Appendix A: Midterm Exam Details

A1. Exam Guidelines

The following guidelines on *AI Reliance versus AI Assistance* were provided to students as part of this experiment.

The purpose of this experiment is to evaluate how students get assistance from Gen AI tools and how their learning improves. This is a realistic scenario in the real world. However, we do not want students to fully rely on the use of such tools. Thus, we will provide guidelines on the use of such tools. For example, we will not allow copy and paste during the exam, which will be considered an honor code violation. The idea is to use AI tools to seek assistance, but not fully rely on responses provided by AI. You will need to review the response/output from a Gen AI tool carefully. Students will be fully responsible for the errors that can arise from the use of such tools.

Moreover, in order to ensure non-mixing between the two cohorts, the course also had an Honor code violation policy as provided below:

Because we will open up the use of AI tools, we will be very strict about honor code violations. We expect all students to carefully review their computer setting to make sure they will take the exam within the required guidelines (e.g., one screen, no additional devices used) and that the honorlock will work on capturing their activity. We will review the honorlock videos for reference on use of tools, but also on identifying potential honor code violations.

Students who will be part of the cohort that will not be allowed to use Gen AI tools, their use of such tools or other external resources, except for course materials, will be considered an honor code violation.

A2. Midterm Exam 1: Multiple Choice Questions

Question 1: Which of the following statements about an ARMA(2,1) process with i.i.d. normal errors are true?

1. The process can be represented as a combination of AR and MA components with appropriate constraints for stationarity.
2. The process can always be expressed as a finite-order AR process.
3. The roots of the characteristic equation for the AR component must lie outside the unit circle for stationarity.
4. The invertibility condition ensures the MA component has a unique representation in terms of past residuals.

Question 2: For the time series $S_t = \sum_{s=1}^{t} X_s$ where X_s $IID(0, \sigma^2$, which of the following statements are correct?

1. The series is non-stationary.
2. The variance of S_t grows linearly with t.
3. The series has constant mean $E(S_t) = 0$.
4. The autocorrelation decreases with increasing lag.

Question 3: For an AR(2) process $X_t = 0.5X_{t-1} + 0.2X_{t-2} + Z_t$ where Z_t is white noise, which of the following statements are true?

1. The process is stationary.
2. The PACF cuts off after lag 2.
3. The roots of the characteristic equation determine stationarity.
4. The ACF oscillates for all lags.

Question 4: In a white noise process, which of the following statements is true?

1. The ACF is zero for all non-zero lags.
2. The PACF decays exponentially.
3. The variance is constant over time.
4. The mean is constant over time.

Question 5: When the width of a moving average window increases in trend estimation, which of the following is true?

1. The estimated trend becomes smoother.
2. Sharp fluctuations are captured less effectively.
3. The moving average model becomes more sensitive to noise.
4. The estimated trend becomes more accurate in estimating the trend.

Question 6: Which of the following statements about LOESS smoothing is true?

1. LOESS is non-parametric and handles non-linear trends.
2. LOESS performs poorly in the presence of seasonal components.
3. LOESS is well-suited for identifying irregular components.
4. LOESS requires specification of a smoothing parameter.

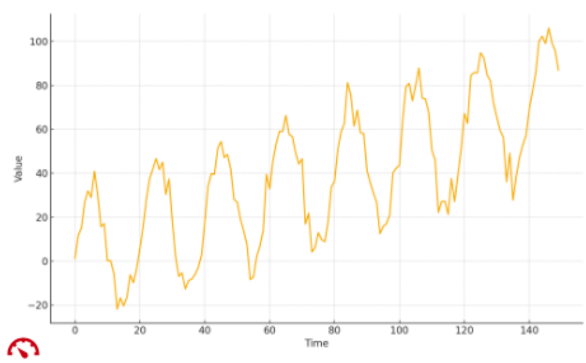
Question 7: In an ARIMA model, over-differencing can lead to which of the following issues?

1. Increased residual variance.
2. Loss of stationarity.
3. Decreased interpretability of the model.
4. Improved model performance.

Question 8: Which of the following statements are true about seasonal differencing in an ARIMA model?

1. Seasonal differencing ensures stationarity for all seasonal patterns.
2. Seasonal differencing removes periodic (seasonal) trends.
3. Overuse of seasonal differencing increases residual variance.
4. Seasonal differencing does not affect non-seasonal trends.

Question 9: Based on the graph of the time series below, which of the following methods is appropriate for modeling?

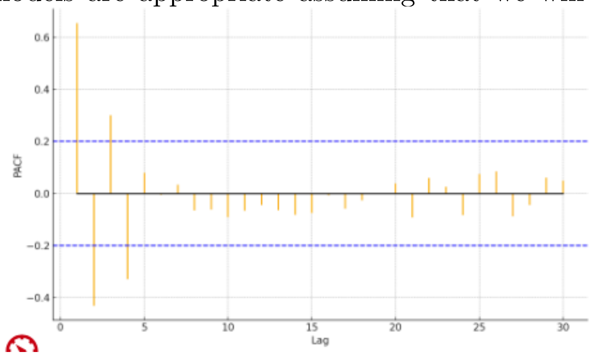


1. Differencing to address non-stationarity.
2. A simple moving average to smooth fluctuations.
3. Using ANOVA to capture seasonality.
4. Applying a regression model to fit the trend.

Question 10: Which of the following statements about Root Mean Squared Error (RMSE) are correct?

1. RMSE is robust to non-normality of residuals.
2. RMSE penalizes smaller errors more than larger ones.
3. RMSE is scale-dependent.
4. RMSE amplifies the effect of large outliers.

Question 11: Consider the PACF plot below that cuts off after lag 4. Which of the following models are appropriate assuming that we will only need to use the PACF to select the model?



1. $ARMA(4,4)$
2. $MA(4)$
3. $ARIMA(4,0,0)$
4. $AR(4)$

Question 12: The presence of a unit root implies non-stationarity, and the Augmented Dickey-Fuller (ADF) test is used to check for this property. Which of the following statements are correct?

1. The ADF test assumes the null hypothesis of stationarity.
2. Rejecting the null hypothesis indicates stationarity.
3. A unit root implies the time series is not stationary.
4. The test can include both trend and drift terms.

Question 13: Consider the $AR(2)$ process: $X_t = 0.7X_{t-1} - 0.1X_{t-2} + \epsilon_t$ where ϵ_t is white noise with mean 0 and constant variance. Which of the following statements about stationarity is correct?

1. This $AR(2)$ process is always non-stationary
2. This $AR(2)$ process is stationary because all roots of the characteristic polynomial lie outside the unit circle.
3. This $AR(2)$ process is non-invertible.
4. We cannot determine stationarity without knowing ϵ_t's distribution.

Question 14: Consider the seasonal $ARIMA_{12}(p=1, d=1, q=1)(P=1, D=1, Q=1)$ model (with periodicity=12). What is the role of the seasonal differencing parameter $(D=1)$?

1. Removes long-term trends
2. Removes seasonal trends
3. Reduces variance
4. Improves prediction accuracy

Question 15: Which of the following are properties of a stationary time series?

1. The mean is constant over time.
2. The variance depends on time.
3. The autocovariance depends only on lag.
4. The time series must be normally distributed.

Question 16: Which of the following metrics could be suitable for evaluating time series forecasting accuracy?

1. AIC (Akaike Information Criterion) for ARMA order
2. Mean Absolute Percentage Error (MAPE)
3. Root Mean Squared Error (RMSE)
4. Precision Measure

Question 17: Which of the following methods applied to a time series can help with stationarity?

1. First-order differencing.
2. Seasonal differencing.
3. Applying a logarithmic transformation
4. Adding seasonal lags.

Question 18: You have four different types of time series, and you want to decide whether an ARIMA model is a typical or practical choice for forecasting each. Which of the following would usually be modeled with ARIMA?

1. A time series with trend with seasonality.
2. A random walk.
3. A white noise process
4. A stationary process.

Question 19: Which of the following expressions is correct?

1. $X_t - \phi_1 X_{t-1} - phi_2 X_{t-2} = (1 - \phi_1 B - \phi_2 B)$
2. $X_t + \phi_1 X_{t-1} - phi_2 X_{t-2} = (1 - \phi_1 B - \phi_2 B^2)$
3.
4. $X_t - \phi_1 X_{t-1} - phi_2 X_{t-2} = (1 - \phi_1 B - \phi_2 B^2)$

Question 20: Which of the following assumptions are made by the KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test?

1. The null hypothesis assumes stationarity.
2. The test detects deterministic trends in a time series.
3. The test assumes the presence of a unit root under the null hypothesis.

4. The test requires specifying a bandwidth parameter for the long-run variance estimate.

Question 21: Which of the following are true about the Ljung-Box test applied to the residuals from an ARIMA model?

1. It tests for residual autocorrelation.
2. It assumes the residuals are white noise in the null hypothesis.
3. It is used to identify overfitting in models.
4. It evaluates whether the model fits the data well.

Question 22: Consider the following $ARMA(2,2)$ model $X_t - a_1 X_{t-1} - bX_{t-2} = Z_t + cZ_{t-1} + dZ_{t-2}$. For what combination of (a,b,c,d) values is this $ARMA(2,2)$ process causal?

1. $(a,b,c,d) = (0.9. -0.2, 0.7, 0.8)$
2. $(a,b,c,d) = (1, 2, 0.7, 0.8)$
3. $(a,b,c,d) = (6, -0.2, -0.7, 0.8)$
4. $(a,b,c,d) = (3, 0, 1.2, 5)$

Question 23: Which of the following are limitations of ARIMA models?

1. They cannot handle non-linear relationships.
2. They require non-stationarity.
3. They do not support seasonal patterns without extension.
4. They perform poorly on short time series.

Question 24: After fitting a model to time series data, residuals are obtained. If we run a Shapiro Wilk test to the residuals, and get a p-value of 0.84. What conclusion would you make?

1. The residuals exhibit autocorrelation
2. The residuals are not normally distributed
3. The residuals do not exhibit autocorrelation
4. The residuals are normally distributed.

Question 25: Which of the following statements about white noise is correct?

1. Gaussian white noise is always stationary, provided it has a constant mean and variance over time.
2. Non-Gaussian white noise is never stationary.
3. White noise terms must be correlated.
4. White noise processes must change variance over time.

Question 26: Which of the following statements about forecasting with ARIMA models is correct?

1. AR terms capture dependencies on past values to model trends and generate multi-step forecasts.
2. Differencing improves forecast stability by reducing the risk of trend-driven overestimations.
3. MA terms handle random shocks and help refine short-term predictions.
4. None of the statements are true

6.A. A3. Midterm Exam 2: Multiple Choice Questions

Question 1: Which statement correctly describes the conditional variance structure in the ARCH model?

1. The variance depends on the future data points only.
2. The variance remains strictly constant at all times.
3. The variance must be negative for large lags.
4. The variance depends on past squared data.

Question 2: Which of the following statements is correct about the GARCH model?

1. GARCH models apply to strictly non-stationary time series.
2. GARCH models time-varying conditional variance.
3. GARCH requires the unconditional variance to be unbounded.
4. GARCH models cannot model time-varying volatility.

Question 3: Which of the following statements best describes how volatility clustering can be incorporated into the residuals of an ARMA model?

1. The conditional variance remains constant over time, so modeling volatility clustering is unnecessary.
2. The residual process Z_t can be expressed as $\sigma_t R_t$ where σ_t^2 is a time-varying conditional variance and R_t a white noise process with mean zero and constant variance.
3. ARMA model residuals are unaffected by past volatility and therefore do not exhibit volatility clustering.
4. The residual must have zero variance at every time t.

Question 4: Which statements accurately describe the use of AIC and BIC when choosing time series models (e.g., ARMA, GARCH, VAR)?

1. Both AIC and BIC balance model fit with model complexity, penalizing models with more parameters.
2. BIC typically imposes a stricter penalty for additional parameters than AIC, favoring simpler models.
3. BIC often selects smaller models when the sample size is large, aiming to avoid overfitting.

4. Neither AIC nor BIC can be used for GARCH models because they only apply to ARMA models.

Question 5: Consider $Z_t^2 = 0.5 + 1.2Z_{t-1}^2$. Which statement is correct about stationarity?

1. It is always weakly stationary because $\alpha_1 = 0.8$, thus less than 1.
2. It is always weakly stationary.
3. It is not weakly stationary because $\alpha_1 = 0.8$, thus less than 1.
4. It is not stationary at all.

Question 6: Which of the following statements best describes the relationship between the presence or absence of univariate AR(p) causality in the individual time series within a VAR(p) model and the stability of the entire VAR(p) system?

1. If every individual time series in a VAR(p) exhibits univariate AR(p) causality, then the entire VAR(p) system must be stable.
2. If even one individual time series in a VAR(p) does not exhibit univariate AR(p) causality, then the entire VAR(p) system must be unstable.
3. The presence or absence of univariate AR(p) causality in the individual time series provides a direct and definitive indication of the stability of the entire VAR(p) system.
4. The presence or absence of univariate AR(p) causality in the individual time series does not, on its own, guarantee or preclude the stability of the entire VAR(p) system, which depends on the eigenvalues of the companion matrix.

Question 7: Which of the following aspects must be considered when checking second-moment stationarity in a multivariate series?

1. Only cross-dependence matters; auto-correlations in each time series are irrelevant.
2. Only the univariate auto-covariances matter; cross-covariances are irrelevant.
3. Both the auto-covariances within each series and cross-covariances among series must be (time) invariant.
4. Stationarity in the second moment does not apply to multivariate data.

Question 8: Which statement is correct about an $ARCH(q)$) process with time-varying conditional variance?

1. Time-varying conditional variance automatically makes it non-stationary.
2. It can still be stationary if the ARCH coefficients satisfy certain constraints (e.g., sum ¡ 1).
3. It must have an infinite unconditional variance.
4. It is never suitable for modeling financial data.

Question 9: Which of the following statements correctly describes deterministic covariates in a $VARX(p)$?

1. They can include trend terms, seasonal dummies, or other user-specified deterministic components.
2. They prohibit including a trend of any kind.
3. They allow for the AR part only, with no covariates at all.
4. They must be purely random, not deterministic.

Question 10: If we are able to predict the future values of Z_t^2 which represents the squared value of a time series, what does this imply?

1. We can accurately predict both how large the changes in volatility will be and whether volatility will increase or decrease.
2. We can predict the direction (increase or decrease) of changes in volatility, but not how large those changes will be.
3. We can estimate how large future volatility will be, but not whether it will increase or decrease.
4. Being able to forecast variance means that we can also perfectly predict the future values of the returns themselves.

Question 11: Which of the following statement is correct for an unrestricted $VAR(p)$ with contemporaneous correlation in the errors?

1. It can use different lags to model each individual time series.
2. It cannot use the standard ordinary least squares to estimate the unrestricted $VAR(p)$.
3. It requires each univariate series to be independent.
4. It can be fitted via separate OLS equations for each time series, each including the same lags of all series.

Question 12: Which statements accurately characterize why heteroskedasticity models (ARCH/GARCH) are introduced for certain time series?

1. A time series can be lack auto-correlation but still display time-dependent volatility.
2. Higher-order ARMA models alone always suffice to capture volatility clusters.
3. The conditional variance may change over time even if the unconditional variance is constant.
4. GARCH models allow past shocks in volatility to persist longer than in the simple ARCH.

Question 13: Which statements correctly describe the structure or implications of a $GARCH(p, q)$ model?

1. The σ_t^2 term depends on σ_{t-1}^2 and older variances, enabling "long memory."
2. If the sum of GARCH coefficients $\alpha_i + \beta_j$ exceeds 1, the series might fail to be stationary.

3. Only the mean equation matters in GARCH; the variance is assumed constant.
4. $GARCH(p,q)$ implies no volatility clustering in financial returns.

Question 14: A series may have zero autocorrelation but still be dependent in higher moments. Which of the following statements accurately reflect this phenomenon?

1. Independence and uncorrelatedness are equivalent for all time series processes.
2. ACF plots of the squared series can uncover dependence that's not identified using the ACF plot of the time series itself.
3. Finding no autocorrelation in Z_t automatically guarantees Z_t^2 is also uncorrelated.
4. Detecting significant correlation in Z_t^2 suggests heteroskedasticity

For questions 15-16 use the following R output from a ARMA(1,2)-GARCH(1,1):

	Estimate	Std. Error	t value
mu	1.67843	0.37760	4.44503
ar1	-0.36387	0.69681	-0.52220
ma1	0.26365	0.70245	0.37532
ma2	0.13467	0.17098	0.78761
omega	10.23761	14.49657	0.70621
alpha1	0.11798	0.19833	0.59488
beta1	0.79752	0.53754	1.48364
shape	3.07252	4.63372	0.66308

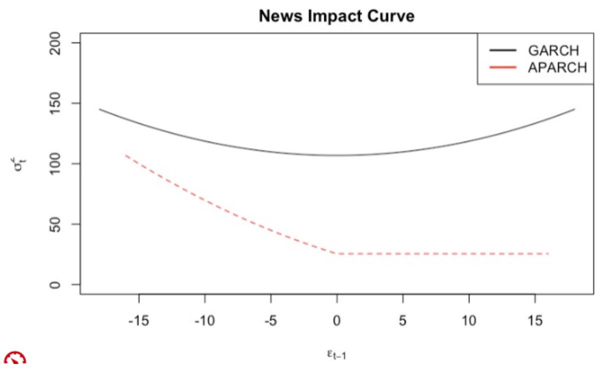
Question 15: From Table 1, we can say that:
Hint: Remember the stationarity conditions given the values of the estimated coefficients from the table.

1. The GARCH(1,1) is stationary.
2. The GARCH(1,1) is not stationary.
3. We may need to estimate a IGARCH
4. None of the above

Question 16: What is the unconditional variance based on the results in Table 1?

1. 11.61
2. 5.34
3. 121.16
4. More information needed

Question 17: From the following News Impact Curve, obtained after modeling a GARCH and a APARCH, which of the following statements is correct:



1. Both models, GARCH and EGARCH assign the same weight to positive and negative shocks
2. APARCH assigns different weights to negative shocks.
3. The kurtosis of both models is lower than 3
4. Leverage effect is successfully modeled by the two models.

Question 18: Which of the following conditions or implications are consistent with weak stationarity in an Autoregressive Conditional Heteroskedasticity of order q (ARCH(q)) process?
Hint: In an ARCH(q) process defined as

$$\sigma_t^2 = \alpha_0 + \alpha_1\sigma_{t-1}^2 + \alpha_2\sigma_{t-2}^2 + ... + +\alpha_q\sigma_{t-q}^2 :$$

• α_0: Represents the constant term or the baseline level of the conditional variance. It must be positive.
• α_i for $i = 1,,,,q$: Represent the coefficients on the lagged squared residuals. These coefficients capture the sensitivity of the current conditional variance to past shocks. They are typically non-negative.

1. For weak stationarity, it is necessary that α_0 and α_i for $i = 1,,,,q$ and the sum of the coefficients on the lagged squared residuals is less than one.
2. Weak stationarity in an ARCH(q) process implies that the conditional variance σ_t^2 must be constant over time.
3. If the sum of the coefficients on the past squared residuals is greater than or equal to one ($\sum_{i=1}^{q}\alpha_i$), weak stationarity of the ARCH(q) process is guaranteed.
4. A key characteristic of weak stationarity is that the unconditional variance of the process remains constant over time.

Question 19: We performed a Portmanteau Test on the residuals of a VAR(4) model. The R output is the following:

```
	Portmanteau Test (asymptotic)

data:  Residuals of VAR object model.var
Chi-squared = 34.071, df = 48, p-value = 0.9356
```

From this result, we would interpret that:

1. The residuals are not normally distributed.
2. The residuals are not serially correlated
3. The residuals distribution has fat tails
4. The residuals have constant variance.

Question 20: Suppose we fit an unrestricted VAR model to two time series called Mic and Int. The selected order was 6, and we applied a Wald test using all lagged variables of Int in the Mic equation (see Output 1) and, using all lagged variables of Mic in the question of Int (see Output 2). Given the following results, with a significance level of 0.05, what can you say about Granger Causality between these variables:

Output 1

```
Wald test:
----------

Chi-squared test:
X2 = 0.5, df = 4, P(> X2) = 0.97
```

Output 2

```
Wald test:
----------

Chi-squared test:
X2 = 8.0, df = 4, P(> X2) = 0.09
```

1. Mic Granger causes Int but Int does not Granger cause Mic
2. Int Grange causes Mic but Mic does not Granger cause Int
3. Both variables Granger Cause each other
4. Neither variables Granger Causes each other

Question 21: Which statements correctly reflect the concept or use of exogenous variables in a VARX model?

1. Exogenous variables can influence the response series but are not influenced by them.
2. In a $VARX(p)$, exogenous variables have no role in forecasting the dependent series.
3. Deterministic trend or seasonal dummies can be treated as exogenous terms.
4. If the exogenous series depends on the dependent series, it ceases to be "exogenous."

Question 22: We estimated an Unrestricted VAR(4) for the time series Mic and Int. The following is the R output of the equation for Mic.

ADD Output

If we want to run a Restricted VAR, how many variables the Mic equation will have with a significant level of 0.1?

1. 6
2. 3
3. 4
4. 5

Question 23: Which statements correctly describe the concept of cointegration?

1. Cointegration occurs when a specific linear combination of two or more individual time series, each of which is non-stationary (meaning their statistical properties like mean and variance change over time), results in a new time series that is stationary (meaning its statistical properties are constant over time). This suggests a long-run relationship where the variables move together and deviations from this relationship are temporary.
2. The concept of cointegration implies the existence of a "long-run equilibrium" relationship among the non-stationary variables. This equilibrium means that while the individual series may wander away from each other in the short run, there are economic or other forces that pull them back towards a stable, long-term relationship. The cointegrating equation estimates this long-run equilibrium.
3. The Engle-Granger procedure is a statistical method used to test for cointegration between time series. It involves first estimating a potential long-run relationship using Ordinary Least Squares (OLS) and then examining the residuals (the deviations from this estimated relationship) for stationarity using a unit root test. If the residuals are found to be stationary, it provides evidence of cointegration.
4. If two time series are found to be cointegrated, simply differencing them to achieve stationarity, while mathematically valid, would eliminate the information about their long-run equilibrium relationship. While differencing can be useful for modeling short-run dynamics, it discards the crucial insight that the series move together in the long run and that deviations from their equilibrium are mean-reverting.

Question 24: Which are valid features or requirements of a Structural VAR?

1. It can capture direct same-period interactions between variables via transformation of the time series.
2. Constraints on certain parameters (e.g., A or B matrices) are needed to ensure identifiability.
3. In SVAR, all error terms remain perfectly uncorrelated at all lags and leads.
4. Without constraints, multiple parameter solutions can fit the same reduced-form VAR.

Question 25: Which statements reflect important conditions or implications for a stationary multivariate time series?

1. The cross-covariance $Cov(Y_t^{(i)}, Y_{t-h}^{(j)})$ can depend on t.
2. Each series must individually have constant mean and finite variance.
3. Cross-covariances must only depend on lag h, not on the specific time t.
4. Stationarity in the second moment requires both auto- and cross-covariances to be time-invariant.

Appendix B: Additional Results

This Appendix B provides additional supporting results for the analysis provided in the main manuscript. These results complement those provided in the manuscript.

B1. Preliminary GenAI Survey

This section presents the set of questions in a survey administered prior to the Gen-AI experiment. The questions were as follows:

- Have you ever used AI tools at work, in your studies or personal life to automate tasks or improve productivity?
- In your academic studies, have you utilized AI tools for research or assignments?
- Have you ever used AI-powered tools for creative projects, such as generating art, music, or writing?
- In your academic or professional work, have you used AI for data analysis or generating insights?
- Have you ever used AI-powered tools for communication, such as language translation or email management?
- What is the primary purpose of using generative AI tools in creative tasks?
- What is the common benefit of using generative AI tools in educational settings?
- In which scenario would using generative AI be particularly advantageous?
- What is a potential drawback of over-reliance on generative AI tools for academic tasks?

Figures B1 and B2 display the distribution of the responses from this pre-survey. Based on this analysis, not all students have interacted with Gen AI tools before the start of the semester.

B2. Student-Level Performance Analysis

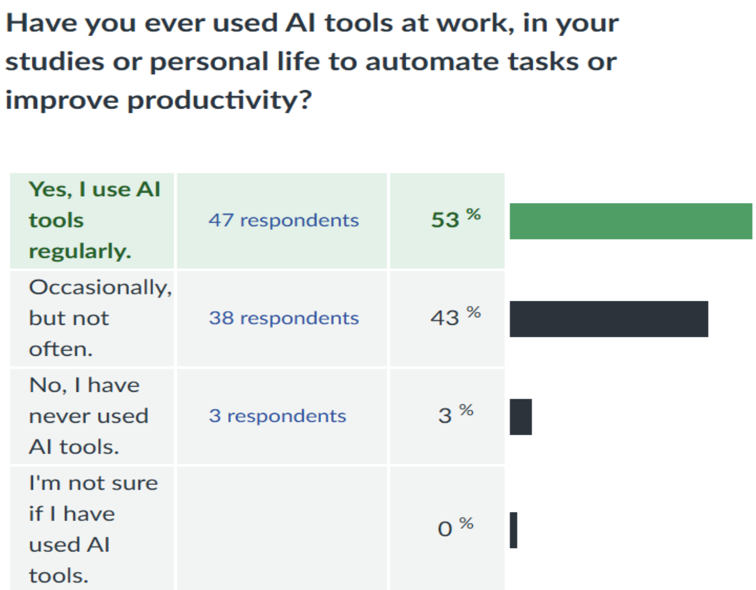
This subsection provides additional supporting figures and results for the performance comparison for the two cohorts of students for the two midterms.

Figure B3 and Table B1 show that the distribution across the four degree programs is not substantively different for the two cohorts. Nonetheless, we include the degree program as a controlling factor in our analysis as it might have an impact because of students' prior exposure.

	Cohort 1				Cohort 2			
Program	Progr 1	Progr 2	Progr 3	Progr 4	Progr 1	Progr 2	Progr 3	Progr 4
Student Count	27	11	6	6	27	11	7	10

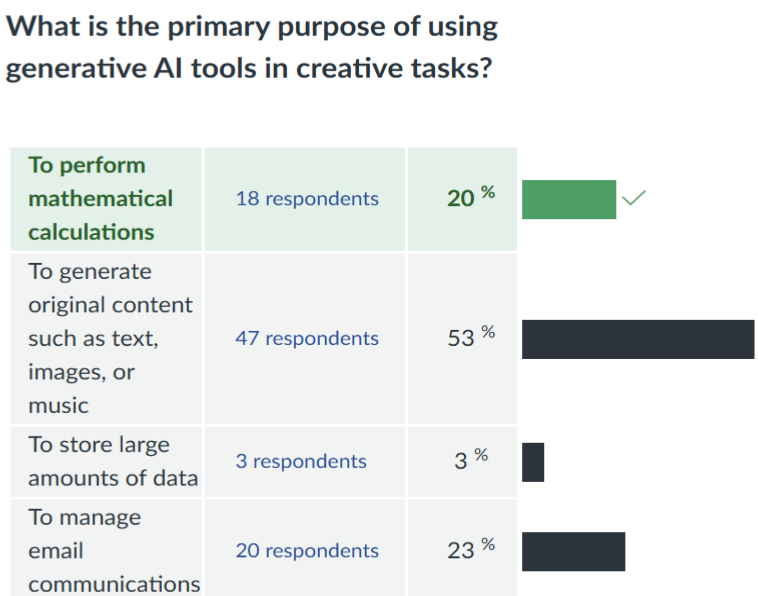
Table B1 Distribution of Student Count across the four programs.

Similarly, Figure B4 shows that students in the two cohorts had similar levels of exposure to AI tools. Cohort 2 has used AI tools for email and mathematical support to a slightly higher degree



0.5

Figure B1 **Survey Responses to Question on Prior Gen AI Use**



0.5

Figure B2 **Survey Responses to Question on Perception of Gen AI Utility**

than Cohort 1, and Cohort 1 has used AI slightly more for original content. To this end, we also include these two factors to control for prior exposure.

Figure B5 presents the frequency of students using different AI tools for questions in each midterm. Note that the use of AI tool is recorded by question by student. ChatGPT has the highest use, however most students decided not to use AI tools at all for many of the questions. Moreover,

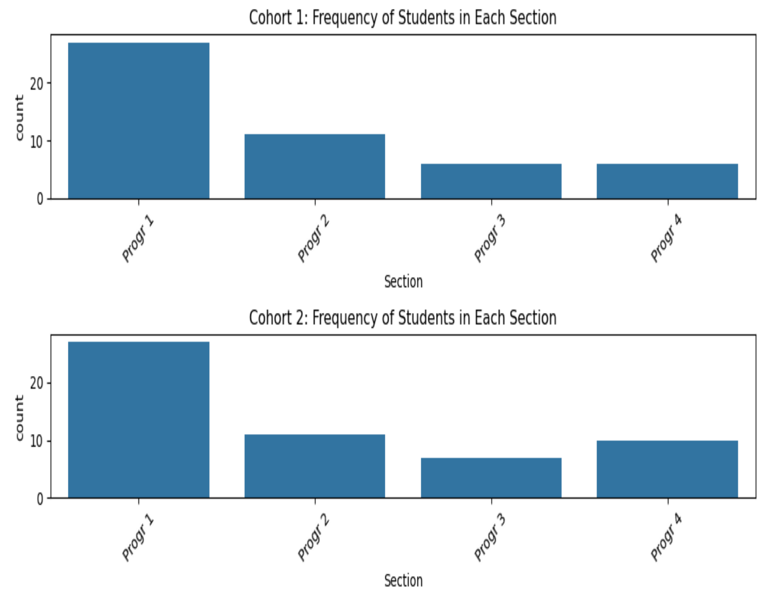


Figure B3 **The distribution of students within each cohort across the four programs.**

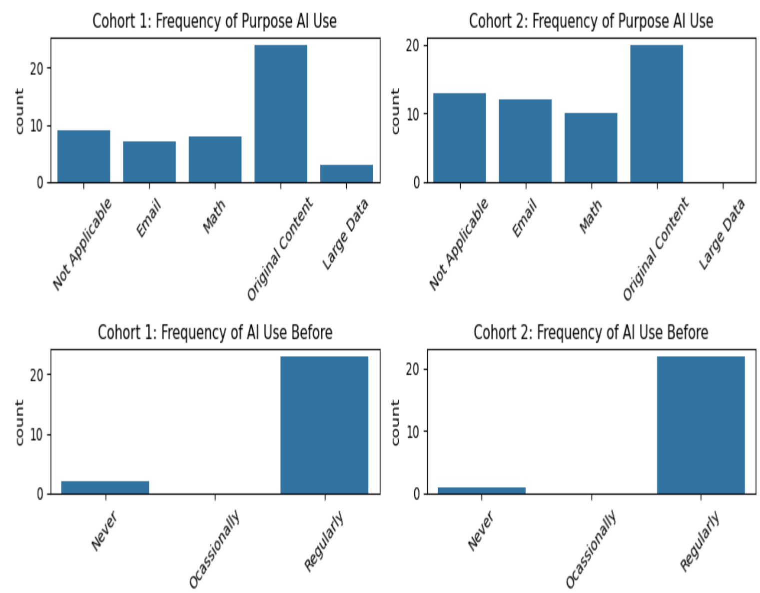


Figure B4 **Students' prior exposure to Use of AI Tools within each cohort.**

apart from AI use, students have not used external resources for the most part. The results are similar for both midterms.

Figure B6 presents the frequency of questions within each category of time spent on question and within each category of time-based engagement, comparing the Gen AI to the Non-Gen AI cohort for each midterm. The results for the two midterms look very similar. However, the temporal mode measures have different distributions for the two cohorts, Gen AI vs Non Gen AI.

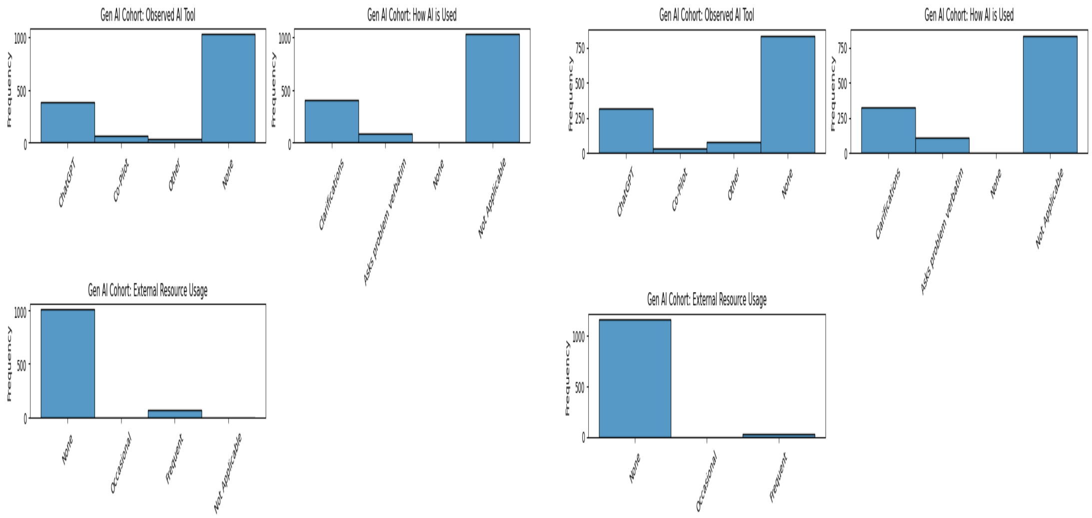


Figure B5 AI Tools used by students in the Gen AI cohort in midterm 1 (left) and midterm 2 (right).

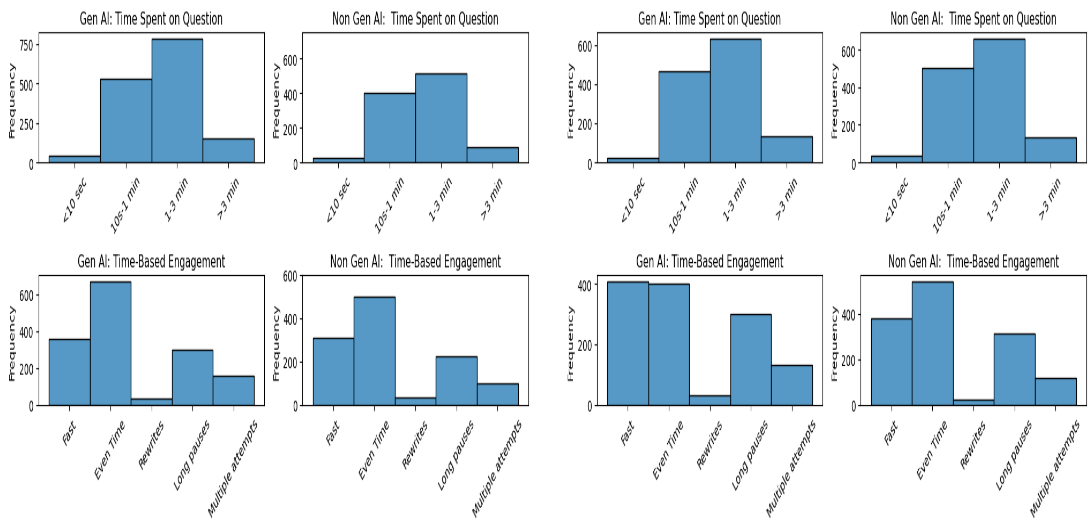


Figure B6 Temporal Mode Measures: Gen AI vs Non Gen AI for both midterms, midterm 1 in left plots and midterm 2 in the right plots.

Figure B2 shows whether students have used course resources to respond the MC questions for both midterms, comparing again the Gen AI and Non Gen AI cohorts. As expected the Non Gen AI students used the course resources to a greater extent.

Figures B3 and B9 provide supporting material for the association analysis of the midterm grades versus the factors collected from the videos. The factors have been renamed to have shorter denominations for ease of visualization of the results. The descriptions of the factors recorded for each student included in the models are as following:

- *Grades*: Midterm grade;

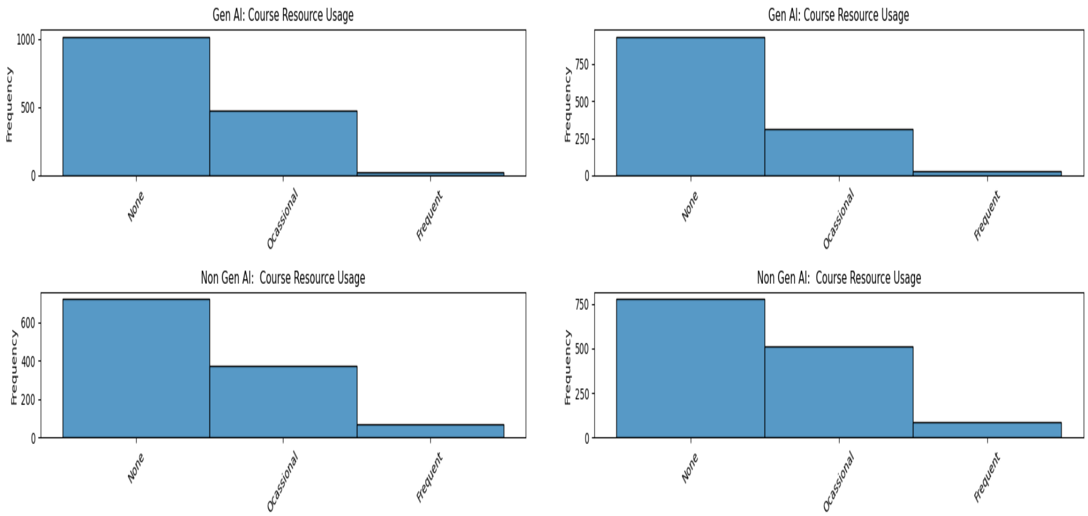


Figure B7 Course Resources: Gen AI vs Non Gen AI for both midterms, midterm 1 in left plots and midterm 2 in the right plots.

- *AIYes*: The rate of use of AI;
- *ChatGPT*:The rate of use of ChatGPT;

- *CoPilot*: The rate of use of CoPilot;
- *Clarifications*: The rate of seeking clarifications using GenAI tools;
- *Verbatim*: The rate of seeking a response to a question by typing verbatim the question into a GenAI tool;
- *Ex_Frequent*: The rate of frequently using external resources (not including GenAI tools) to respond to exam questions;
- *Co_Occasional*: The rate of occasionally using course resources to respond to exam questions;
- *Co_Frequent*: The rate of frequently using course resources to respond to exam questions;
- *Time1min*: The rate of answering a question under 1 minute;
- *Time3min*: The rate of answering a question between 1 and 3 minutes;
- *Time3minmore*: The rate of answering a question in more than 3 minutes
- *Multiple_Attempts*: The rate of attempting multiple times in answering questions;
- *Long_Pause*: The rate of taking long pauses in answering questions;
- *Frequent_Rewrite*: The rate of frequent in answering questions;
- *AIPurpose_Math*: Binary variable stating whether the student found the purpose of using GenAI for mathematical derivations;
- *AIPurpose_Original*: Binary variable stating whether the student found the original purpose of GenAI use;

- *AIPurpose_Email*: Binary variable stating whether the student found the GenAI useful in writing emails;
- *Progr1*: Program degree 1;
- *Progr2*: Program degree 2;
- *Progr3*: Program degree 3.

The rates above are derived for each student across the midterm questions within one midterm exam.

Figure B6 shows the marginal association measured by the (linear) correlation measure between the quantitative factors, including the response (Grades) and the predicting factors. The figure compares the association for the two cohorts (Gen AI versus Non-Gen AI) and for the two midterms.

The correlations between Grades and all quantitative predicting factors are provided on the top row (first column). Comparing the associations with the grades of the GenAI cohort from the two midterms, the correlation values maintain their sign across the two midterms, but with values closer to zero for Midterm 1 across most predicting factors. This is not found however when comparing the correlations for the Non GenAI cohort. For the GenAI cohort, use of CoPilot and use of the course materials, a time longer than 3 minutes, long pauses and frequent rewrites all have a negative association with the grades for both midterms. The use ChatGPT is positive for Midterm 2. For the Non GenAI cohort, all correlations with the grades are negative, except response time less than 1 minute and multiple attempts, are negative.

Table B2 shows the values of the coefficients from the lasso regression applied to the student-level midterm scores. In the table, the value 0 indicates that the corresponding variable has not been selected by LASSO and $-$ indicates that the corresponding variable has not been captured in the model.

Figure B2 shows the paths of the regression coefficients varying with the change in the regularization parameter from the LASSO regression applied to the four models, Gen AI for both midterms and Non-Gen AI for both midterms.

Table B3 shows the values of the coefficients from the lasso regression applied to the logistic regression model (binary factor - whether the student answered the question correctly or not). In the table, the value 0 indicates that the corresponding variable has not been selected by LASSO and $-$ indicates that the corresponding variable has not been captured in the model.

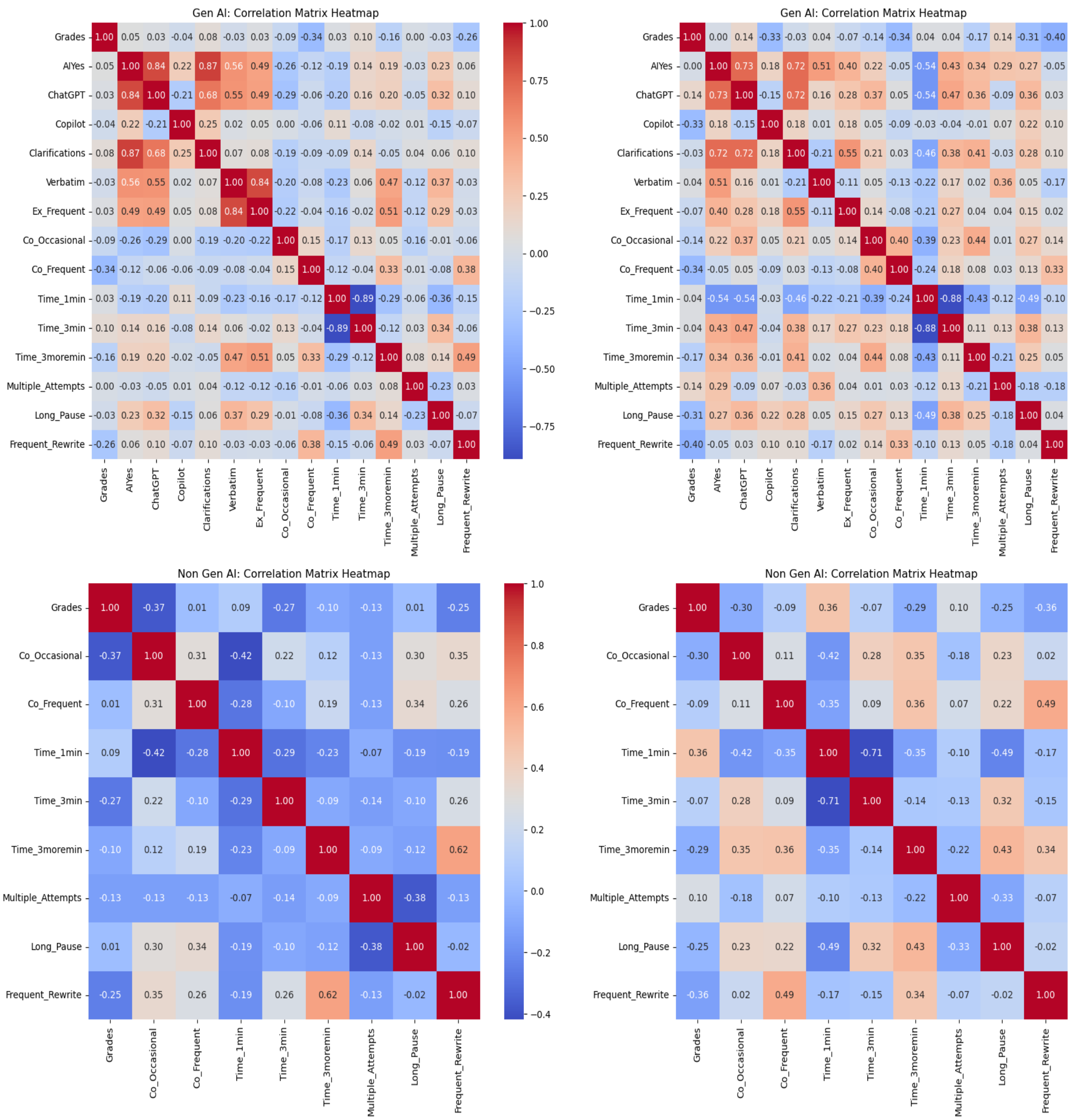


Figure B8 **Correlation Matrix Plot: Marginal association measured using the correlation measure comparing it across the two midterms for the Gen AI cohort (upper plots) and for the Non-Gen AI cohort (lower plots). Thr left plots correspond to Midterm 1 and right plots to Midterm 2.**

Figure B10 shows the paths of the regression coefficients varying with the change in the regularization parameter from the LASSO regression applied to the four logistic regression models, Gen AI for both midterms and Non-Gen AI for both midterms.

Predictor	GenAI M1	Non-GenAI M1	GenAI M2	Non-GenAI M2
AIYes	0	—	0	—
ChatGPT	0	—	1.47	—
Copilot	0	—	-5.85	—
Clarifications	0	—	0	—
Verbatim	0	—	0	—
Ex_Frequent	0	—	0	—
Co_Occasional	0	-2.68	0	-1.28
Co_Frequent	-18.93	0	-30.13	0
Time_1min	0	0	0	3.94
Time_3min	0	-1.64	2.53	0
Time_3moremin	0	0	-0.91	0
Multiple_Attempts	0	0	1.36	0
Long_Pause	0	0	-5.74	0
Frequent_Rewrite	-7.01	-2.63	-14.66	-23.79
AIBefore_Regular	0	0	0.30	0
AIPurpose_Math	0	1.01	3.00	0
AIPurpose_Original	0	-2.32	-1.96	0
AIPurpose_Email	0	-1.60	0	0
Progr1	-1.20	0	-0.68	0
Progr2	0	0	0	0
Progr3	0	-0.34	0	0

Table B2 LASSO Coefficients: LASSO regularized regression applied to the midterm scores, comparing it across the two midterms for the Gen AI cohort and for the Non-Gen AI cohort.

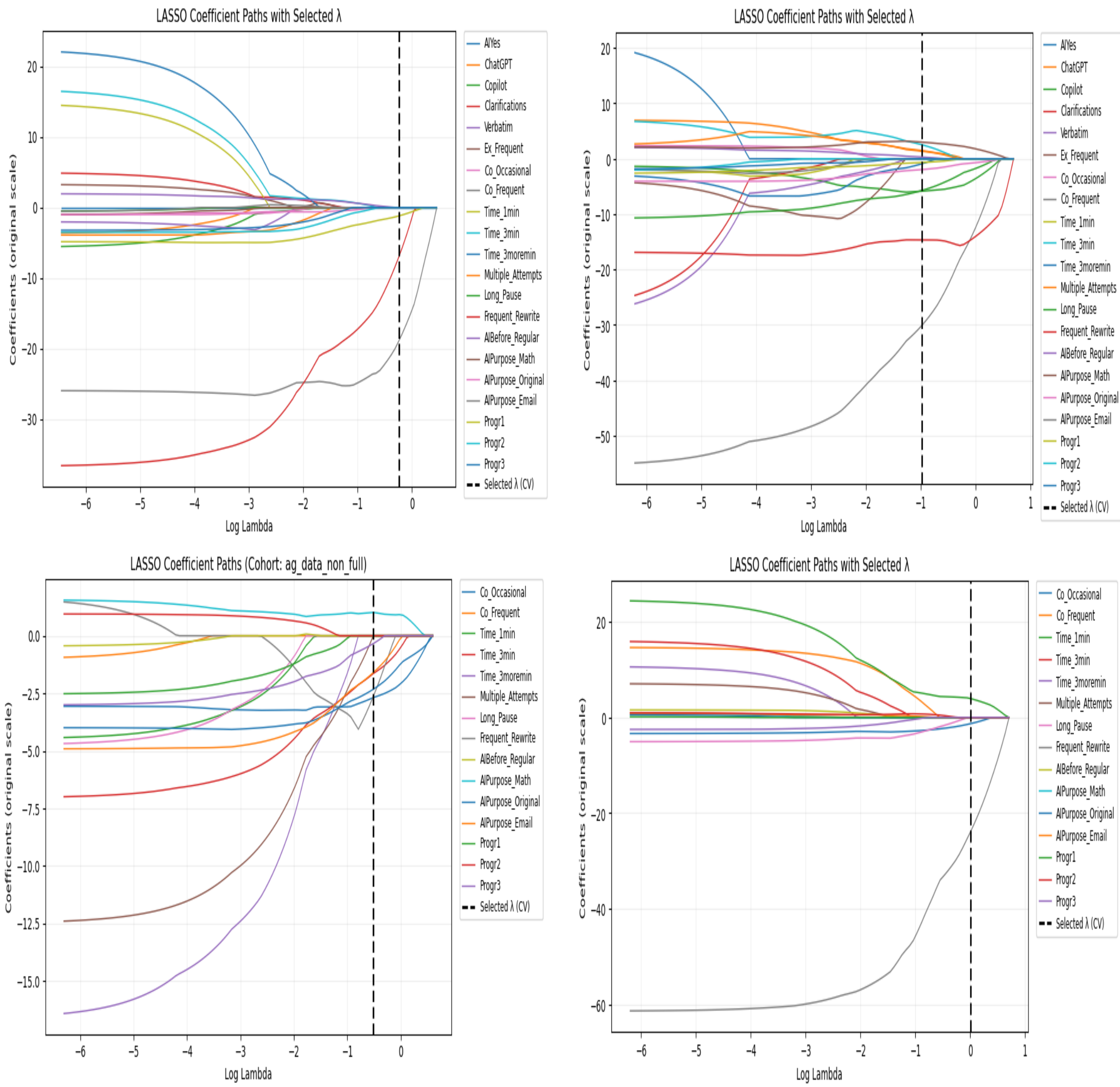


Figure B9 **LASSO Coefficient Path Plots: Coefficient path plots derived from application of LASSO regularized regression to the midterm scores, comparing it across the two midterms for the Gen AI cohort (upper plots) and for the Non-Gen AI cohort (lower plots).**

Predictor	GenAI M1	Non-GenAI M1	GenAI M2	Non-GenAI M
AI Used?_Yes	0	—	0	—
ObservedAITool_Co-Pilot	0	—	0	—
ObservedAITool_None	0	—	0	—
ObservedAITool_Other	0	—	0	—
HowAIisUsed_Clarifications	0	—	0	—
HowAIisUsed_Not Applicable	0	—	0	—
HowAIisUsed_Asks problem verbatim	—	—	0	—
ExternalResourceUsage_None	0	—	0	—
ExternalResourceUsage_Occasional	0	—	—	—
ExternalResourceUsage_Ocassional	—	—	0	—
CourseResourceUsage_None	0	0.08	0	0
CourseResourceUsage_Occasional	0	0	—	—
CourseResourceUsage_Ocassional	—	—	0	0
TimeSpentonQuestion_<1 min	0	0.01	0	0
TimeSpentonQuestion_>3 min	0	0	0	0
TimeSpentonQuestion_10s-1min	—	0	—	—
TimeSpentonQuestion_1-3min	—	0	—	—
TimeSpentonQuestion_3	—	0	—	—
TimeBasedEngagement_Fast	0.05	0.09	0.00	0
TimeBasedEngagement_Long pauses	0	0	0	0
TimeBasedEngagement_Multiple attempts	0	-0.07	0	0
SectionExtracted_MSA	0	0	—	—
SectionExtracted_O01	0	0	—	—
SectionExtracted_OAN	-0.02	0	—	—
Section_MSA	—	—	0	0
Section_O01	—	—	0	0
Section_OAN	—	—	0	0
UseAIBefore_Not Applicable	0	0	0	0
UseAIBefore_Occasionally	0	0	0	—
UseAIBefore_Regularly	0	0	0	—
UseAIBefore_Occasionally, but not often.	—	—	—	0
UseAIBefore_Yes, I use AI tools regularly.	—	—	—	0
PurposeUseAI_Math	0	0	—	—
PurposeUseAI_Original	0	0	0	0
PurposeUseAI_Not Applicable	0	0	0	0
PurposeUseAI_To store large amounts of data	—	0	—	—
PurposeUseAI_Math&Data	—	—	0.00	0
Grades	0.62	0.38	—	—
Points Awarded	1.12	0.28	—	—
MidGrade	—	—	0.78	0.57
ControlQGrade	—	—	0.55	0

Table B3 LASSO Coefficients: LASSO regularized regression applied to the logistic regression, comparing it across the two midterms for the Gen AI cohort and for the Non-Gen AI cohort.

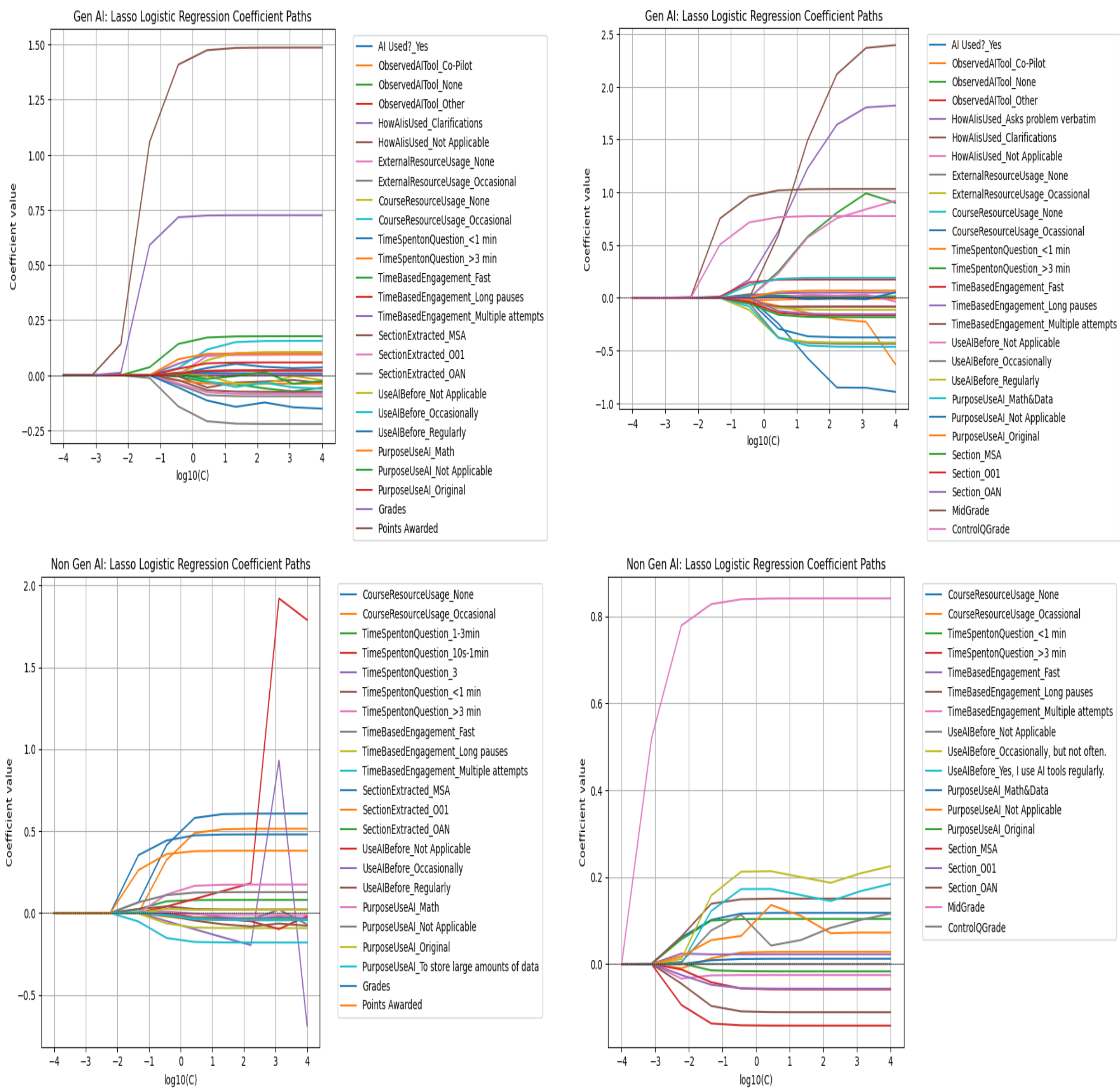


Figure B10 **LASSO Coefficient Path Plots: Coefficient path plots derived from application of LASSO regularized regression to the logistic regression, comparing it across the two midterms for the Gen AI cohort (upper plots) and for the Non-Gen AI cohort (lower plots).**