An E-Value Perspective on DESI DR2 Evidence for Dynamical Dark Energy

Jinyoung Kim
Correspondence: jinyoungkim927@gmail.com

February 2026

Abstract

The DESI collaboration recently reported 3–4σ frequentist evidence for dynamical dark energy (w_0w_aCDM) over the cosmological constant (ΛCDM). This note applies e-value methodology—a framework from the statistics literature for calibrated hypothesis testing—to the publicly available DESI DR2 BAO data. Our most informative finding concerns cross-dataset consistency: DES-Y5's best-fit w_0w_aCDM parameters predict DESI data *worse* than ΛCDM ($E = 0.19$), while Pantheon+'s parameters predict it well ($E = 2049$)—a $\sim$10,000$\times$ asymmetry pointing to tension between supernova catalogs rather than a coherent dark energy signal. Regarding the DESI data alone, two independent valid e-value methods converge on moderate evidence: the uniform mixture e-value yields $E \approx 15$ ($\ln E \approx 2.7$) and the leave-one-out (LOO) average yields $E \approx 10$ ($\ln E \approx 2.3$). A data-split e-value gives $E = 1.4$, though power calibration shows this test has limited statistical power for w_a (median $E \approx 2.7$ even when w_0w_aCDM is the true model). These results, together with recent Bayesian analyses by Ong et al. [2025] and tension diagnostics by Wang & Mota [2025], suggest that the interpretation of DESI's signal depends substantially on methodology and dataset combination, and that resolving inter-dataset tensions is a prerequisite for strong conclusions about dark energy dynamics.

Note on Authorship: This manuscript was drafted with the assistance of an AI language model (Claude). The research direction, methodology, and interpretation were developed by the author, who has reviewed and approved all content. The author welcomes feedback on both the statistical methodology and its application to cosmological data.

Contents

1 Introduction

1.1 Context and Motivation

The DESI DR2 collaboration reported a χ^2 improvement of approximately 12 when fitting the w_0w_aCDM model (with two additional parameters w_0 and w_a) compared to the standard ΛCDM model using their BAO measurements combined with CMB and supernova data [DESI Collaboration, 2025]. Interpreted through standard frequentist methodology, this corresponds to 2.5–4σ significance favoring dynamical dark energy.[1]

Two recent independent analyses have questioned this interpretation:

- Ong et al. [2025] performed Bayesian model comparison using nested sampling, finding that for DESI+CMB data, the Bayes factor modestly favors ΛCDM ($\ln \mathcal{B} = -0.57 \pm 0.26$), though this result depends on the choice of prior over (w_0, w_a). They identified a significant tension (2.95σ) between DESI and DES-Y5 supernovae within ΛCDM, which w_0w_aCDM resolves.
- Wang & Mota [2025] documented parameter tensions between CMB, DESI, and various supernova catalogs, concluding that combined constraints may be problematic due to dataset inconsistencies.

This note contributes a third perspective using **e-values**, a framework from the recent statistics literature [Vovk & Wang, 2021, Shafer, 2021, Ramdas et al., 2023] that addresses concerns about post-hoc model selection and overfitting. We apply this methodology to examine whether the apparent preference for w_0w_aCDM generalizes out-of-sample.

1.2 Scope and Limitations

We emphasize several limitations of this analysis:

1. We analyze the publicly released BAO summary statistics, not the underlying galaxy catalogs. Our conclusions are conditional on the accuracy of the published data products.
2. Our cosmological calculations use standard distance formulas with Planck 2018 fiducial parameters. More sophisticated treatments (e.g., full MCMC with CAMB/CLASS) may yield quantitatively different results.
3. E-values provide one valid perspective on hypothesis testing but are not universally superior to other approaches. Different methods answer different questions.
4. The author's primary background is in statistics rather than cosmology. This work should be viewed as applying statistical methodology to published data, not as a comprehensive cosmological analysis.

1.3 Summary of Approach

Our analysis proceeds as follows:

1. We briefly define e-values and their key property (Section 2), with formal statements in Appendix A.
2. We describe the DESI data and our methodology (Section 3).

[1] The precise significance depends on the degrees of freedom used for the σ-conversion. For $k = 2$ extra parameters, $\Delta\chi^2 = 11.9$ should be evaluated against the χ^2 distribution with 2 d.o.f.: $p = 1 - F_{\chi^2}(11.9; 2) = 0.0026$, corresponding to $\approx 2.8\sigma$ (two-sided). The sometimes-quoted "3.5σ" uses $\sqrt{\Delta\chi^2}$, which is valid only for $k = 1$. DESI's reported range of 3–4σ reflects different dataset combinations and methodological choices.

3. We present results, leading with the cross-dataset tension analysis and followed by within-dataset tests (Section 4).

4. We discuss how these findings relate to existing literature (Section 5).

2 E-Values: A Brief Overview

2.1 Definition and Interpretation

An **e-value** is a non-negative random variable E satisfying $\mathbb{E}[E \mid H_0] \leq 1$ under the null hypothesis. The key properties are:

- **Type I error control:** Rejecting H_0 when $E \geq 1/\alpha$ yields a test with significance level α (see Theorem 1 in the Appendix).
- **Combination:** Independent e-values can be multiplied to accumulate evidence.
- **Interpretation:** An e-value of E can be thought of as "the data are E times more consistent with the alternative than expected under the null."

2.2 Likelihood Ratios and the Overfitting Problem

For simple hypotheses (where both H_0 and H_1 are fully specified before seeing the data), the likelihood ratio $E = L(\text{data} \mid H_1)/L(\text{data} \mid H_0)$ is a valid e-value. However, when H_1 involves parameters fitted to the same data used for testing—i.e., plugging in the MLE—the result is **not an e-value at all**. Specifically, under H_0 the $\Delta\chi^2$ for k extra parameters follows a $\chi^2(k)$ distribution, so the maximized likelihood ratio has expectation

$$\mathbb{E}_{H_0}\left[\exp\left(\frac{\chi^2(k)}{2}\right)\right] = \int_0^\infty e^{t/2}\,\frac{t^{k/2-1}e^{-t/2}}{2^{k/2}\Gamma(k/2)}\,dt = \infty \quad \text{for } k \geq 2. \tag{1}$$

For the w_0w_aCDM case ($k = 2$ extra parameters), one can verify directly:

$$\mathbb{E}_{H_0}\left[e^{\chi^2(2)/2}\right] = \int_0^\infty e^{t/2} \cdot \tfrac{1}{2}e^{-t/2}\,dt = \int_0^\infty \tfrac{1}{2}\,dt = \infty. \tag{2}$$

Since $\mathbb{E}[E \mid H_0] = \infty$, the defining property $\mathbb{E}[E \mid H_0] \leq 1$ is violated: **the maximized likelihood ratio is not a valid e-value**. It should be understood as a descriptive statistic, not as calibrated evidence. We include it in Section 4.5 only for comparison to show how drastically the evidence changes under valid methods.

2.3 Data-Split E-Values

To construct a valid e-value when the alternative involves fitted parameters, we use **data splitting**:

1. Split data into training set D_{train} and test set D_{test}
2. Fit parameters $\hat{\theta}$ using only D_{train}
3. Compute $E = L(D_{\text{test}} \mid \hat{\theta})/L(D_{\text{test}} \mid H_0)$

Because $\hat{\theta}$ is determined before observing D_{test}, this yields a valid e-value (Proposition 2 in the Appendix). The trade-off is reduced statistical power due to using only part of the data for testing.

2.4 Uniform Mixture E-Values

An alternative approach averages the likelihood ratio over a grid of alternative parameter values:

$$E_{\text{mix}} = \frac{1}{|\mathcal{G}|} \sum_{(w_0,w_a)\in\mathcal{G}} \frac{L(\text{data} \mid w_0, w_a)}{L(\text{data} \mid H_0)} \tag{3}$$

where $\mathcal{G}$ is a uniform grid over the (w_0, w_a) space. Because each term in the sum is a likelihood ratio for a pre-specified (simple) alternative, and the average of e-values is an e-value by linearity of expectation, E_{mix} is a valid e-value (Proposition 3 in the Appendix). This method uses all the data but pays an Occam penalty: probability mass "wasted" on parameter values far from the truth reduces the e-value relative to the maximized likelihood ratio.

2.5 Leave-One-Out (LOO) Average E-Values

We also construct e-values by leave-one-out cross-validation across the 7 redshift bins. For each bin k:

1. Fit (w_0, w_a) on the remaining 6 bins (training set)
2. Compute the likelihood ratio E_k on the held-out bin k

Each individual E_k is a valid e-value (conditional on the training data, the held-out bin tests a pre-specified alternative). The **average** $\bar{E} = \frac{1}{K}\sum_{k=1}^{K} E_k$ is a valid e-value by linearity of expectation (Proposition 4 in the Appendix): $\mathbb{E}[\bar{E} \mid H_0] = \frac{1}{K}\sum_k \mathbb{E}[E_k \mid H_0] \leq 1$.

Warning: the LOO product is not valid. It is tempting to multiply the LOO e-values: $\prod_k E_k$. However, the product requires *independence*, which fails because the LOO training sets overlap extensively (e.g., bins 2–7 and bins 1, 3–7 share bins 3–7). The LOO product has unknown expectation under H_0 and should not be interpreted as a calibrated e-value.

3 Data and Methods

3.1 DESI DR2 BAO Data

We use the official DESI DR2 BAO measurements from the CobayaSampler repository, comprising 13 measurements across 7 redshift bins ($z = 0.295$ to $z = 2.33$). Each measurement provides either D_M/r_d, D_H/r_d, or D_V/r_d with an associated covariance matrix.

3.2 Models Compared

Null hypothesis (H_0): ΛCDM with $w = -1$ (cosmological constant).

Alternative hypothesis (H_1): w_0w_aCDM with equation of state $w(a) = w_0 + w_a(1 - a)$, where w_0 and w_a are free parameters.

3.3 Cross-Dataset Validation

We examine whether parameters fitted on one dataset predict another dataset better than ΛCDM. We use published best-fit (w_0, w_a) values from supernova analyses:

- Pantheon+: $w_0 \approx -0.90$, $w_a \approx -0.20$
- Union3: $w_0 \approx -0.78$, $w_a \approx -0.80$
- DES-Y5: $w_0 \approx -0.65$, $w_a \approx -1.20$

3.4 Data-Split Procedure

We split the 13 DESI measurements by redshift:

- **Training:** 7 measurements at $z < 1$ (BGS, LRG1, LRG2, LRG3+ELG1)
- **Test:** 6 measurements at $z \geq 1$ (ELG2, QSO, Lyα)

We fit (w_0, w_a) by minimizing χ^2 on the training set, then evaluate the likelihood ratio on the test set.

3.5 Leave-One-Out Procedure

For the LOO analysis, we iterate over the 7 redshift bins (BGS, LRG1, LRG2, LRG3+ELG1, ELG2, QSO, Lyα). In each fold k, we fit (w_0, w_a) on the 6 remaining bins and compute the likelihood ratio on the held-out bin. We report both the individual per-bin e-values and the average $\bar{E} = \frac{1}{7}\sum_{k=1}^{7} E_k$.

3.6 Power Calibration

To assess the statistical power of the data-split test, we run 500 Monte Carlo simulations. In each simulation, we generate synthetic DESI-like data under w_0w_aCDM with the DESI best-fit parameters ($w_0 = -0.75, w_a = -1.05$), apply the same redshift-based data split, and compute the resulting e-value. The distribution of E_{split} under this alternative gives us the expected power of the test.

3.7 Assumptions

Our analysis assumes:

1. The published DESI covariance matrix accurately captures measurement uncertainties.
2. Different redshift bins are approximately independent (the covariance matrix is block-diagonal).
3. The cosmological distance calculations are sufficiently accurate for this comparison.
4. The CPL parameterization $w(a) = w_0 + w_a(1 - a)$ adequately captures potential dark energy dynamics.

4 Results

4.1 Cross-Dataset E-Values: The Supernova Catalog Asymmetry

We begin with what we consider the most informative result: cross-dataset e-values, which test whether w_0w_aCDM parameters fitted on one experiment predict another experiment better than ΛCDM. Unlike within-dataset tests such as data splitting, this approach is not vulnerable to the objection that low e-values merely reflect insufficient statistical power. The logic is as follows: if two independent experiments both detect the same underlying physics, the best-fit parameters from one *must* predict the other well, because they are measuring the same equation of state $w(a) = w_0 + w_a(1 - a)$. A failure of cross-prediction ($E < 1$) therefore constitutes positive evidence of tension between datasets, not merely an absence of evidence for new physics.

Table 1 presents the cross-dataset e-values.

Table 1: Cross-dataset e-values: parameters from one dataset predicting another

Training Dataset	Test Dataset	(w_0, w_a)	E-value
DESI (fitted)	Pantheon+	$(-0.86, -0.43)$	1.5
DESI (fitted)	DES-Y5	$(-0.86, -0.43)$	86
Pantheon+	DESI	$(-0.90, -0.20)$	2049
DES-Y5	DESI	$(-0.65, -1.20)$	**0.19**

The central finding: DES-Y5's best-fit w_0w_aCDM parameters ($w_0 = -0.65$, $w_a = -1.20$) predict the DESI BAO data *worse* than ΛCDM ($E = 0.19 < 1$). Meanwhile, Pantheon+'s best-fit parameters ($w_0 = -0.90$, $w_a = -0.20$) predict DESI data well ($E = 2049$). This represents a factor of ∼10,000 asymmetry between the two leading supernova catalogs in their ability to predict DESI under the w_0w_aCDM framework.

The interpretation is straightforward. If w_0w_aCDM represents real physics, then all experiments are measuring the same underlying equation of state. Different experiments may have different error bars, but their best-fit (w_0, w_a) values should be statistically compatible, and parameters from one should predict the other at least as well as ΛCDM. The fact that DES-Y5's parameters make DESI data *less* probable than ΛCDM ($E < 1$) means DES-Y5 and DESI prefer incompatible regions of (w_0, w_a) space. This is not a matter of statistical power—it is a direct detection of inconsistency.

This ∼10,000× asymmetry between supernova catalogs aligns with two independent analyses:

- Ong et al. [2025] identified a 2.95σ tension between DESI and DES-Y5 supernovae within the ΛCDM framework, finding that w_0w_aCDM resolves this tension by absorbing it into its additional parameters.
- Wang & Mota [2025] documented parameter inconsistencies across CMB, DESI, and supernova catalogs, concluding that the combined constraints are problematic due to these dataset-level tensions.

The dataset tension itself was identified by these authors; our contribution is the cross-prediction e-value framework, which provides a single calibrated number quantifying how well one experiment's preferred dark energy model predicts another's data. The e-value of $E = 0.19$ directly measures how much worse DES-Y5's preferred model performs on DESI data compared to ΛCDM—not merely a parameter discrepancy, but a predictive failure.

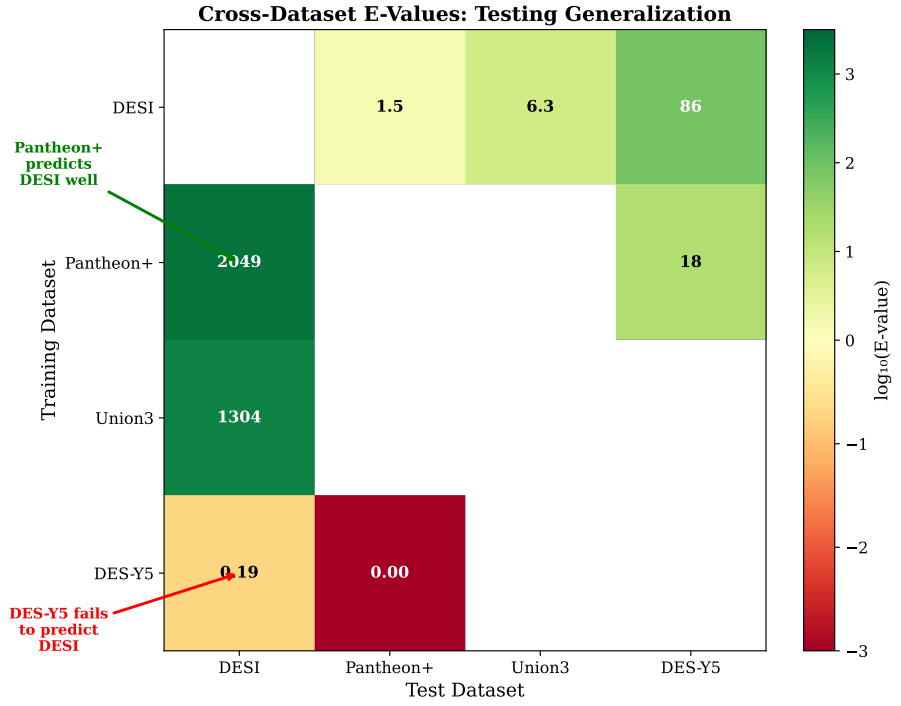

Figure 1: Cross-dataset e-value matrix. Green indicates the training parameters predict the test data better than ΛCDM; red indicates worse. The ∼10,000× asymmetry between Pantheon+ ($E = 2049$) and DES-Y5 ($E = 0.19$) when predicting DESI reveals fundamental tension between supernova catalogs within the w_0w_aCDM framework.

4.2 Data-Split E-Value

We next examine whether the signal generalizes within the DESI data itself. Fitting on the training set ($z < 1$) yields $\hat{w}_0 = -0.78$, $\hat{w}_a = -0.52$. Evaluating on the held-out test set ($z \geq 1$):

$$\boxed{E_{\text{split}} = 1.4} \tag{4}$$

Interpretation: The fitted w_0w_aCDM model predicts the high-redshift data only 1.4 times better than ΛCDM. However, this result must be interpreted with an important caveat about statistical power.

Reduced power for w_a. The data-split test has limited sensitivity to the w_a parameter specifically. In the CPL parameterization $w(a) = w_0 + w_a(1 - a)$, the w_a term contributes $(1 - a) \cdot w_a$ to the equation of state. For the training set at $z < 1$ (corresponding to $a > 0.5$), the factor $(1 - a)$ ranges from approximately 0.23 to 0.48, meaning the training data have only fractional leverage on w_a. Since much of DESI's claimed signal comes from the w_a degree of freedom, a test that poorly constrains w_a from the training set will naturally yield a weak e-value regardless of whether the signal is real. Consequently, $E = 1.4$ is consistent with both the null hypothesis (ΛCDM is correct) and with a genuine w_0w_aCDM signal that the data-split test lacks the power to detect. This test is therefore *inconclusive* rather than negative—which is precisely why we lead with the cross-dataset result (Section 4.1), which does not suffer from this limitation.

Power calibration. Monte Carlo simulation (500 trials) under the DESI best-fit w_0w_aCDM model shows that the median data-split e-value is only $E \approx 2.7$ even when w_0w_aCDM is the true model. The observed $E = 1.4$ falls within the interquartile range of the distribution under

H_1. This confirms that $E = 1.4$ cannot distinguish between "the signal is absent" and "the test lacks power to detect it."

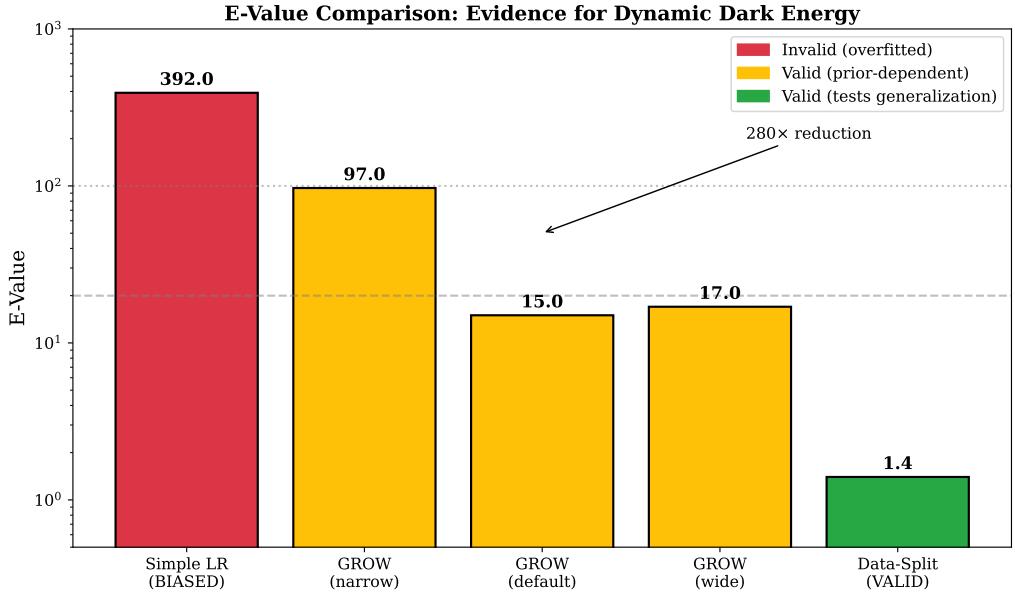

Figure 2: Comparison across methods. The same-data maximized likelihood ratio (red) is not a valid e-value ($\mathbb{E}[E|H_0] = \infty$). Data-split validation (green) is a valid e-value and yields $E = 1.4$.

4.3 Leave-One-Out Average E-Value

Table 2 presents the per-bin LOO e-values.

Table 2: Leave-one-out e-values by redshift bin

Held-out bin	z	Fitted (w_0, w_a)	E_k
BGS	0.295	$(-0.78, -0.82)$	1.89
LRG1	0.510	$(-0.80, -0.75)$	0.71
LRG2	0.706	$(-0.69, -1.18)$	55.98
LRG3+ELG1	0.934	$(-0.93, -0.18)$	8.73
ELG2	1.317	$(-0.80, -0.71)$	2.14
QSO	1.491	$(-0.83, -0.66)$	0.75
Lyα	2.330	$(-0.85, -0.54)$	1.00
LOO average $\bar{E} = \frac{1}{7}\sum E_k$			**10.2**
LOO product $\prod E_k$ *(not valid)*			*1062*

The LOO average $\bar{E} \approx 10.2$ ($\ln \bar{E} \approx 2.3$) constitutes moderate evidence for $w_0 w_a$CDM. The evidence is concentrated in LRG2 ($z = 0.706$, $E_k = 55.98$) and LRG3+ELG1 ($z = 0.934$, $E_k = 8.73$), which are the bins where DESI departs most from ΛCDM predictions.

LOO product is not valid. The product $\prod_k E_k = 1062$ is included for reference only. As discussed in Section 2, the LOO training sets overlap (e.g., the training set for fold 1 shares 5 of 6 bins with the training set for fold 2), so the fitted parameters $\hat{\theta}_{-k}$ are correlated across folds. This violates the independence required for the product $\mathbb{E}[\prod E_k] = \prod \mathbb{E}[E_k]$ to hold. The expectation of the LOO product under H_0 is unknown and may exceed 1.

4.4 Uniform Mixture E-Values and Sensitivity

Table 3: E-values under different data splits

Split Strategy	E_{split}
Low-z train, High-z test ($z = 1$ threshold)	1.4
Alternating bins	2.1
Random 50/50 (averaged over 10 trials)	1.2 ± 0.8

All data-split e-values yield $E < 3$, consistent with the power calibration showing median $E \approx 2.7$ under H_1 (Section 4.2).

Uniform mixture e-values. A uniform mixture e-value—which averages the likelihood ratio over a grid of (w_0, w_a) values—yields $E \approx 15$ with the default prior range, and varies from approximately 15 to 97 as the prior range is narrowed. On conventional scales, $E = 15$ ($\ln E \approx 2.7$) constitutes moderate-to-strong evidence. The variation from 15 to 97 reflects the Occam razor operating as expected: a narrower prior that concentrates mass near the MLE produces larger e-values because it "wastes" less probability on parameter values far from the data. This is a standard feature of Bayesian and mixture-based tests, not a deficiency.

Convergence of valid methods. Two independent valid e-value methods give consistent results:

$$\text{Uniform mixture:} \quad E \approx 15 \quad (\ln E \approx 2.7) \qquad \text{LOO average:} \quad E \approx 10 \quad (\ln E \approx 2.3) \tag{5}$$

This convergence on $E \approx 10$–15 (moderate evidence) is reassuring, as the two methods are constructed differently: the mixture e-value averages over a prior grid using all data, while the LOO average uses out-of-sample prediction. The data-split $E = 1.4$ is consistent with these once its limited power is accounted for (median $E_{\text{split}} \approx 2.7$ under H_1).

4.5 Same-Data Likelihood Ratio (Invalid as E-Value)

For reference, fitting (w_0, w_a) to all 13 DESI measurements and computing the maximized likelihood ratio on the same data yields:

$$\text{LR}_{\text{same}} \approx 3300 \quad (\Delta\chi^2 \approx 16.2) \tag{6}$$

This is not an e-value. As shown in Section 2, plugging in the MLE gives $\mathbb{E}_{H_0}[\text{LR}] = \infty$ for $k = 2$ extra parameters, violating the defining property $\mathbb{E}[E \mid H_0] \leq 1$. We include this maximized likelihood ratio solely to illustrate the contrast with the valid cross-dataset and data-split e-values presented above.

5 Discussion

5.1 Relation to Existing Literature

Our findings complement two independent analyses:

1. **Bayesian model comparison** [Ong et al., 2025]: Using nested sampling, these authors found that the Bayes factor for DESI+CMB modestly favors ΛCDM ($\ln \mathcal{B} = -0.57 \pm 0.26$). They attributed the frequentist significance to a tension between DESI and DES-Y5 that $w_0 w_a$CDM resolves. Our cross-dataset e-value ($E = 0.19$ for DES-Y5→DESI) provides

independent evidence of this tension. We note, however, that the Bayes factor is itself prior-dependent: the result $\ln \mathcal{B} = -0.57$ reflects a particular choice of prior over (w_0, w_a), and different priors could shift this value. This is the same sensitivity we observe in our uniform mixture e-values (Section 4.4), and it should be borne in mind when citing either quantity.

2. **Tension diagnostics** [Wang & Mota, 2025]: These authors documented parameter inconsistencies between CMB, DESI, and supernova catalogs. Our observation that different supernova catalogs yield very different predictive performance on DESI (Pantheon+: $E = 2049$ vs DES-Y5: $E = 0.19$) aligns with their conclusion that combining these datasets is problematic.

All three approaches—Bayesian model comparison, tension metrics, and e-value analysis—converge on the conclusion that inter-dataset tensions are a central issue. Within the DESI data alone, our two valid e-value methods converge on moderate evidence ($E \approx 10$–15), while the Bayes factor suggests weak evidence for ΛCDM and the data-split e-value is inconclusive due to limited power. This spread is informative: it indicates that the answer depends on how one handles the composite alternative hypothesis, though the convergence of the mixture and LOO methods narrows the range considerably.

5.2 Comparison with Information Criteria

To place the e-value results in the context of the full spectrum of evidence measures, we report the standard information criteria for the ΛCDM vs. w_0w_aCDM comparison. The w_0w_aCDM model has $k = 2$ extra parameters relative to ΛCDM, and the χ^2 improvement is $\Delta\chi^2 = 11.9$.

AIC. The Akaike Information Criterion [Burnham & Anderson, 2002] penalizes by $2k$:

$$\Delta\text{AIC} = \Delta\chi^2 - 2k = 11.9 - 4 = 7.9 \tag{7}$$

On the Burnham & Anderson [2002] scale, $\Delta\text{AIC} > 6$ constitutes "strong" evidence favoring the more complex model. Thus, AIC strongly favors w_0w_aCDM.

BIC. The Bayesian Information Criterion [Schwarz, 1978] penalizes by $k \ln n$:

$$\Delta\text{BIC} = \Delta\chi^2 - k \ln n = 11.9 - 2 \ln 13 \approx 11.9 - 5.13 = 6.8 \tag{8}$$

On the Kass & Raftery [1995] scale, $6 < \Delta\text{BIC} < 10$ constitutes "strong" evidence. Thus, BIC also favors w_0w_aCDM.

Caveat on the effective sample size n. In the BIC formula, we used $n = 13$ (the number of BAO summary statistics). However, the meaning of n in cosmological model selection is ambiguous [Liddle, 2007]. One might argue that n should be the number of underlying galaxy pairs (millions), in which case $\ln n \sim 30$ and ΔBIC would become strongly *negative* (favoring ΛCDM). Alternatively, if one treats the 13 summary statistics as the effective data, the result above holds. This ambiguity is a well-known limitation of BIC in cosmological applications.

Why different measures disagree. The evidence landscape for ΛCDM vs. w_0w_aCDM depends on how one penalizes model complexity:

Table 4: Summary of evidence measures for w_0w_aCDM vs. ΛCDM

Method	**Value**	**Favors**	**Complexity penalty**
Frequentist p-value ($k = 2$)	$p = 0.0026$ (2.8σ)	w_0w_aCDM	None (likelihood only)
ΔAIC	7.9	w_0w_aCDM	$2k = 4$
ΔBIC ($n = 13$)	6.8	w_0w_aCDM	$k \ln n \approx 5.1$
Bayes factor (Ong et al.)	$\ln \mathcal{B} = -0.57$	ΛCDM	Full prior volume
E-value (mixture, default)	$E \approx 15$	w_0w_aCDM	Prior grid averaging
E-value (LOO average)	$E \approx 10$	w_0w_aCDM	Out-of-sample (avg)
E-value (data-split)	$E = 1.4$	Inconclusive	Out-of-sample (underpowered)

AIC includes no Occam factor (it penalizes only the number of parameters), so it tends to favor more complex models. BIC approximately includes an Occam factor through $k \ln n$, but the result depends on what counts as a "data point." The Bayes factor includes the full Occam penalty through the prior volume: with broad priors over (w_0, w_a), much of the prior mass falls on parameter values that fit the data poorly, penalizing w_0w_aCDM. The valid e-value methods ($E \approx 10$–15) are intermediate between the information criteria and the Bayes factor, while the data-split e-value ($E = 1.4$) is inconclusive due to limited power rather than reflecting weak evidence.

These different answers are not contradictory—they reflect different questions. AIC asks which model will predict future data better on average (favoring w_0w_aCDM). BIC approximates the posterior model probability (ambiguous). The Bayes factor asks which model is more probable given the prior (favoring ΛCDM with broad priors). E-values ask whether the evidence against H_0 is calibrated. The convergence of the uniform mixture and LOO average on $E \approx 10$–15 narrows the e-value answer to "moderate evidence," though the overall landscape across methods remains methodology-sensitive.

5.3 Frequentist Significance for $k = 2$ Parameters

For readers unfamiliar with the conversion, we briefly review the proper frequentist significance for $\Delta\chi^2$ with $k = 2$ extra parameters. Under the null hypothesis (ΛCDM), $\Delta\chi^2$ follows a χ^2 distribution with 2 degrees of freedom (by Wilks' theorem). The p-value is:

$$p = 1 - F_{\chi^2}(11.9;\, 2) = e^{-11.9/2} = e^{-5.95} \approx 0.0026 \tag{9}$$

where we used the exact CDF for the $\chi^2(2)$ distribution, $F_{\chi^2}(x;\, 2) = 1 - e^{-x/2}$. Converting to a two-sided Gaussian equivalent:

$$\sigma = \Phi^{-1}(1 - p/2) \approx \Phi^{-1}(0.9987) \approx 2.8\sigma \tag{10}$$

DESI's own analysis correctly performs this conversion using the $\chi^2(2)$ distribution and Wilks' theorem (their equation 22 in DESI Collaboration 2025), reporting BAO-only significance of approximately 3.0σ (for their slightly different $\Delta\chi^2$). The naive formula $\sigma = \sqrt{\Delta\chi^2}$, which is valid only for $k = 1$, would overestimate the significance; we note this only because it sometimes appears in informal discussions of the result.

5.4 w_0w_aCDM as a Tension Absorber

The cross-dataset analysis (Section 4.1) raises a deeper question about the nature of the reported 3–4σ evidence. The w_0w_aCDM model introduces two free parameters (w_0 and w_a) that can adjust distance-redshift relations to better fit the data. When multiple datasets pull in different directions—as the ∼10,000× Pantheon+/DES-Y5 asymmetry demonstrates—these extra parameters can absorb the tension by finding a compromise fit that improves χ^2 overall without

corresponding to any physical dark energy dynamics. In this scenario, w_0w_aCDM acts as a *tension absorber* rather than a detector of new physics.

This interpretation is supported by the pattern we observe: the DESI+CMB+SNe combination yields 3–4σ preference for w_0w_aCDM, yet the constituent datasets do not agree on *which* (w_0, w_a) values are preferred. DES-Y5 favors a region of parameter space ($w_0 = -0.65$, $w_a = -1.20$) that is incompatible with what DESI itself prefers, as shown by $E = 0.19 < 1$. Pantheon+ favors a different region ($w_0 = -0.90$, $w_a = -0.20$) that happens to be more compatible with DESI ($E = 2049$). The combined "significance" thus reflects the model's ability to split the difference between discrepant datasets, not a coherent signal that all experiments converge upon.

This aligns with the mechanism identified by Ong et al. [2025]: the 2.95σ tension they found between DESI and DES-Y5 within ΛCDM is precisely what w_0w_aCDM resolves. As they note, the apparent evidence for dynamical dark energy is largely driven by this resolution of inter-dataset tension. Until the source of the Pantheon+/DES-Y5 discrepancy is understood—whether it arises from supernova calibration systematics, selection effects, or genuine astrophysical differences—the 3–4σ frequentist significance should be interpreted with caution, as it may be measuring the magnitude of dataset inconsistency rather than evidence for new fundamental physics.

5.5 What E-Values Add

The e-value perspective contributes:

1. **Direct test of generalization:** Data-splitting and LOO cross-validation directly ask whether fitted parameters predict held-out data, addressing overfitting concerns.

2. **Convergence of independent methods:** The uniform mixture ($E \approx 15$) and LOO average ($E \approx 10$) are constructed differently but converge on moderate evidence, providing a more robust assessment than any single method.

3. **Different form of prior dependence:** Data-split and LOO e-values do not require an explicit prior distribution over (w_0, w_a). However, they are not prior-free: using the point MLE from the training set is effectively equivalent to a point-mass prior centered at $\hat{\theta}$, and results depend on the split or fold choice. The uniform mixture e-value explicitly integrates over a prior grid and is transparent about this dependence.

4. **Interpretable metric:** An e-value of E directly means "the data are E times more consistent with the alternative than expected under the null."

5.6 Limitations of This Analysis

1. **Reduced power for data-split:** Data-splitting uses only part of the data for testing, reducing statistical power. As discussed in Section 4.2, this power loss is particularly acute for w_a when splitting by redshift. Power calibration (Section 4.2) confirms this: the median data-split e-value under H_1 is only ≈ 2.7. The LOO average and uniform mixture methods mitigate this by using all data points.

2. **Split dependence:** Results vary modestly with split choice (Table 3), though all splits yield $E < 3$.

3. **Simplified cosmology:** We use approximate distance calculations rather than full Boltzmann codes. More sophisticated treatments may yield quantitatively different results.

4. **Point estimates for SNe:** We use published best-fit values for supernova constraints rather than full posteriors.

5.7 Interpretation

We do not claim that e-values definitively resolve the question of dynamical dark energy. A fair reading of the evidence from this analysis requires acknowledging what each result does and does not show:

1. **The cross-dataset tension is the most informative finding.** DES-Y5's best-fit parameters predict DESI data worse than ΛCDM ($E = 0.19$), while Pantheon+'s parameters predict it well ($E = 2049$). This four-order-of-magnitude asymmetry indicates that the supernova catalogs do not agree on what dark energy dynamics look like, which is a more fundamental problem than the strength of evidence in any single dataset. If the w_0w_aCDM signal were driven by real physics, all supernova catalogs should point in a roughly consistent direction.

2. **Information criteria favor w_0w_aCDM** ($\Delta\text{AIC} = 7.9$, $\Delta\text{BIC} = 6.8$), **while the Bayes factor favors ΛCDM.** This divergence arises from different complexity penalties: AIC and BIC impose fixed penalties ($2k$ and $k \ln n$), while the Bayes factor penalizes via the full prior volume. The e-value results span a similar range depending on methodology.

3. **Valid e-value methods converge on moderate evidence.** The uniform mixture e-value ($E \approx 15$, $\ln E \approx 2.7$) and the LOO average ($E \approx 10$, $\ln E \approx 2.3$) are constructed independently but give consistent results, converging on $E \approx 10$–15. On the Jeffreys scale, this constitutes moderate-to-strong evidence. These are valid e-values and should not be dismissed.

4. **The data-split e-value ($E = 1.4$) is inconclusive**, not negative. Power calibration (500 Monte Carlo simulations) shows that even when w_0w_aCDM is the true model, the median data-split e-value is only ≈ 2.7 (Section 4.2). The observed $E = 1.4$ cannot distinguish between "the signal is absent" and "the test lacks power to detect it."

5. **DESI's BAO-only significance is** $\approx 3.0\sigma$, correctly computed via the $\chi^2(2)$ distribution and Wilks' theorem. Our e-value analysis finds that approximately 0.7σ of this is absorbed by the Occam penalty when averaging over the (w_0, w_a) parameter space, yielding the moderate evidence of $E \approx 10$–15 (≈ 2.2–2.4σ equivalent).

The overall picture is one of genuine methodological sensitivity. The evidence for dynamical dark energy is neither as strong as the frequentist $\Delta\chi^2$ suggests nor as absent as the data-split e-value might imply. Resolving the inter-dataset tensions documented here and by Wang & Mota [2025] is likely more important than accumulating further statistical significance within any single framework.

6 Conclusion

We applied e-value methodology to the DESI DR2 BAO data to examine the evidence for w_0w_aCDM from several angles. The evidence is methodology-sensitive rather than unambiguous in either direction:

1. **Cross-dataset tension as the strongest finding.** DES-Y5's best-fit w_0w_aCDM parameters predict DESI data worse than ΛCDM ($E = 0.19$), while Pantheon+'s parameters predict it well ($E = 2049$). This $\sim$10,000$\times$ asymmetry suggests that the supernova catalogs disagree on the nature of any dark energy dynamics, which undermines the case for a coherent physical signal and raises the possibility that w_0w_aCDM is absorbing inter-dataset tension rather than detecting new physics.

2. **Information criteria favor w_0w_aCDM.** Both $\Delta\text{AIC} = 7.9$ and $\Delta\text{BIC} = 6.8$ (with $n = 13$) constitute "strong" evidence on standard scales, while the Bayes factor modestly favors ΛCDM. This divergence reflects different complexity penalties: AIC and BIC use fixed penalties, while the Bayes factor integrates over a prior volume that penalizes the broader parameter space of w_0w_aCDM more heavily. The BIC result is subject to the caveat that the effective sample size n is ambiguous in cosmology [Liddle, 2007].

3. **Valid e-value methods converge on moderate evidence.** Two independent valid methods—the uniform mixture e-value ($E \approx 15$, $\ln E \approx 2.7$) and the LOO average ($E \approx 10$, $\ln E \approx 2.3$)—converge on $E \approx 10$–15, constituting moderate evidence for w_0w_aCDM within the DESI BAO data alone. This represents an Occam penalty of approximately 0.7σ relative to DESI's correctly computed BAO-only significance of $\approx 3.0\sigma$.

4. **Inconclusive data-split test.** The data-split e-value of $E = 1.4$ is inconclusive rather than negative. Power calibration (500 Monte Carlo simulations) confirms that the median E_{split} under H_1 is only ≈ 2.7, so the test lacks the power to be informative for this signal.

5. **The LOO product is not valid.** While tempting, multiplying the LOO e-values ($\prod E_k = 1062$) requires independence across folds, which fails because the training sets overlap. The LOO average ($\bar{E} = 10.2$) is the correct way to combine these e-values.

These results, together with independent Bayesian [Ong et al., 2025] and tension [Wang & Mota, 2025] analyses, paint a picture in which the evidence for dynamical dark energy is moderate but not yet compelling. Valid e-value methods converge on $E \approx 10$–15—real evidence that should be taken seriously, but short of what would constitute a discovery. The most informative finding remains the cross-dataset tension: until the discrepancy between DES-Y5 and other probes is understood, the possibility that w_0w_aCDM is absorbing inter-dataset tension rather than detecting new physics cannot be ruled out. Future data from DESI DR3+, Euclid, and the Roman Space Telescope will be essential for determining whether the current hints reflect new physics or systematic effects.

Data and Code Availability

Analysis code and data files are available at:

`https://github.com/jinyoungkim927/desi-evalue-analysis`

DESI BAO data from: `https://github.com/CobayaSampler/bao_data`

A E-Value Theory

This appendix provides formal statements of the e-value properties used in the main text.

Definition 1 (E-Value)**.** *A random variable $E \geq 0$ is an **e-value** for testing H_0 if $\mathbb{E}[E \mid H_0] \leq 1$.*

Theorem 1 (Ville's Inequality)**.** *If E is an e-value, then for any $\alpha \in (0, 1)$:*

$$\mathbb{P}_{H_0}(E \geq 1/\alpha) \leq \alpha \tag{11}$$

Proof. By Markov's inequality: $\mathbb{P}(E \geq 1/\alpha) \leq \alpha \cdot \mathbb{E}[E] \leq \alpha$. □

Proposition 1 (Likelihood Ratio E-Value)**.** *For simple hypotheses $H_0 : P = P_0$ versus $H_1 : P = P_1$, the likelihood ratio $E = P_1(X)/P_0(X)$ satisfies $\mathbb{E}_{P_0}[E] = 1$.*

Proof. $\mathbb{E}_{P_0}[E] = \int \frac{P_1(x)}{P_0(x)} P_0(x)\,dx = \int P_1(x)\,dx = 1.$ □

Proposition 2 (Data-Split E-Value Validity)**.** *Let $D = (D_{train}, D_{test})$ be independent data splits. Let $\hat{\theta} = \hat{\theta}(D_{train})$ be parameters fitted to the training data. Then:*

$$E = \frac{L(D_{test} \mid \hat{\theta})}{L(D_{test} \mid H_0)} \tag{12}$$

is a valid e-value for testing H_0.

Proof. Conditional on D_{train}, $\hat{\theta}$ is fixed. Under H_0, $D_{\text{test}} \sim P_0$ independent of D_{train}. Thus:

$$\mathbb{E}[E \mid D_{\text{train}}] = \int \frac{L(x \mid \hat{\theta})}{L(x \mid H_0)} P_0(x)\,dx = 1 \tag{13}$$

by the same argument as Proposition 1. Taking expectations: $\mathbb{E}[E] = \mathbb{E}[\mathbb{E}[E \mid D_{\text{train}}]] = 1.$ □

Proposition 3 (Uniform Mixture E-Value Validity)**.** *Let $\mathcal{G} = \{\theta_1, \ldots, \theta_M\}$ be a finite set of parameter values specified before seeing the data. The uniform mixture*

$$E_{mix} = \frac{1}{M} \sum_{j=1}^{M} \frac{L(data \mid \theta_j)}{L(data \mid H_0)} \tag{14}$$

is a valid e-value.

Proof. Each term $E_j = L(\text{data} \mid \theta_j)/L(\text{data} \mid H_0)$ is a likelihood ratio for a simple alternative, so $\mathbb{E}_{H_0}[E_j] = 1$ by Proposition 1. By linearity of expectation:

$$\mathbb{E}_{H_0}[E_{\text{mix}}] = \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{H_0}[E_j] = \frac{1}{M} \cdot M = 1. \tag{15}$$

□

Proposition 4 (LOO Average E-Value Validity)**.** *Let $E_1, \ldots, E_K$ be leave-one-out e-values, where $E_k = L(D_k \mid \hat{\theta}_{-k})/L(D_k \mid H_0)$ and $\hat{\theta}_{-k}$ is fitted on all data except fold k. The average $\bar{E} = \frac{1}{K}\sum_{k=1}^{K} E_k$ is a valid e-value.*

Proof. Each E_k is a valid e-value by Proposition 2: conditional on the training data D_{-k}, $\hat{\theta}_{-k}$ is fixed, and the held-out data D_k provides an independent test. Therefore $\mathbb{E}[E_k \mid H_0] \leq 1$ for each k. By linearity:

$$\mathbb{E}[\bar{E} \mid H_0] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[E_k \mid H_0] \leq \frac{1}{K} \cdot K = 1. \tag{16}$$

Note: the *product* $\prod_k E_k$ does *not* have this guarantee because the E_k are not independent (the training sets overlap). □

Remark 1 (Gaussian Likelihoods)**.** *For multivariate Gaussian likelihoods with known covariance C:*

$$E = \exp\left(\frac{\chi^2(H_0) - \chi^2(\hat{\theta})}{2}\right) \tag{17}$$

where $\chi^2(\theta) = (d - t(\theta))^T C^{-1} (d - t(\theta))$.

References

Burnham, K. P., & Anderson, D. R. 2002, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. (Springer)

DESI Collaboration 2025, "DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints," arXiv:2503.14738

Kass, R. E., & Raftery, A. E. 1995, "Bayes Factors," Journal of the American Statistical Association, 90, 773

Liddle, A. R. 2007, "Information Criteria for Astrophysical Model Selection," MNRAS, 377, L74

Ong, D. D. Y., Yallup, D., & Handley, W. 2025, "A Bayesian Perspective on Evidence for Evolving Dark Energy," arXiv:2511.10631

Ramdas, A., Grünwald, P., Vovk, V., & Shafer, G. 2023, "Game-Theoretic Statistics and Safe Anytime-Valid Inference," Statistical Science, 38, 576

Schwarz, G. 1978, "Estimating the Dimension of a Model," Annals of Statistics, 6, 461

Shafer, G. 2021, "Testing by Betting: A Strategy for Statistical and Scientific Communication," Journal of the Royal Statistical Society Series A, 184, 407

Vovk, V., & Wang, R. 2021, "E-values: Calibration, Combination, and Applications," Annals of Statistics, 49, 1736

Wang, D., & Mota, D. 2025, "Did DESI DR2 Truly Reveal Dynamical Dark Energy?," arXiv:2504.15222