predictions and case assignments were unchanged. MAE and RMSE are in percentage points; R^2 is unitless. Narrow axis ranges highlight small changes; interpret values using the axis labels. This analysis addresses only the T045 ambiguity, not uncertainty in all interpolated targets.